Guiding Large Language Models for Biomedical Entity Linking via Restrictive and Contrastive Decoding

Anonymous ACL submission

Abstract

Biomedical entity linking (BioEL) aims at mapping biomedical mentions to pre-defined entities. While extensive research efforts have been devoted to BioEL, applying large language models (LLMs) for BioEL has not been fully explored. Previous attempts have revealed difficulties when directly applying LLMs to the task of BioEL. Possible errors include generating non-entity sentences, invalid entities, or incorrect answers. To this end, we introduce LLM4BioEL, a concise yet effective framework that enables LLMs to adapt well to the BioEL task. LLM4BioEL employs restrictive decoding to ensure the generation of valid entities and utilizes entropy-based contrastive decoding to incorporate additional biomedical knowledge without requiring further tuning. Besides, we implement few-shot prompting to maximize the in-context learning capabilities of LLM. Extensive experiments demonstrate the effectiveness and applicability of LLM4BioEL across different BioEL tasks and with different LLM backbones, and the best-performing LLM4BioEL variant outperforms the traditional and LLM-based BioEL baselines.

1 Introduction

Biomedical entity linking (BioEL) serves as the foundation for tasks like biomedical KG construction (Zhang et al., 2020; Yu et al., 2022), KG-based answering (Shi et al., 2023; Yang et al., 2024b), and automatic diagnosis (Qiao et al., 2020; Shi et al., 2022; Zhao et al., 2024). BioEL aims at recognizing the biomedical mentions and linking them to standard entities with valid concept unique IDs (CUIs) in the given medical knowledge graphs (KGs), such as UMLS (Bodenreider, 2004). In contrast to the general-domain entity linking, BioEL is characterized by a wide range of diverse and fine-grained biomedical concepts. A single biomedical entity can exhibit multiple morphological variations, such as "autosomal recessive disorder" and "diseases inborn genetic", while different biomedical entities may share similar surface forms, like "neoplasm of autonomic nerve disorder" and "neoplasm of vagus nerve disorder". These complexities present greater challenges for BioEL methods, requiring the ability to understand and capture their nuanced relationships and distinctions.

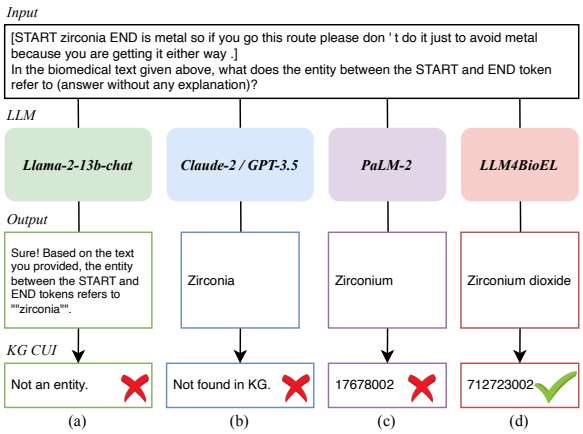


Figure 1: An example BioEL input and different LLMs' output, where (a) represents the non-entity output that is not an entity as required, (b) represents the invalid output that is not found in the given medical KG, (c) represents the incorrect output, and (d) represents the correct output.

Most current BioEL methods are either *discriminative* or *generative* methods. The *discriminative* methods employ BERT-based models to encode the biomedical mentions and entities and retrieve the most similar entities using embedding similarities. Some *discriminative* methods further utilize cross-encoders to rerank the retrieved entities via modeling the fine-grained mention-entity interactions. The *generative* methods directly generate the linked entities based on the task-specific language models, such as pre-training BART on BioEL datasets. A recent benchmarking study (Jahan et al., 2023) investigated and evaluated LLMs' perfor-

mance on various biomedical tasks in an end-to-end manner. Considering the sensitivity to prompts, Jahan et al. (2023) studied how to construct prompts for LLMs to simulate common biomedical tasks effectively. For BioEL, the prompt is designed as the "Input" in Figure 1, where LLMs are probed to directly generate the correct entities. We present the outputs of four different LLMs and our method, LLM4BioEL, with the same input in Figure 1.[1] We observe three scenarios when LLM's output is incorrect. (a) LLM does not fully follow the instruction and the output is not an entity as required, such as Llama-2-13b-Chat; (b) LLM follows the instruction but outputs an invalid entity that is not in the given biomedical KG, such as Claude-2[2] and GPT-3.5;[3] (c) LLM outputs an entity with a valid CUI that is incorrect answer, such as PaLM-2 (Anil et al., 2023). The ideal and correct output is shown in Figure 1 (d), where the output entity corresponds to the correct CUI.

In summary, adapting LLMs to BioEL presents two main challenges. First, LLMs are unfamiliar with pre-defined biomedical entities, and different BioEL tasks often utilize different medical KGs, necessitating rapid adaptation to various entity sets. The long-tailed distribution in BioEL datasets can hinder LLMs' generalization (Lin et al., 2024b), and the techniques like fine-tuning and in-context learning may not be suitable for BioEL. Second, the large scale and ambiguity of biomedical entities make it challenging for LLMs to accurately link to the accurate entities without injecting domain-specific knowledge (Xie et al., 2024).

To address these challenges, we propose a concise yet effective framework, LLM4BioEL, designed to guide LLMs for BioEL through two distinct decoding strategies. The first strategy, **restrictive decoding**, constrains the logit distribution to relevant tokens associated with the predefined entities. This approach effectively prevents LLMs from generating non-entity tokens and thus ensures valid responses in the context of BioEL. The second strategy, **contrastive decoding**, leverages the inherent knowledge embedded within LLMs and the external knowledge obtained from a trained retriever, which captures the semantic relationships between mentions and entities. Contrastive decoding has been validated to enhance LLMs' truthfulness and factuality (Chuang et al., 2024), and in this work, we employ entropy to derive the contrasted predictions. When LLM token distribution is relatively uniform (high entropy), the retriever's knowledge is prioritized; when it is less uniform (low entropy), LLM4BioEL utilizes the inherent knowledge of LLMs. This dynamic dual approach facilitates adaptive knowledge injection during the decoding process, allowing for external knowledge-aware outputs. To enhance performance, we utilize LLMs' in-context learning (ICL) capabilities by organizing few-shot prompts with relevant examples retrieved by the same retriever, improving adaptability and ensuring access to pertinent information. It is worth noting that LLM4BioEL is a decoding-enhanced framework that can be applied to any open-source LLM as it requires no extra fine-tuning or modification to the architecture.

Our contributions are three-fold:

- We present the first attempt to directly adapt large language models (LLMs) for the biomedical entity linking (BioEL) task. The code is available at `https://anonymous.4open.science/r/LLM4BioEL-6D46/`.

- LLM4BioEL introduces a novel combination of restrictive decoding and entropy-based contrastive decoding, ensuring the generation of valid outputs while dynamically incorporating biomedical knowledge. Additionally, we leverage in-context learning (ICL) to enhance the efficacy of LLM4BioEL.

- Our comprehensive experiments reveal the effectiveness and applicability of LLM4BioEL. The top-performing variant of LLM4BioEL surpasses the performance of training-based *discriminative* and *generative* BioEL methods, as well as other LLM-based methods, underscoring its competitiveness.

2 LLM4BioEL

2.1 Preliminary

Given a pre-identified mention m and a biomedical knowledge graph with entity set $\mathcal{E}$, the target of biomedical entity linking (BioEL) is to link m to the correct entity $e \in \mathcal{E}$. To adapt to the question-answering format of LLMs, Jahan et al. (2023) constructed the BioEL datasets that are organized as $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$, where N denotes the number

[1]The outputs of the four LLMs are taken from the released results of Jahan et al. (2023).

[2]`https://www.anthropic.com/index/claude-2`

[3]`https://platform.openai.com/docs/models/gpt-3-5`

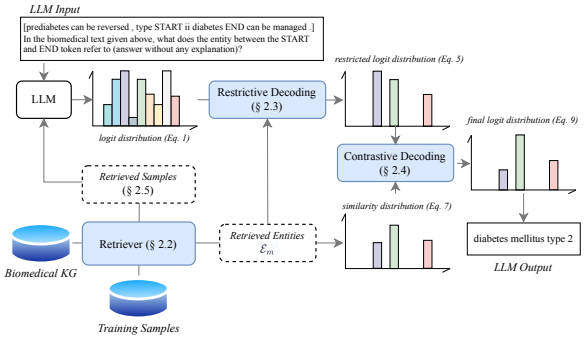


Figure 2: Overall architecture of LLM4BioEL.

of samples in dataset, and x contains the mention m and y refers to the entity e. For simplicity, we will use (x, y) to denote (x_i, y_i) throughout the paper. An example of (x, y) is provided below.

> LLM input x: *[START zirconia END is metal so if you go this route please don't do it just to avoid metal because you are getting it either way.] \n \n In the biomedical text given above, what does the entity between the START and END token refer to (answer without any explanation)?*
>
> LLM output y: *Zirconium dioxide.*

We denote an LLM as $P_\theta(y^1, y^2, .., y^t|x)$ where θ denotes the LLM's parameters and y^t denotes the tth token to be generated. For input x, the greedy decoding process can be denoted as follows:

$$y* = \arg\max P_\theta(y \mid x). \tag{1}$$

The overall architecture of LLM4BioEL is shown as Figure 2 and the following subsections will detail each component.

2.2 Retriever

LLM4BioEL introduces a semantic retriever to obtain an entity subset $\mathcal{E}_m$ given mention m to enable restrictive decoding and also calculate semantic similarities $sim(m, e)$ given m and e to enable contrastive decoding. Following the previous work (Xu et al., 2023; Lin et al., 2024a), we employ a bi-encoder based on SapBERT (Liu et al., 2021) to generate dense vectors for both mentions and entities. The mention embedding $f(m)$ of m is denoted as:

$$f(m) = \texttt{SapBERT}(m)\texttt{[CLS]}, \tag{2}$$

where `[CLS]` denotes the special token used to derive a fixed-size vector. The entity embedding $f(e)$ of e is computed similarly. The score of a mention-entity pair (m, e) is denoted as follows:

$$sim(m, e) = g(f(m), f(e)), \tag{3}$$

where g is the cosine similarity and is utilized for external knowledge injection during contrastive decoding (see § 2.4). During inference, we pre-calculate $f(e)$ for each $e \in \mathcal{E}$, select top-k entities for each mention, $k = |\mathcal{E}_m|$, and use FAISS (Johnson et al., 2019) for fast retrieval. Similarly, we use the same retriever to calculate similarities of mention-mention pairs, $sim(m_i, m_j) = g(f(m_i), f(m_j))$, for constructing in-context prompts (see § 2.5).

We leverage contrastive learning to train the retriever, which aims at optimizing the agreement between true mention-entity pairs and the disagreement between false ones. The loss for each true pair (m, e) is computed as:

$$\mathcal{L}(m, e) = -\log\left(\frac{\delta(m, e)}{\delta(m, e) + \sum\limits_{e' \in \mathcal{H}(e)} \delta(m, e')}\right), \tag{4}$$

where $\delta(m, e) = \exp(sim(m, e)/\tau)$, τ is a temperature hyper-parameter, and $\mathcal{H}(e) \subset \mathcal{E} \backslash \{e\}$ is a set of negatives that excludes e. We obtain $\mathcal{H}(e)$ by combining in-batch negative sampling and hard negative sampling (*i.e.*, highest-scoring incorrect entities), which has been shown beneficial for entity retrieval (Wu et al., 2020; Gao et al., 2021).

2.3 Restrictive Decoding

Restrictive decoding aims at guiding LLMs to output valid entities with the given biomedical KG, as shown in Fig 2. Typically, restrictive decoding, also named constrained decoding, modifies the original decoding process to ensure the output adheres to specific constraints (De Cao et al., 2022; Beurer-Kellner et al., 2024; Park et al., 2024). In the context of BioEL, LLMs are required to directly generate pre-defined biomedical entities, which is consistent with the concept of restrictive decoding.

Given full set $\mathcal{E}$ and mention m, we leverage the retriever (see § 2.2) to obtain an entity subset $\mathcal{E}_m = \{e_1, e_2, \ldots, e_k\}$ where k is the number of retrieved entities $k = |\mathcal{E}_m|$. We define the tokenization process as $\Omega(\cdot)$. For an entity e_i, we obtain its token list as $\Omega(e_i) = [s_i^1, s_i^2, \ldots, s_i^{q_i}]$, where q_i denotes the number of tokens. The token list is then padded with $l - q_i$ "end-of-text" tokens to reach a fixed dimension of l. We process all the retrieved entities to obtain the matrix of tokens $Y_m \in \mathbb{R}^{l \times k}$ where each column refers to an entity and each element Y_m^{ij} refers to s_j^i. As restrictive decoding becomes effective starting from $t = 1$, each row of

Y_m represents the candidate tokens for restrictive decoding when generating the tth token, denoted as $Y_m^t \in \mathbb{R}^{1\times k}$. Note that each token may correspond to multiple entities; for instance, "diabetes" is the 1st token for entities "diabetes mellitus type 1" and "diabetes mellitus type 2". For timestep t in the decoding process, we constrain the logit distribution with the tokens Y_m^t obtained from the retrieved entities, and we rewrite Eq 1 as follows:

$$y^{*t} = \arg\max(P_\theta(y \mid \mathbf{y}^{<t}, x) \cdot \delta^t), \quad (5)$$

where $\mathbf{y}^{<t}$ represents the generated tokens before t, i.e., $\mathbf{y}^{<t} = (y^1, y^2, \ldots, y^{t-1})$ and δ^t determines whether to filter the token logit:

$$\delta^t = \begin{cases} 1.0 & \text{if } y \in Y_m^t; \\ 0.0 & \text{otherwise.} \end{cases} \quad (6)$$

With restrictive decoding, LLMs are guided to output a limited set of tokens for each decoding step, increasing the likelihood of valid responses and mitigating the issue of invalid entity outputs.

2.4 Contrastive Decoding

Based on the restricted logit distributions, LLM4BioEL introduces **contrastive decoding** to contrast the inherent knowledge within LLMs and external knowledge brought by the retriever (see § 2.2), thereby enhancing entity disambiguation abilities. At timpstep t, we denote the logit distribution produced by LLMs as $P_\theta(y \mid \mathbf{y}^{<t}, x)$. the logit distribution obtained by the retriever as $P(y|m, \mathcal{E}_m)$, which represents the distribution of semantic similarities among entity token y given mention m. For simplicity, we take the entity-level semantic similarity as the token-level similarity, and for an entity e and its token $y \in \Omega(e)$, $P(y|m, \mathcal{E}_m)$ is computed as follows:

$$P(y|m, \mathcal{E}_m) \propto \frac{sim(m, e)}{\sum_{e' \in \mathcal{E}_m} sim(m, e')}, \quad (7)$$

where $sim(m, e)$ denotes the similarity computation between mention m and entity e (taken from Eq 3). Token y may correspond to multiple entities and the maximized similarity value among these entities is taken. Higher similarity values suggest increased probabilities of the correct entities, which can be regarded as external knowledge brought by the retriever to guide the LLMs' decoding process.

However, LLMs may exhibit inherent knowledge regarding some biomedical concepts, which can be correctly linked without the use of external knowledge. We thus utilize the entropy of the logit distribution $H(P_\theta(y \mid \mathbf{y}^{<t}, x))$ to express LLMs' uncertainty under the given question (Kuhn et al., 2023; Kim et al., 2024b). Intuitively, when LLMs are uncertain about some biomedical concepts, the entropy $H(P_\theta(y \mid \mathbf{y}^{<t}, x))$ tends to be higher, indicating that external knowledge should be prioritized to assist LLMs rather than relying on internal knowledge. Conversely, a lower entropy suggests LLMs are more confident in the predictions, allowing LLMs to utilize their inherent knowledge to answer the questions. Therefore, we design an entropy-based parameter to balance the logit distributions of LLMs and the retriever. Since the ranges of logit distributions differ greatly, the distributions are normalized before calculating the entropies:

$$\alpha^t = \frac{H(P_\theta(y \mid \mathbf{y}^{<t}, x))}{H(P_\theta(y \mid \mathbf{y}^{<t}, x)) + H(P(y|m, \mathcal{E}_m))}. \quad (8)$$

Guided by α^t, LLM4BioEL enables an adaptive adjustment in the extent to which LLMs leverage external knowledge for prediction. Thus, we can reformulate Eq 5 to derive the contrastive prediction as follows:

$$y^{*t} = \arg\max \big((1 - \alpha^t) \cdot P_\theta(y \mid \mathbf{y}^{<t}, x) \cdot \delta^t + \alpha^t \cdot P(y \mid m, \mathcal{E}_m)\big). \quad (9)$$

Through contrastive decoding, LLM4BioEL adaptively guides the answer generation by contrasting and injecting external knowledge brought by the retriever.

2.5 In-context Learning (ICL)

Semantically similar samples can serve as informative inputs to LLMs and some previous studies proposed to retrieve similar samples to construct better few-shot prompts (Rubin et al., 2022; Liu et al., 2022). Inspired by these findings, we leverage ICL capabilities to improve LLM4BioEL further. Formally, some training samples are taken from $\mathcal{B}$ and linearized to incorporate into the input x, which formulates the output $y*$ as follows:

$$y* = \arg\max P_\theta(y \mid \underbrace{x_1, y_1, \ldots, x_n, y_n}_{context}, x), \quad (10)$$

where each pair (x_j, y_j) is selected from $\mathcal{B}$ and n denotes the number of samples. The selec-

tion requires retrieving top-n similar mentions using $sim(m_i, m_j)$ where mention m_i is within input x and m_j is within input x_j. We use the same retriever (see § 2.2) to create a datastore with key-value pairs $(f(m_i), e_i)$ for the ith instance (m_i, e_i), where $f(m_i)$ is the mention embedding from Eq 2. Consequently, we replace x in Eq 9 with $(x_1, y_1, \ldots, x_n, y_n, x)$ to incorporate ICL. Notably, if retrieved samples provide informative clues, LLM4BioEL will prioritize inherent knowledge, leading to a lower value of α in Eq 9.

3 Experiments

3.1 Experimental Setup

Datasets. We adopt 3 BioEL datasets for evaluation, including NCBI (Doğan et al., 2014), BC5CDR (Li et al., 2016), and COMETA (Basaldella et al., 2020). Please refer to Appendix § A for more details.

Metrics. We report Hits@1 along with the newly designed Hits@KG for evaluation. The metric Hits@KG aims to compute the ratio of valid generated entities in the biomedical KG, and higher Hits@KG indicates that LLM's output is more in line with the biomedical KG. Different from Jahan et al. (2023), we report metrics of directly retrieving CUIs of the LLM's output as LLM has been required to "answer without any explanation". The metrics are thus re-calculated for the LLMs taken from Jahan et al. (2023). For the details of our evaluation method and the difference from that used in Jahan et al. (2023), please refer to Appendix § E. We also report the averaged decoding throughput (Token/s).

LLM backbones. To show the applicability of LLM4BioEL, we adopt four LLMs of different scales in our experiments: 1) Qwen-2-1.5b-instruct (Yang et al., 2024a), a relatively smaller LLM; 2) Mistral-7b-v0.3-instruct (Jiang et al., 2023), a widely-used general-domain LLM; 3) Llama-3-8b-instruct (AI@Meta, 2024) and 4) Llama-3-70b-instruct (AI@Meta, 2024), another widely-used general-domain LLMs. Please refer to Appendix § B for more details.

Baseline settings. We compare LLM4BioEL with three baseline settings: 1) greedy decoding, 2) DoLa (Chuang et al., 2024), a contrastive decoding strategy to improve the factuality of LLMs, and 3) ICL with 10-shot prompts, where the prompts are built with randomly shuffled training instances. The DoLa baseline contrasts "high" layers to enhance short-answer tasks (Chuang et al., 2024).[4]

Implementation details. The hyper-parameters of LLM4BioEL include the number of retrieved entities k and the number of few-shot prompts n. We apply the grid search strategy on the evaluation split for best-performing k out of [1,10] and n out of [0,10]. Note that we search n out of [10,80] for the COMETA dataset due to greater task difficulty and less contextual information in the question. All the experiments are done with greedy search as the decoding strategy.

3.2 Main Results

We report the experimental results of LLM4BioEL in Table 1. In greedy decoding, Llama-3-8b and Llama-3-70b noticeably outperform Qwen-2-1.5b and Mistral-7b-v0.3 with Llama-3-70b achieving comparable or superior performance against the three *LLM-based* BioEL methods. However, Llama-3-70b lacks domain-specific knowledge, resulting in invalid entities and Hits@KG of approximately 60.3%, 64.8%, and 41.5% on NCBI, BC5CDR, and COMETA, respectively. The DoLa (Chuang et al., 2024) decoding strategy shows no performance enhancements for LLMs on BioEL. Conversely, ICL with 10-shot prompts leads to performance drops for Qwen-2-1.5b and Llama-3-8b, while Llama-3-70b shows an average performance gain of 28.4% over greedy decoding, attributed to its larger parameter scale.

LLM4BioEL demonstrates substantial improvements from greedy decoding across different datasets using different LLM backbones. For example, with LLM4BioEL, Llama-3-8b improves the averaged Hits@1 from 36.3% to 89.9% and Mistral-7b-v0.3 improves from 2.0% to 89.8%. LLM4BioEL effectively enables LLMs to produce valid entities within the biomedical KG, achieving 100.0% Hits@KG across different backbones. While different LLMs perform differently in greedy decoding, they all can reach relatively comparative performance using LLM4BioEL, highlighting its wide applicability. As for decoding throughput, Qwen-2-1.5b is the fastest, followed by Mistral-7b-v0.3 and Llama-3-8b, while Llama-3-70b is the slowest in greedy decoding. Although DoLa impacts the throughput with negligible cost for Qwen-2-1.5b and Mistral-7b-v0.3, it significantly increases the decoding time for Llama-3-8b.

[4] We encountered a GPU out-of-memory issue with Llama-3-70b when using DoLa strategy, and therefore, we do not report this part of results.

Table 1: Experimental results of LLM4BioEL using different LLM backbones.

LLMs	NCBI		BC5CDR		COMETA		Avg.	
	Hits@1	Hits@KG	Hits@1	Hits@KG	Hits@1	Hits@KG	Hits@1	Token/s
Qwen-2-1.5b	20.0	23.7	13.4	17.9	3.1	4.9	12.2	13.0 (× 1.00)
+ DoLa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.4 (× 0.80)
+ ICL (10-shot)	0.0	0.0	0.0	0.0	0.0	0.1	0.0	12.4 (× 0.95)
+ LLM4BioEL	92.4	100.0	92.2	100.0	82.7	100.0	89.1	4.8 (× 0.37)
Mistral-7b-v0.3	0.9	0.9	3.9	4.1	1.0	1.2	2.0	10.2 (× 1.00)
+ DoLa	0.0	0.0	0.5	0.5	0.2	0.2	0.2	10.3 (× 1.01)
+ ICL (10-shot)	0.9	0.9	3.9	4.1	28.3	32.7	11.1	9.1 (× 0.89)
+ LLM4BioEL	92.9	100.0	92.4	100.0	84.1	100.0	89.8	4.2 (× 0.41)
Llama-3-8b	36.3	39.5	49.8	56.4	22.9	35.5	36.3	5.4 (× 1.00)
+ DoLa	27.7	30.4	28.2	32.4	7.3	10.0	21.1	0.1 (× 0.02)
+ ICL (10-shot)	12.8	14.7	26.9	28.8	42.1	52.4	27.3	3.6 (× 0.67)
+ LLM4BioEL	93.2	100.0	92.2	100.0	84.4	100.0	89.9	3.3 (× 0.61)
Llama-3-70b	57.3	60.3	61.1	64.8	34.9	41.5	51.1	1.1 (× 1.00)
+ ICL (10-shot)	77.8	79.1	72.7	76.0	46.3	56.7	65.6	0.6 (× 0.55)
+ LLM4BioEL	93.8	100.0	92.4	100.0	84.8	100.0	90.3	1.0 (× 0.91)

Applying in-context learning (ICL) with 10-shot prompts leads to a modest decrease in throughput, with reduction factors ranging from 0.55 to 0.95. LLM4BioEL exhibits varying effects on decoding efficiency, resulting in reduction factors of 63%, 59%, 39%, and 9%, indicating that larger language models (LLMs) tend to be less affected. These differences in the impact of LLM4BioEL across various LLMs can be attributed to variations in model architecture and optimization strategies.

3.3 Comparison with State-of-the-art Methods

We experimentally compare LLM4BioEL against the state-of-the-art (SOTA) BioEL methods using the His@1 metric on three datasets. The BioEL baselines include a) 8 *discriminative* methods, including BioSyn (Sung et al., 2020), ResCNN (Lai et al., 2021), SapBERT (Liu et al., 2021), Cross-domain (Varma et al., 2021), Clustering-based (Angell et al., 2021), Prompt-BioEL (Xu et al., 2023), BioFEG (Sui et al., 2023) and BioPro (Zhu et al., 2023), b) 7 *generative* methods, including GenBioEL (Yuan et al., 2022b), BART-base/large (Yuan et al., 2022a), BioBART-base/large (Yuan et al., 2022a), GenBioEL+ANGEL (Kim et al., 2024a), and BioBART+ANGEL (Kim et al., 2024a), and c) 8 *LLM-based* methods, including GPT-3.5/PaLM-2/Claude-2 (Jahan et al., 2023), GPT-4 (Achiam et al., 2023), DeepSeek-R1-distilled LLMs,[5] and DeepSeek-R1 (DeepSeek-AI et al., 2025).

As shown in Table 2, LLM4BioEL demonstrates comparative performance to SOTA baselines across different datasets. Overall, LLM4BioEL (Llama-3-70b) shows the best performance, and LLM4BioEL (Llama-3-8b) achieves comparable performance to the second-best Prompt-BioEL (Xu et al., 2023), among which our lower BC5CDR performance likely stems from competitors' UMLS pretraining (aligned with BC5CDR's medical subset) and our general-domain LLM's reduced specificity on specialized terminology. Regarding Qwen-2-1.5b and Mistral-7b-v0.3, both demonstrate performance comparable to other SOTA baselines, showing the significant improvements brought by LLM4BioEL. LLM4BioEL shows the flexibility to switch between different biomedical KGs and LLM backbones, as evidenced by the comparative performance of its different variants. Furthermore, LLM4BioEL is a robust framework that offers enhanced generalization across diverse tasks when compared to other baselines. It is important to note that directly applying LLMs into BioEL without proper guidance leads to unsatisfactory performance. LLMs like GPT-3.5, PaLM-2, Claude-2, and GPT-4 all underperform *discriminative* and *generative* baselines by a large margin. Similar findings can be observed with the DeepSeek-R1 models, which demonstrate various but limited abilities to reason about biomedical entities accurately.

3.4 Ablation Studies

This subsection investigates different components in LLM4BioEL. We experiment LLM4BioEL with

[5] including DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-R1-1.5b), DeepSeek-R1-Distill-Qwen-7B (DeepSeek-R1-7b), and DeepSeek-R1-Distill-Llama-8B (DeepSeek-R1-8b).

Table 2: Comparison of LLM4BioEL against *discriminative*, *generative*, and *LLM-based* BioEL baselines on three BioEL datasets, where the best performance is **in bold** and the second best is underlined. The symbol * denotes the re-calculated metrics and † denotes the reproduced results.

Methods	NCBI	BC5CDR	COMETA	Avg.
Discriminative BioEL methods				
BioSyn	91.1	86.3†	71.3	82.9
ResCNN	92.4	88.8†	80.1	87.1
SapBERT	92.3	89.7†	75.1	85.7
Cross-domain	–	89.3	–	–
Clustering-based	–	91.3	–	–
Prompt-BioEL	92.6	93.7	83.7	90.0
BioFEG	–	93.4	–	–
BioPro	94.5	–	–	–
Generative BioEL methods				
GenBioEL	91.9	93.3	81.4	88.9
BART-base	88.5	91.6	78.3	86.1
BART-large	90.2	92.5	80.7	87.8
BioBART-base	89.3	93.0	79.6	87.3
BioBART-large	89.9	93.3	81.8	88.3
GenBioEL+ANGEL	92.5	94.4	82.4	89.8
BioBART+ANGEL	91.9	94.7	82.2	89.6
LLM-based BioEL methods				
GPT-3.5	–	–	27.3*	–
PaLM-2	–	–	29.5*	–
Claude-2	–	–	37.2*	–
GPT-4	59.4	66.3	40.3	55.3
DeepSeek-R1-1.5b	2.3	2.3	2.0	2.2
DeepSeek-R1-7b	9.0	7.9	4.9	7.3
DeepSeek-R1-8b	2.0	1.9	2.9	2.3
DeepSeek-R1	62.6	71.4	37.6	57.2
LLM4BioEL (Ours)				
Qwen-2-1.5b	92.4	92.2	82.7	89.1
Mistral-7b-v0.3	92.9	92.4	84.1	89.8
Llama-3-8b	93.2	92.2	84.4	89.9
Llama-3-70b	93.8	92.4	84.8	**90.3**

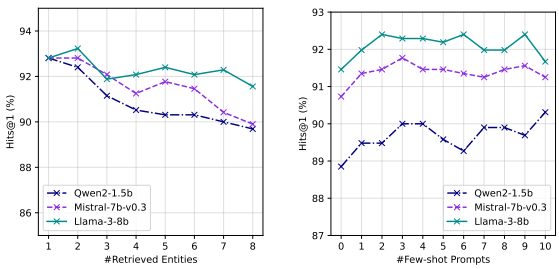
Llama-3-8b as LLM backbone by 1) removing contrastive decoding (*w/o contrastive*), 2) replacing fixed-value $\alpha = 0.5$ in Eq 9 (*w/ fixed* α), 3) removing few-shot prompting (*w/o few-shot*), and 4) replacing few-shot examples with randomly selected examples (*w/ random prompts*). As listed in Table 3, we observe performance degradation of removing contrastive decoding and using the fixed value of α in the contrasting process. In addition, we observe that few-shot prompting has a greater impact on COMETA (9.4% drop) than on NCBI (1.8% drop) or BC5CDR (1.5% drop), suggesting that LLM4BioEL relies more on few-shot prompting to address limited contextual information provided by the original prompt in COMETA. The comparison between LLM4BioEL and LLM4BioEL *w/ random prompts* further confirms the effectiveness of the prompting design proposed in § 2.5. For instance, when using random prompts instead of retrieval-based prompts, Hits@1 drops by 1.2%, 0.2%, and 3.3% for NCBI, BC5CDR, and COMETA, respectively.

Table 3: Ablation study of LLM4BioEL (Hits@1).

Settings	NCBI	BC5CDR	COMETA
LLM4BioEL	93.2	92.2	84.4
w/o contrastive	91.8	91.9	83.6
w/ fixed α	91.9	92.2	84.1
w/o few-shot	91.5	90.8	76.3
w/ random prompts	92.1	92.0	81.4

3.5 Hyper-parameter Sensitivity Analysis

This subsection discusses hyper-parameters sensitivity on the NCBI dataset, including the number of retrieved entities k and the number of few-shot prompts n. We report the Hits@1 metrics for three LLM backbones, with varying values of k while keeping n fixed, and vice versa. As shown in Figure 3 (a), Hits@1 is observed to reach its highest values mostly at $k = 2$, and as k increases, Hits@1 generally shows a decreasing trend. This suggests that retrieving more entities introduces additional noise, making entity disambiguation more challenging and affecting the contrastive process. For hyperparameter n, Figure 3 (b) shows significant increases of Hits@1, when comparing zero-shot ($n = 0$) and 10-shot ($n = 10$) prompting. This shows clear evidence of the retrieval-based few-shot prompts. While n continues to rise, the performance gain is less obvious than from $n = 0$ to $n = 10$. Overall, LLM4BioEL reaches the highest Hits@1 when $n = 3/3/2$ on NCBI for three backbones. Please refer to Appendix § C for the experiments on BC5CDR and COMETA datasets.



(a) Impact of k on NCBI. (b) Impact of n on NCBI.

Figure 3: Hyper-parameter sensitivity experiments.

3.6 Comparison of Different Retrievers

The retriever assists in various components of LLM4BioEL, including restricting logit distribution (§ 2.3), contrasting logit distribution (§ 2.4), and building few-shot prompts (§ 2.5). To assess the impacts of different retrievers, we compare our retriever with other models: SapBERT (Liu et al., 2021), a specialized BERT variant for biomedical text processing, and SimCSE (Gao et al., 2021), a general-domain sentence embedding model. As shown in Table 4, we find performance improvements in LLM4BioEL when applying more capable retrievers. Specifically, the retriever (§ 2.2) performs the best, followed by SapBERT and SimCSE, and the performance drops are even more pronounced on the COMETA dataset. This underscores the importance of a capable retriever for the performance of LLM4BioEL.

Table 4: Comparison of different retrievers (Hits@1).

Settings	NCBI	BC5CDR	COMETA
LLM4BioEL			
w/ Retriever (§ 2.2)	93.2	92.2	84.4
w/ SapBERT	92.2	88.7	68.3
w/ SimCSE	88.9	83.1	56.5

4 Related Work

4.1 Biomedical Entity Linking

The current BioEL methods can be broadly categorized into *discriminative* methods and *generative* methods (Shi et al., 2023). ***Discriminative* BioEL methods** focus on training bi-encoders or cross-encoders to enhance the retrieval of relevant biomedical entities. For instance, Liu et al. (2021) introduced a self-alignment pretraining strategy to refine biomedical entity representations. Lai et al. (2021) presented a lightweight yet effective CNN, demonstrating that complex models are not always necessary. Furthermore, some studies (Xu et al., 2020; Angell et al., 2021; Xu et al., 2023) have explored using cross-encoders to capture subtle mention-entity relationships. Besides, some studies (Lin et al., 2024a,b) proposed using similar instance references during training or prediction to address the long-tailed distribution issue. ***Generative* BioEL methods** bypass retrieval by directly generating linked entities. For instance, Yuan et al. (2022b) enhanced generation with knowledge base pre-training and synonym-aware fine-tuning, while Yuan et al. (2022a) developed BioBART for strong biomedical NLG benchmarks. Kim et al. (2024a) further improved generative models by incorporating negative samples, enhancing their ability to distinguish similar entities. Recent studies (Wang et al., 2023b; Xie et al., 2024) have investigated the potential of large language models for in-context learning in biomedical concept linking via prompting, yet are limited by retrieval-dependent candidate selection. Our LLM4BioEL framework addresses this by enabling direct entity generation from biomedical KGs, better suited for clinical applications where predefined candidate sets may not be available.

4.2 Biomedical Large Language Models

In recent years, large language models (LLMs) like ChatGPT, PaLM-2, Claude-2, and Llama have shown promising potential in biomedical tasks (Jahan et al., 2023; Liu et al., 2024). To enhance their domain-specific capabilities, several methods have been developed, including extra pre-training with biomedical data. Luo et al. (2022) pre-trained a model on biomedical literature, while Wang et al. (2023a) fine-tuned LLMs with diverse medical data for clinical tasks. Additionally, Christophe et al. (2024) fine-tuned Llama-3 models with medical instruction data and implemented multi-stage preference alignment. Although these models have shown strong performance in various biomedical tasks, their application in BioEL remains underexplored. This work introduces a universal framework tailored for BioEL to enhance the performance of open-source LLMs.

5 Conclusion

In this paper, we introduced LLM4BioEL, a straightforward yet effective framework for BioEL. LLM4BioEL consists of restrictive decoding to ensure valid output, entropy-based contrastive decoding to adaptively integrate external knowledge, and few-shot prompting to enhance the linking performance. We experimented with four different LLMs and validated the performance improvement of LLM4BioEL compared to other baselines. We also demonstrated its comparative performance to other state-of-the-art BioEL baselines, and its flexibility to switch between different tasks and backbones without fine-tuning. For future exploration, we plan to investigate LLM4BioEL in the general-domain EL and investigate the combination of decoding strategies for other knowledge-intensive tasks.

Limitations

This section discusses the limitations of our work. First, LLM4BioEL requires preprocessing steps for entity retrieval and tokenization, which introduce additional computational costs to the decoding process of the LLMs. Second, the decoding-enhanced nature of LLM4BioEL limits its applicability to closed-source models, such as ChatGPT. Lastly, the performance of LLM4BioEL is influenced by the effectiveness of the retriever (as discussed in § 2.2). In the future, we aim to explore advanced retrieval techniques to reduce this dependency, such as hybrid retrieval combined with rule-based filtering that applies semantic type constraints, or synonym expansion.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 Model Card.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based Inference for Biomedical Entity Linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A Corpus for Medical Entity Linking in the Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3122–3137.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation. *arXiv preprint arXiv:2403.06988*.

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

Clément Christophe, Tathagata Raha, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Nada Saadi, Hamza Javed, Umar Salman, Svetlana Maslenkova, Marco Pimentel, Ronnie Rajan, and Shadab Khan. 2024. Med42-v2 - A Suite of Clinically-aligned Large Language Models.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of Biomedical Informatics*, 47.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks. *arXiv preprint arXiv:2310.04270*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3).

Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. 2024a. Learning from Negative Samples in Generative Biomedical Entity Linking. *Preprint*, arXiv:2408.16493.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang goo Lee, and Taeuk Kim. 2024b. Adaptive Contrastive Decoding in Retrieval-Augmented Generation for Handling Noisy Contexts. *Preprint*, arXiv:2408.01084.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *Preprint*, arXiv:2302.09664.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional

Neural Networks. In *Findings of the Association for Computational Linguistics: EMNLP*.

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

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024a. Biomedical Entity Linking as Multiple Choice Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2390–2396. ELRA and ICCL.

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024b. Improving Biomedical Entity Linking with Retrieval-Enhanced Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11461–11465.

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*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114. Association for Computational Linguistics.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking Large Language Models on CMExam – A Comprehensive Chinese Medical Exam Dataset. *Advances in Neural Information Processing Systems*, 36.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. 2024. Grammar-Aligned Decoding. *arXiv preprint arXiv:2405.21047*.

Zhi Qiao, Zhen Zhang, Xian Wu, Shen Ge, and Wei Fan. 2020. MHM: Multi-modal Clinical Data based Hierarchical Multi-label Diagnosis Prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1841–1844. Association for Computing Machinery.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. 2023. Knowledge-graph-enabled Biomedical Entity Linking: A Survey. *World Wide Web*, 26(5):2593–2622.

Xiaoming Shi, Sendong Zhao, Yuxuan Wang, Xi Chen, Ziheng Zhang, Yefeng Zheng, and Wanxiang Che. 2022. Understanding Patient Query With Weak Supervision From Doctor Response. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2770–2777.

Xuhui Sui, Ying Zhang, Xiangrui Cai, Kehui Song, Baohang Zhou, Xiaojie Yuan, and Wensheng Zhang. 2023. BioFEG: Generate Latent Features for Biomedical Entity Linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11584–11593, Singapore. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023b. Exploring the In-context Learning Ability of Large Language Model for Biomedical Concept Linking. *Preprint*, arXiv:2307.01137.

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*, pages 6397–6407.

Yuzhang Xie, Jiaying Lu, Joyce Ho, Fadi Nahab, Xiao Hu, and Carl Yang. 2024. PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2589–2593. Association for Computing Machinery.

Dongfang Xu, Zeyu Zhang, and Steven Bethard. 2020. A Generate-and-Rank Framework with Semantic Type Regularization for Biomedical Concept Normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving Biomedical Entity Linking with Cross-Entity Interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, et al. 2024a. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024b. KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques. *Preprint*, arXiv:2403.05881.

Sheng Yu, Zheng Yuan, Jun Xia, Shengxuan Luo, Huaiyuan Ying, Sihang Zeng, Jingyi Ren, Hongyi Yuan, Zhengyun Zhao, Yucong Lin, Keming Lu, Jing Wang, Yutao Xie, and Heung-Yeung Shum. 2022. BIOS: An Algorithmically Generated Biomedical Knowledge Graph. *Preprint*, arXiv:2203.09975.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. BioBART: Pretraining and Evaluation of a Biomedical Generative Language Model. *arXiv preprint arXiv:2204.03905*.

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*.

Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, YueJia Xiang, and Yefeng Zheng. 2020. An Industry Evaluation of Embedding-based Entity Alignment. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 179–189. International Committee on Computational Linguistics.

Yutian Zhao, Huimin Wang, Xian Wu, and Yefeng Zheng. 2024. MKeCL: Medical Knowledge-Enhanced Contrastive Learning for Few-shot Disease Diagnosis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 11394–11404. ELRA and ICCL.

Tiantian Zhu, Yang Qin, Ming Feng, Qingcai Chen, Baotian Hu, and Yang Xiang. 2023. BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:374–385.

A Datasets

Table 5 represents the dataset statistics and task prompt used in the experiments.

Table 5: Dataset statistics and task prompt.

Datasets	Entity Types	Entities $\lvert\mathcal{E}\rvert$	Dataset Split Train / Valid / Test
NCBI	Disease	14,967	5,784 / 787 / 960
BC5CDR	Disease & Chemical	268,162	9,285 / 9,515 / 9,654
COMETA	Clinical Terms	350,830	13,489 / 2,176 / 4,350

Task Prompt

[TEXT_S <START> ENTITY <END> TEXT_E]
In the biomedical text given above, what does the entity between the START and END token refer to (answer without any explanation)?

B LLM Descriptions

This section provides descriptions of the 5 instruction-tuned LLMs used in the evaluation:

- **Qwen-2-1.5b** (Yang et al., 2024a): As part of the Qwen series of models, Qwen-2-1.5b is an instruction-tuned language model with 1.5 billion parameters. Qwen-2-1.5b is designed for easy deployment and quick application, and the model checkpoints can be accessed via `https://huggingface.co/Qwen/Qwen2-1.5B-Instruct`.

- **Mistral-7b-v0.3** (Jiang et al., 2023): Mistral-7b-v0.3, developed by Mistral AI, is a large language model with 7 billion parameters. It can follow instructions, complete requests, and generate creative content. The model checkpoints can be accessed via `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`.

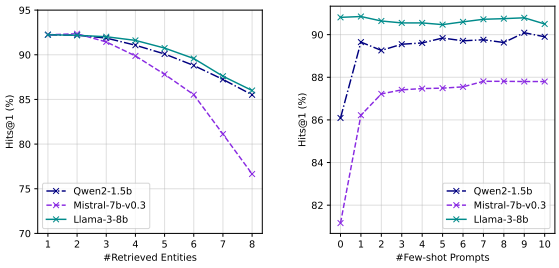
- **Llama-3-8b/70b** (AI@Meta, 2024): Llama-3 family of large language models is a collection of pre-trained and instruction-tuned generative text models in 8B and 70B sizes. The instruction-tuned Llama-3 models are optimized for dialogue use cases. The model checkpoints can be accessed via `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct` and `https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct`.

- **Llama-3-Med42-8b** (Christophe et al., 2024): Med42-v2 is a suite of open-access clinical large language models instruction- and

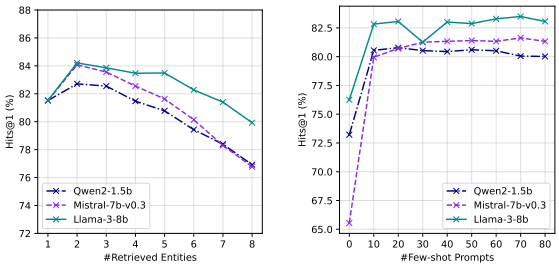
preference-tuned by M42 to expand access to medical knowledge and to provide high-quality answers to medical questions. Llama-3-Med-8b is built upon Llama-3-8b with 8 billion parameters. The model checkpoints can be accessed via https://huggingface.co/m42-health/Llama3-Med42-8B.

C Hyper-parameter Sensitivity Analysis (Cont')

The experiments conducted on the BC5CDR and COMETA datasets yield conclusions consistent with § 3.5. LLM4BioEL reaches the highest Hits@1 $n = 9/7/1$ on BC5CDR, and $n = 20/70/70$ on COMETA, for three backbones, respectively.



(a) Impact of k on BC5CDR. (b) Impact of n on BC5CDR.



(c) Impact of k on COMETA. (d) Impact of n on COMETA.

Figure 4: Hyper-parameter sensitivity experiments.

C.1 Comparison of General-domain LLM and Domain-specific LLM

Another method for adapting LLMs to BioEL is domain-specific fine-tuning. This subsection investigates the suitability of LLM4BioEL for fine-tuned LLMs and compares general-domain LLMs with domain-specific ones. We use Llama-3-Med42-8b[6] as the domain-specific LLM, as it is instruction-tuned from Llama-3-8b using domain-specific corpora, including medical flashcards and exam questions (Christophe et al., 2024). Following the setup in § 3.1, we report the performance of greedy decoding and LLM4BioEL settings. Table 6 shows the Hits@1 (H@1) and Hits@KG (H@$\mathcal{E}$) metrics. Surprisingly, Llama-3-Med42-8b underperforms compared to Llama-3-8b in greedy decoding, likely due to the instruction gap between its fine-tuning data and BioEL data, as it focuses on dialogue-oriented medical scenarios rather than task-oriented ones. In contrast, LLM4BioEL significantly enhances the performance of both Llama-3-8b and Llama-3-Med42-8b, achieving average Hits@1 of 89.9% and 89.4%, respectively. Specifically, LLM4BioEL (Llama-3-8b) excels on NCBI and COMETA, while LLM4BioEL (Llama-3-Med42-8b) shows comparable performance to LLM4BioEL (Llama-3-8b) on BC5CDR. These results indicate that LLM4BioEL is effective for domain-specific LLMs, even when their original fine-tuning objectives are not directly related to BioEL.

Table 6: Comparison of general-domain Llama-3-8b and domain-specific Llama-3-Med42-8b.

LLMs	NCBI		BC5CDR		COMETA	
	H@1	H@$\mathcal{E}$	H@1	H@$\mathcal{E}$	H@1	H@$\mathcal{E}$
Llama-3-8b	36.3	39.5	49.8	56.4	22.9	35.5
+ LLM4BioEL	93.2	100.0	92.2	100.0	84.4	100.0
Llama-3-Med42-8b	2.1	2.6	2.4	2.7	10.6	13.5
+ LLM4BioEL	92.3	100.0	92.2	100.0	83.6	100.0

C.2 Investigation of different BioEL Task Formats

Compared to directly generating the correct entities, some previous work attempted to probe LLMs to conduct multiple-choice question answering (MCQA) (Lin et al., 2024a). This subsection investigates the performance comparison of different task formats, namely 1) entity generation and 2) entity selection. As for entity selection (MCQA format), each question is provided by five options recalled by the Retriever (§ 2.2). Similar to the ICL (10-shot) setting, we also provide 10-shot examples to probe LLMs in the MCQA setting. Table 8 reports the results of four LLM backbones. It clearly shows that modeling BioEL as MCQA brings performance gains compared to ICL, but it still underperforms LLM4BioEL. Although the MCQA task seems to simplify it, we find that LLMs exhibit selection bias in MCQA and are susceptible to the influence of option positioning. Besides, LLMs need to learn to associate the symbol with the chosen answer option; otherwise, the output symbol

[6] https://huggingface.co/m42-health/Llama3-Med42-8B

Table 7: Case studies of LLM4BioEL against other baselines using Llama-3-8b as LLM backbone.

ID	LLM Input	Methods	LLM Output	KG CUI
(1)	[blodshot eyes , START wrinkles END , hair loss , hunger attacks , cravings , weight loss , heartburn , caffeine , addiction .] \n \n In the biomedical text given above, what does the entity between the START and END token refer to (answer without any explanation)?	greedy decoding	symptoms	N/A
		DoLa	The entity is "wrinkles".	N/A
		ICL (10-shot)	wrinkles	N/A
		LLM4BioEL	wrinkle	27367001
		ground truth	wrinkle	27367001
(2)	[syncope caused by hyperkalemia during use of a combined therapy with the angiotensin-converting enzyme inhibitor and spironolactone. a 76 year-old woman with a history of coronary artery bypass grafting and prior myocardial infarction was transferred to the emergency room with START loss of consciousness END due to marked bradycardia caused by hyperkalemia. the concentration of serum potassium was high, and normal sinus rhythm was restored after correction of the serum potassium level. the cause of hyperkalemia was considered to be several doses of spiranolactone, an aldosterone antagonist, in addition to the long-term intake of ramipril, an ace inhibitor .....] \n \n In the biomedical text given above, what does the entity between the START and END token refer to (answer without any explanation)?	greedy decoding	Loss of consciousness	D014474
		DoLa	The entity between the START and END tokens refers to "START loss of consciousness" or "loss of consciousness". It describes the acute symptom experienced by the 76-year-old woman in the emergency room.	N/A
		ICL (10-shot)	Here are the answers:\n1. focal glomerulosclerosis\n2. nonsmall cell lung carcinoma\n3. coccidioidomycosis\n4. parameth asone\n5. tamoxifen\n6. potassium\n7. glutathione\n8. warfarin\n9. angiotensin	N/A
		LLM4BioEL	loss of consciousness	D014474
		ground truth	loss of consciousness	D014474

may lack meaningfulness and exhibit a degree of randomness.

Table 8: Comparison of different BioEL task format.

Methods	NCBI	BC5CDR	COMETA	Avg.
Qwen-2-1.5b				
ICL	0.0	0.0	0.0	0.0
ICL + MCQA	48.1	46.0	16.2	36.8
LLM4BioEL	92.4	92.2	82.7	89.1
Mistral-7b-v0.3				
ICL	0.9	3.9	28.3	11.1
ICL + MCQA	85.8	88.8	69.9	81.6
LLM4BioEL	92.9	92.4	84.1	89.8
Llama-3-8b				
ICL	12.8	26.9	42.1	27.3
ICL + MCQA	80.2	77.4	54.2	70.6
LLM4BioEL	93.2	92.2	84.4	89.9
Llama-3-70b				
ICL	77.8	72.7	46.3	65.6
ICL + MCQA	87.0	89.5	59.9	78.8
LLM4BioEL	93.8	92.4	84.8	90.3

D Case Studies

This subsection presents two case studies in Table 7. We list the ground truth with the output of Llama-3-8b using greedy decoding, DoLa, ICL with 10-shot prompting (ICL (10-shot)), and LLM4BioEL, same as § 3.1. We also list the KG CUI, which is directly obtained using the LLM output, to show the validity and correctness of the LLM output. Overall, LLM4BioEL is shown capable of generating valid and correct biomedical entities. In contrast, while applying DoLa appears to improve the actuality of output, it fails to adhere to the instructions, as seen in Case (2). when using ICL with 10-shot prompting, Llama-3-8b generates invalid biomedical entities (Case (1)) and non-entity sequences (Case (2)).

E Evaluation Method for LLM4BioEL

In this work, we utilize a different evaluation method from that used in Jahan et al. (2023) and report the re-calculated metrics in our experiments. We emphasize the importance of conducting an equal comparison between *LLM-based* BioEL methods and *discriminative* and *generative* methods, following the approach used in Xu et al. (2023); Lin et al. (2024b,a); Kim et al. (2024a) to evaluate the *LLM-based* BioEL methods. Specifically, we directly utilize the LLM output to retrieve KG CUI, which returns N/A if the output is not a valid entity. We define Hits@1 as 1 if the corresponding KG CUI is the same as the ground truth and 0 otherwise. In contrast, Jahan et al. (2023) evaluated the performance of LLMs on BioEL using a straightforward method: they define Hits@1 as 1 if a) the LLM output equals the ground truth; b) the LLM output exists within the ground truth; or c) the ground truth exists within the LLM output. This method heavily biases the evaluated performance as it counts more false positives. For instance, if the LLM output is "neoplasm of oesophagus" and the ground truth is "benign neoplasm of oesophagus", the evaluation method in (Jahan et al., 2023) would determine this as correct, which is not accurate.

To illustrate the difference between the two evaluation methods, we report both metrics for all datasets and LLM backbones in Table 9. We observe that the evaluation method in (Jahan et al.,

Table 9: Comparison of LLM4BioEL and other *LLM-based* BioEL baselines using different evaluation methods.

Models	Evaluation method in Jahan et al. (2023)				Evaluation method in this work			
	NCBI	BC5CDR	COMETA	Avg.	NCBI	BC5CDR	COMETA	Avg.
GPT-3.5	52.2	54.9	43.5	50.2	-	-	27.3	-
PaLM-2	38.4	52.1	48.8	46.5	-	-	29.5	-
Claude-2	70.2	78.0	53.3	67.2	-	-	37.2	-
GPT-4	81.0	81.3	55.5	72.6	59.4	66.3	40.3	55.3
DeepSeek-R1-1.5b	20.0	26.2	15.3	20.5	2.3	2.3	2.0	2.2
DeepSeek-R1-7b	27.3	29.7	15.8	24.3	9.0	7.9	4.9	7.3
DeepSeek-R1-8b	21.4	14.4	15.6	17.1	2.0	1.9	2.9	2.3
DeepSeek-R1	76.3	82.3	48.2	68.9	62.6	71.4	37.6	57.2
Qwen-2-1.5b	53.5	51.3	27.2	44.0	20.0	13.4	3.1	12.2
+ DoLa	60.7	63.7	35.1	53.2	0.0	0.0	0.0	0.0
+ ICL (10-shot)	27.6	20.3	14.4	20.8	0.0	0.0	0.0	0.0
+ LLM4BioEL	89.4	91.0	84.3	88.2	92.4	92.2	82.7	89.1
Mistral-7b-v0.3	64.9	73.4	46.8	61.7	0.9	3.9	1.0	2.0
+ DoLa	68.4	74.4	47.2	63.3	0.0	0.5	0.2	0.2
+ ICL (10-shot)	64.9	73.4	47.4	61.9	0.9	3.9	28.3	11.1
+ LLM4BioEL	88.4	91.2	84.6	88.1	92.9	92.4	84.1	89.8
Llama-3-8b	64.3	68.3	35.3	55.9	36.3	49.8	22.9	36.3
+ DoLa	51.7	46.6	20.9	39.7	27.7	28.2	7.3	21.1
+ ICL (10-shot)	56.9	53.5	55.0	55.1	12.8	26.9	42.1	27.3
+ LLM4BioEL	87.5	91.0	84.9	87.8	93.2	92.2	84.4	89.9
Llama-3-70b	77.6	79.7	51.8	69.7	57.3	61.1	34.9	51.1
+ ICL (10-shot)	86.6	84.5	57.9	76.3	77.8	72.7	46.3	65.6
+ LLM4BioEL	89.2	91.1	85.2	88.5	93.8	92.4	84.8	90.3

2023) usually obtains higher metrics than our method, and on the COMETA dataset, for instance, the averaged Hits@1 drops from 43.5% to 27.3%, from 48.8% to 29.5%, and from 53.3% to 37.2% for GPT-3.5, PaLM-2, and Claude-2, respectively. For GPT-4 and DeepSeek-R1 models, similar Hits@1 degradation is observed. Besides, for Mistral-7b-v0.3 in greedy decoding, the evaluation method in (Jahan et al., 2023) obtains the averaged Hits@1 of 61.7% while ours obtains 2.0%, and for Qwen-2-1.5b in ICL with 10-shot prompts, their method obtains the average Hits@1 of 20.8% while ours obtains 0.0%. Our evaluation method considers the synonym relationships in the biomedical KG, which means that our method considers a generated entity as correct if it is a synonym of the ground truth. Therefore, LLM4BioEL typically produces higher Hits@1 metrics using our evaluation method than Jahan et al. (2023); for instance, LLM4BioEL (Llama-3-8b) achieves averaged Hits@1 of 87.8% with Jahan et al. (2023) but 89.9% using our evaluation method.