PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking

Yuzhang Xie
yuzhang.xie@emory.edu
Emory University, Atlanta, USA

Jiaying Lu
jiaying.lu@emory.edu
Emory University, Atlanta, USA

Joyce Ho
joyce.c.ho@emory.edu
Emory University, Atlanta, USA

Fadi Nahab
fnahab@emory.edu
Emory University, Atlanta, USA

Xiao Hu
xiao.hu@emory.edu
Emory University, Atlanta, USA

Carl Yang
j.carlyang@emory.edu
Emory University, Atlanta, USA

Abstract

Linking (aligning) biomedical concepts across diverse data sources enables various integrative analyses, but it is challenging due to the discrepancies in concept naming conventions. Various strategies have been developed to overcome this challenge, such as those based on string-matching rules, manually crafted thesauri, and machine learning models. However, these methods are constrained by limited prior biomedical knowledge and can hardly generalize beyond the limited amounts of rules, thesauri, or training samples. Recently, large language models (LLMs) have exhibited impressive results in diverse biomedical NLP tasks due to their unprecedentedly rich prior knowledge and strong zero-shot prediction abilities. However, LLMs suffer from issues including high costs, limited context length, and unreliable predictions. In this research, we propose PromptLink, a novel biomedical concept linking framework that leverages LLMs. Empirical results on the concept linking task between two EHR datasets and an external biomedical KG demonstrate the effectiveness of PromptLink. Furthermore, PromptLink is a generic framework without reliance on additional prior knowledge, context, or training data, making it well-suited for concept linking across various types of data sources. The source code of this study is available at https://github.com/constantjxyz/PromptLink.

CCS Concepts

• **Applied computing** → **Health care information systems**; • **Information systems** → *Retrieval models and ranking*;

Keywords

Biomedical Concept Linking, Few-Shot Prompting, Large Language Models for Resource-Constrained Field, Retrieve & Re-Rank

ACM Reference Format:
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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657904

SIGIR '24, July 14–18, 2024, Washington, DC, USA.

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657904

1 Introduction

Biomedical concept linking studies the intricate task of linking closely related concepts across different data sources by leveraging their semantic meanings and underlying biomedical knowledge, as exemplified in Figure 1 [29]. This linking process is crucial for enabling integrative analyses, as biomedical concepts obtained from diverse sources offer multifaceted views of biomedical knowledge and data [19, 32]. For example, the electronic health record (EHR), which is regarded as a valuable asset for comprehensive patient health analysis, contains various digital medical information including tabular data, clinical notes, and other types of patient data [1, 33, 39]. Similarly, the knowledge graph (KG), playing an important role in biomedical research, provides structured knowledge, such as definitions of concepts and their interrelationships [21]. However, the cross-source biomedical linking task is challenging due to discrepancies in the biomedical naming conventions used in different systems [15]. For example, a KG may mention a disease as "Ellis-Van Creveld syndrome", while an EHR may refer to the same disease as "Chondroectodermal dysplasia". This inconsistency presents a strong barrier to cohesive data analysis.

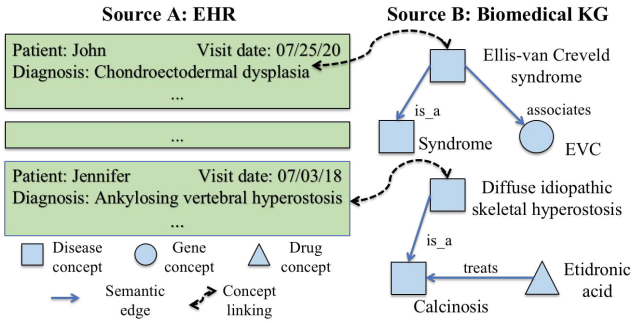

Figure 1: A toy example of biomedical concept linking. *Left:* concepts in the EHR. *Right:* concepts in the biomedical KG.

The challenge of biomedical concept linking has motivated the development of various methods. **Conventional methods** focus on setting *string-matching rules* [7, 13] and leveraging *constructed thesauri* [3, 9, 27]. However, their reliance on fixed rules and crafted thesauri limits coverage and generalizability in real-world scenarios [30]. Addressing these limitations, **machine learning-based methods** have been widely explored, avoiding the manual design of rules or thesauri. These methods essentially transform biomedical concepts from raw text into embeddings (latent vector representations), which are then used to compute similarity scores via distance functions (*e.g.* cosine similarity) or learning-based scoring functions

(*e.g.* bilinear attention [14]). Various models have been used to obtain biomedical concept embeddings, including *pre-trained language models* (PLMs) [34] that capture fine-grained semantic relations through extensive training on biomedical corpora [2, 16, 17, 38], and *graph neural networks* (GNNs) [42] that capture both semantics and relations of biomedical concepts [4, 10, 18]. Despite the notable achievements of these ML-based linking methods, they are data-hungry and require significant supervision signals when adapted into novel downstream applications. They face challenges due to the costly data annotation and model training processes.

Recently, large language models (LLMs) have exhibited impressive performances in various NLP tasks, due to their unprecedentedly rich prior knowledge and language capabilities [31, 35, 43], enabling various applications in a zero-shot learning setting [19]. Therefore, LLMs provide a promising solution for linking related concepts across different systems. Meanwhile, LLMs also face challenges including the design of effective and cost-efficient prompts within the context length limits [40], and the NIL prediction capability of reliably rejecting all candidates when correct concepts are absent, instead of returning relatively close but incorrect ones [23].

In this paper, we propose **PromptLink**, leveraging LLMs for the cross-source biomedical concept linking task. Considering LLMs' high cost and context length constraints, we first employ a pre-trained SAPBERT language model to generate biomedical-aware concept embeddings and retrieve top candidates based on the cosine similarities of these embeddings. We then design a novel two-stage prompting mechanism for the GPT-4 model to derive reliable linking predictions. The first stage efficiently filters out irrelevant candidates, thereby minimizing the response token numbers required in the subsequent stage. The second stage generates the final linking results and incorporates a self-verification prompt to address the NIL prediction challenge, effectively rejecting all candidates when none are relevant. In the experiments, PromptLink demonstrates exceptional performance, surpassing various existing concept linking methods by over 5% in two scenarios of biomedical concept linking between EHR and external biomedical KG, which could be attributed to LLM's intrinsic strong biomedical knowledge. Moreover, PromptLink works as a zero-shot framework due to the utilization of pre-trained language models, eliminating the need for a training process. It is also a versatile framework that performs well even when only concept names, without concept context or topological structure, are provided. Given its various advantages, PromptLink boasts strong generalization capabilities, making it suitable for a wide range of biomedical research and applications.

2 Biomedical Concept Linking

2.1 Problem Definition

The biomedical concept linking links biomedical concepts across sources based on semantic meanings and biomedical knowledge, using only concept names to cover broader applications. This task differs from existing tasks such as entity linking [30], entity alignment [19], and ontology matching [11] that require extra contextual or topological information. In this study, we link the EHR concepts to corresponding concepts in a biomedical KG. We define an EHR database $\mathcal{D}$, a biomedical KG $\mathcal{G}$, and the linking task as follows:

Definition 1 (EHR)**.** An EHR database $\mathcal{D}$ is a relational database $\mathcal{D} = (P, A, V)$, with P being patient identifiers, A patient attributes, $V \in P \times A$ the values of these attributes. Additionally, M represents multi-token biomedical concepts associated with patient attributes.

Definition 2 (Biomedical KG)**.** A biomedical KG is a multi-relation graph $\mathcal{G} = (C, R, RT)$, where C are concepts, R are relation names, and $RT \in C \times R \times C$ are the relational triples among them.

Definition 3 (Biomedical Concept Linking)**.** Link identified biomedical concepts from an EHR $\mathcal{D}$ to a biomedical KG $\mathcal{G}$ based on semantic meanings and biomedical knowledge, forming linkages $LK = \{(m_i, c_j) | m_i \in M_{\mathcal{D}}, c_j \in C_{\mathcal{G}} \cup \text{NIL}\}$. If a concept m from D is not in G, link it to a special "*NIL*" entity, indicating it is unlinkable.

2.2 PromptLink

We propose PromptLink, a novel LLM-based solution for cross-source biomedical concept linking, as illustrated in Figure 2. Addressing LLMs' high cost and limited input text length, we first employ a biomedical-specialized pre-trained language model to generate concept embeddings and retrieve top candidates via cosine similarities. Subsequently, we employ a two-stage prompting mechanism with GPT-4 to generate the final linking predictions.

Concept representation and candidate generation. After preprocessing text by lowercasing and removing punctuation, we use a pre-trained LM (specifically *SapBERT* [17]), to create embeddings $\boldsymbol{h} \in \mathbb{R}^{1\times d}$ for EHR concepts m and KG concepts c, represented as $\boldsymbol{h_m} = PLM(m)$ and $\boldsymbol{h_c} = PLM(c)$, respectively. For concepts that span multiple tokens, the token-level embeddings are averaged to create the concept embedding. This model helps to project the semantic meanings and prior biomedical knowledge into the embedding space. For candidate generation, we compute cosine similarity $S \in [0, 1]$ between pairs of EHR concept embedding $\boldsymbol{h_m}$ and KG concept embedding $\boldsymbol{h_c}$, represented as: $S = cos(\boldsymbol{h_m}, \boldsymbol{h_c})$. Given each input query EHR concept m_i, We select the top-K (K=10) KG concepts $[c_1, c_2, \ldots, c_K]$ with the highest similarities as candidates for further GPT-based linking prediction.

Linking prediction using two-stage prompts. The next step of our framework is generating linking predictions of query m_i over the top-K candidate $[c_1, c_2, \ldots, c_K]$ using GPT-4 model, leveraging its text comprehension ability, logical reasoning ability, and prior biomedical knowledge [5, 31]. In this step, we design a novel two-stage prompt for our task, as can be seen in Figure 2. Combining the two prompts utilizes their strengths and mitigates weaknesses. The first stage focuses on concept pairs to filter out unrelated candidates. The second stage evaluates all candidates in a broader context to identify the closest match or reject all unmatch candidates.

In the first stage, the LLM is prompted to check if a concept pair (m_i, c_j) should be linked. By defining the response structure, the LLM can return answers in specified formats. To improve the prompt response quality, we adopt the self-consistency [36] prompting strategy that repeatedly prompts the same question to the LLM multiple times. Specifically, we prompt each concept pair (m_i, c_j) for $n = 5$ times, thus obtaining the belief score $B_{i,j} \in [0, 1]$ by:

$$B_{i,j} = \frac{\text{number of "yes" responses}}{n}.$$

Considering the belief scores across different candidates, we derive a comprehensive filter strategy to exclude irrelevant candidates, using parameter τ (set as $0.8 \times n$). This approach ensures that irrelevant candidates are not considered in the next stage, optimizing both efficiency and effectiveness. The approach is described as follows:

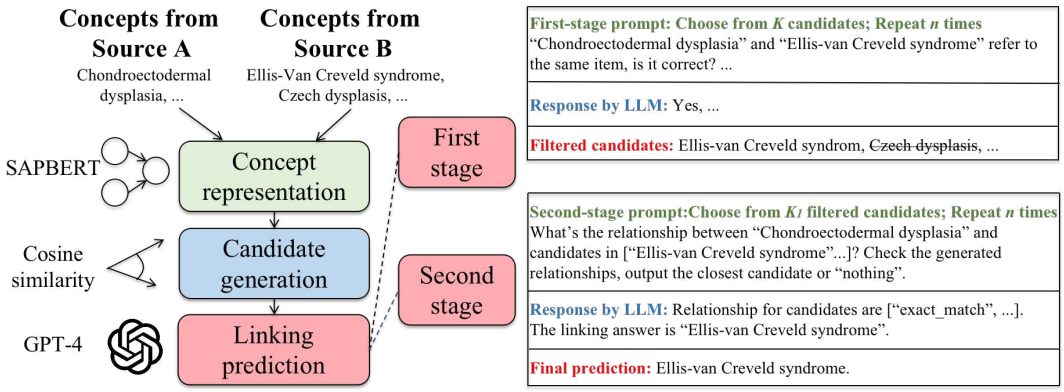

Figure 2: Overview of our proposed PromptLink framework.

- If $max(B_{i,1}, \cdots, B_{i,K}) \geq \tau$, this indicates some candidates closely align with the query concept. In such cases, candidates with belief scores of zero will be filtered out as they are deemed irrelevant to the query concept and there are closely aligning alternatives. This removes many irrelevant candidates, thereby optimizing both efficiency and effectiveness for the subsequent stage.
- Otherwise, the range of different candidates' belief scores is not wide enough to justify filtering. Thus, all K candidates will be subjected to double-checking by the second-stage prompt.

In the second stage, the LLM evaluates the K_1 candidates retained from the first stage's filtering process $[c'_1, c'_2, \ldots, c'_{K_1}]$, where $K_1 \leq K$, using a compositional prompt that consists of two consecutive questions to perform complex reasoning. Specifically, the LLM is asked to (1) label the relationship between the query concept and all candidate concepts as "exact match", "related to", or "different from"; (2) use self-verification prompts to either identify the closest candidate or dismiss all candidates if none are close, thus the final concept linking result of this prompt is K_2 (usually $K_2 = 1$) item from $[c'_1, \ldots, c'_{K_1}] \cup [\text{NIL}]$. In this stage, we also use the self-consistency strategy that prompts one question for the same n times. Subsequently, we calculate the occurrence frequency $f_{i,j} \in [0, 1]$ for answers in $[c'_1, c'_2, \ldots, \text{NIL}]$ and retrieve the final linking result for query EHR concept m_i as follows:

- If $f_{i,\text{NIL}} > 0.5$, this indicates a high probability that no candidates are appropriate. Thus, "NIL" is chosen as the final prediction.
- Otherwise, the candidate c_j with the highest frequency $f_{i,j}$ is decided as the final prediction. In case of a tie, the one c_j with higher embedding similarity $S_{i,j}$ to the query concept m_i is chosen.

3 Experiments & Discussions

3.1 Implementation Details

Datasets. In our experiments, we curate two biomedical concept linking benchmark datasets: *MIID* (**M**IMIC-**I**II-**i**BKH-**D**isease) and *CISE* (**C**RADLE-**i**BKH-**S**ide-**E**ffect). MIID comprises 1,493 diagnosis concepts from MIMIC-III [12], which is an EHR dataset including over 53,423 patient records, and 18,697 disease concepts from iBKH [32], which is a KG dataset with 2,384,501 entities. To construct MIID, we first remove exact matches between MIMIC-III diagnosis concepts and iBKH disease concepts. Then, we link the remaining MIMIC concepts to iBKH using mappings between ICD-9 [8] and UMLS CUI [28] codes. We use the linked concept pairs as ground-truth labels only for evaluation purposes. CISE contains 1,500 CRADLE [39] diagnosis concepts and 4,251 iBKH drug side-effect concepts, constructed by using CUI [28] and SNOMED CT [6] codes. Ground-truth matched pairs are also only used for evaluation purposes.

Experimental Settings. Following the definition in Sec. 2.1 and recognizing the scarcity of supervision in the biomedical domain, we mainly focus on the biomedical concept linking under the zero-shot setting. Additionally, our biomedical concept linking task solely relies on concept names for broad real-world application coverage. Given this characteristic of our data, graph-based linking methods, such as selfKG [18], are not applicable as they need topological information to establish concept alignment. Similarly, thesauri-based methods, such as MetaMap [3], are unsuitable as they only establish links between concepts existing in the pre-defined vocabulary. Therefore, the following baseline methods are compared:

- Conventional methods: **Cosine Distance**, **Jaccard Distance**, **Levenshtein Distance** [24], **Jaro-Winkler Distance** [37], **BM25** [25]. These methods measure the concept pairs' string similarity and relevance and then obtain the linking prediction result.
- Machine learning-based methods: Pre-trained language models are used to generate concept embedding and linking prediction results (according to embedding cosine similarity). Specifically, we select representative PLMs including **BioBERT** [16], **BioGPT** [20], **BioClinicalBERT** [2], **BioDistilBERT** [26], **KRISS-BERT** [41], **ada002** [22], and **SAPBERT** [17].

3.2 Concept Linking Experiment Results

Table 1 shows the accuracy of our proposed PromptLink along with baseline methods, when every method links a query EHR concept m_i with their predicted top-1 KG concept c_j. As can be seen, PromptLink outperforms competing approaches across both datasets in terms of zero-shot accuracy, underscoring the superiority of our LLM-based concept linking methodology. Among the compared methods, SAPBERT, a SOTA biomedical entity linking method, achieves the second-highest performance. Moreover, conventional methods lag behind machine learning methods, which leverage embeddings from pre-trained language models to effectively match conceptually similar but lexically distinct entities like "Ellis-Van Creveld syndrome" and "Chondroectodermal dysplasia".

3.3 Ablation Studies

Prompt Effectiveness and Efficiency. We conduct ablation studies to reveal the effectiveness and cost-efficiency of the prompt used

Table 1: Comparison of the zero-shot accuracy for different methods on MIID and CISE.

Method	Acc-MIID	Acc-CISE
Cosine Distance	0.2981	0.2907
Jaccard Distance	0.2123	0.3280
Levenshtein Distance	0.1995	0.3033
Jaro-Winkler Distance	0.3141	0.3693
BM25	0.4722	0.3993
BioBERT	0.3423	0.5280
BioClinicalBERT	0.3007	0.5007
BioGPT	0.3530	0.5093
BioDistilBERT	0.4240	0.5293
KRISSBERT	0.5265	0.5787
ada002	0.5968	0.6773
SAPBERT	0.7213	0.8167
PromptLink	**0.7756**	**0.8880**

Table 2: Ablation results with different prompting methods used by PromptLink on the MIID dataset.

Prompting Methods	Acc	Token Cost
Before Prompting	0.7213	N/A
First-stage Prompt	0.7595	995,836 ($36.59)
Second-stage Prompt	0.7634	1,681,987 ($88.69)
Two-stage Prompts	**0.7756**	1,594,996 ($66.25)

in our approach, as shown in Table 2. This comparison uses the same input data and 10 linking candidates across various prompts. In the table, the "Before prompting" denotes the performance of using only embedding similarity obtained from the pre-trained LM, while other methods use LLM to predict linking results based on LM-generated candidates. From Table 2, the "Before Prompting" method achieves the worst accuracy, demonstrating that linking performance could be improved by using LLM. Notably, PromptLinkwith both two-stage prompts achieves the best accuracy with the second-highest cost (~ 1.7M total tokens, costing approximately $66.25), indicating that the combined effect of the prompts substantially enhances accuracy, with the costs being moderated by the first stage's proficiency in eliminating unrelated candidates.

NIL Prediction. Another ablation study examines PromptLink's NIL prediction ability. In our built MIID and CISE datasets, each query EHR concept m_i is designed to have a ground-truth linking KG concept c_j. To reflect the real-world unlinkable scenario, we extend our MIID dataset into "*MIID-NIL*" which contains a proportion (25%) of unlikable EHR concept m_i. In Figure 3, the overall accuracy of PromptLink in the MIID-NIL dataset is 0.8145. Specifically for the unlikable concepts, PromptLink outputs the expected "NIL" with 0.9290 accuracy, which validates the NIL prediction ability of our proposed method. Existing methods highly rely on the hard-coded threshold. For example, we could threshold SAPBERT's generated embeddings' cosine similarity, then output the KG concept with the highest similarity above the threshold or "NIL" when none are above. However, this straightforward idea, requiring a manually set threshold, is less effective than PromptLink. As shown in Figure 3, SAPBERT achieves lower accuracy (maximum value 0.7920) no matter what the threshold value is, which corresponds to our assumptions. When the threshold value is low, SAPBERT generates many wrong predictions to unlinkable query concepts; otherwise, SAPBERT continues to output "NIL" for many linkable concepts.

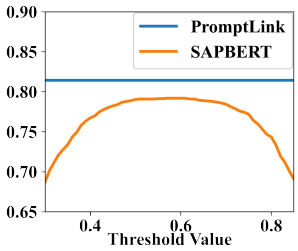

Figure 3: Accuracy on MIID-NIL: Traditional ML-based methods outputting matching scores have varying NIL prediction performance based on the selected threshold, while PromptLink does not need a threshold yet consistently performs better.

3.4 Case Studies

In case studies on linking EHR concepts to MIID's KG disease concepts, three scenarios are presented: (1) concepts assessed by both ground-truth labels and a clinician; (2) concepts evaluated by a clinician due to missing ground-truth labels; (3) irrelevant concepts judged by a clinician. The linking results of PromptLink and SAPBERT are presented in Table 3. Overall, PromptLink could link biomedical concepts more accurately and appropriately. For casse I-V, PromptLink's linking results are justified by the ground-truth label and clinician. Specifically, for cases I and II, PromptLink accurately links the EHR concepts to conceptually similar but lexically distinct KG concepts, while SAPBERT links to lexically similar but conceptually different KG concepts. This difference showcases the effective use of LLM's biomedical knowledge. SAPBERT also shows inaccuracies in cases III-IV, and provides a broader prediction in case V, whereas PromptLink's predictions are more accurate and specific. For cases VI-IX, where linking ground truth labels are lacking, PromptLink's predictions also align more accurately with EHR concepts than SAPBERT's, according to a clinician's review. In cases VI and VII, PromptLink closely matches the EHR concepts, while SAPBERT's predictions are overly specific. In cases VIII and IX, PromptLink correctly and automatically identifies no matching KG disease concepts, while SAPBERT fails to resolve that NIL prediction challenge unless manual thresholds are set and adjusted.

Table 3: Analyzed cases.

ID	EHR Concept	PromptLink's Prediction	SAPBERT's Prediction
I	Chondroectodermal dysplasia	Ellis-van Creveld syndrome 😀🏷️	Cranioectodermal dysplasia
II	Dermatophytosis of hand	Tinea manuum 😀🏷️	Hand dermatosis
III	Late syphilis, unspecified	Tertiary syphilis 😀🏷️	Secondary syphilis
IV	Hypopotassemia	Hypokalemia 😀🏷️	Hypocupremia nos
V	Epidemic vertigo	Vestibular neuronitis 😀🏷️	Vertigo
VI	Postprocedural fever	Postoperative complications 😀	Postcardiotomy syndrome
VII	Acquired cardiac septal defect	Heart septal defect 😀	Atrial heart septal defect
VIII	Height of bed	NIL 😀	Binge eating disorder
IX	Level one	NIL 😀	Glaucoma 1 open angle

Note: "😀" indicates this prediction is justified by the clinician. "🏷️" indicates this prediction is justified by the ground-truth label.

4 Conclusion

In this study, we introduce PromptLink, a novel framework leveraging LLMs and multi-stage prompts for effective biomedical concept linking. Compared with previous concept linking methods, PromptLink achieves better linking accuracy, attributed to LLM's intrinsic strong biomedical knowledge. PromptLink further employs multi-stage prompts to maintain cost-efficiency and handle the NIL prediction problem. Moreover, PromptLink functions as a zero-shot framework, requiring no training and demonstrating strong flexibility and generalizability across biomedical systems. Promising future work can focus on further enhancing the prompt effectiveness, reducing costs, and minimizing manual efforts, aiming to extend PromptLink's application to broader systems.

References

[1] Noura S Abul-Husn and Eimear E Kenny. 2019. Personalized medicine and the power of electronic health records. *Cell* 177, 1 (2019), 58–69.

[2] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly Available Clinical BERT Embeddings. *NAACL HLT 2019* (2019), 72.

[3] 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 (2010), 229–236.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[6] Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.

[7] Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 297–302.

[8] Centers for Disease Control. 2007. International Classification of Diseases-9-CM. Available at http://www.cdc.gov/nchs/icd.htm. Accessed Feb, 2024.

[9] Carol Friedman, Hongfang Liu, Lyudmila Shagina, Stephen Johnson, and George Hripcsak. 2001. Evaluating the UMLS as a source of lexical knowledge for medical language processing.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 189.

[10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[11] Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. Logmap: Logic-based and scalable ontology matching. In *10th International Semantic Web Conference*. 273–288.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[13] Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. 2013. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association* 20, 5 (2013), 876–881.

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems* 31 (2018).

[15] Isaac S Kohane, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, Mario Cannataro, James J Cimino, Noelia García-Barrio, et al. 2021. What every reader should know about studies using electronic health record data but may be afraid to ask. *Journal of medical Internet research* 23, 3 (2021), e22219.

[16] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[17] 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*. 4228–4238.

[18] Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, and Jie Tang. 2022. Selfkg: Self-supervised entity alignment in knowledge graphs. In *Proceedings of the ACM Web Conference 2022*. 860–870.

[19] Jiaying Lu, Jiaming Shen, Bo Xiong, Wenjing Ma, Steffen Staab, and Carl Yang. 2023. Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval - Short Paper*.

[20] 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 (2022), bbac409.

[21] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1910–1919.

[22] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005* (2022).

[23] Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 43–54.

[24] Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 5 (1998), 522–532.

[25] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[26] Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, and David A Clifton. 2023. On the effectiveness of compact biomedical transformers. *Bioinformatics* 39, 3 (2023), btad103.

[27] 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 (2010), 507–513.

[28] Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association* 81, 2 (1993), 217.

[29] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web* 13, 3 (2022), 527–570.

[30] 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* (2023), 1–30.

[31] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[32] Chang Su, Yu Hou, Manqi Zhou, Suraj Rajendran, Jacqueline RMA Maasch, Zehra Abedi, Haotan Zhang, Zilong Bai, Anthony Cuturrufo, Winston Guo, et al. 2023. Biomedical discovery through the integrative biomedical knowledge hub (iBKH). *Iscience* 26, 4 (2023).

[33] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. 2018. Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering* 2018 (2018).

[34] Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pre-trained language models in biomedical domain: A systematic survey. *Comput. Surveys* 56, 3 (2023), 1–52.

[35] Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023. Exploring the in-context learning ability of large language model for biomedical concept linking. *arXiv preprint arXiv:2307.01137* (2023).

[36] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

[37] William E Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. (1990).

[38] 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*. 8452–8464.

[39] Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2022. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*. PMLR, 259–278.

[40] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534* (2023).

[41] 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*. 868–880.

[42] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open* 1 (2020), 57–81.

[43] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models Are Human-Level Prompt Engineers.(2023). *ProQuest Number: INFORMATION TO ALL USERS* 30490868 (2023).