LLM as Entity Disambiguator for Biomedical Entity-Linking

Christophe Ye **Cassie S. Mitchell**

Georgia Institute of Technology

cye73@gatech.edu,
cassie.mitchell@bme.gatech.edu

Abstract

Entity linking involves normalizing a mention in medical text to a unique identifier in a knowledge base, such as UMLS or MeSH. Most entity linkers follow a two-stage process: first, a candidate generation step selects high-quality candidates, and then a named entity disambiguation phase determines the best candidate for final linking. This study demonstrates that leveraging a large language model (LLM) as an entity disambiguator significantly enhances entity linking models' accuracy and recall. Specifically, the LLM disambiguator achieves remarkable improvements when applied to alias-matching entity linking methods. Without any fine-tuning, our approach establishes a new state-of-the-art (SOTA), surpassing previous methods on multiple prevalent biomedical datasets by up to 16 points in accuracy. We released our code on GitHub at https://github.com/ChristopheYe/llm_disamb.

1 Introduction

The biomedical domain is an information-rich and highly specialized field, characterized by vast amounts of domain-specific knowledge and intricate terminologies. Unlike named entity recognition (NER), which focuses on identifying entity mentions in text, biomedical entity linking (EL) goes a step further by mapping these mentions to unique identifiers in a structured knowledge base (KB). This distinction is critical, as biomedical texts often contain extensive synonyms, polysemy, and abbreviations, where the same concept may be expressed in multiple ways, or a single term may refer to different entities depending on context. Effective EL is essential for resolving these ambiguities, enabling more precise information retrieval (Lee et al., 2016), improving knowledge discovery (Wang et al., 2018), and facilitating downstream tasks such as automated data annotation and curation. By strengthening EL systems, researchers can unlock more efficient data integration, enhance text mining applications, and accelerate advancements in biomedical research and healthcare.

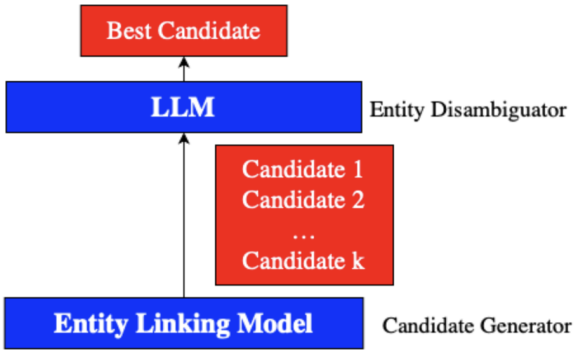

Figure 1: LLM as Entity Disambiguator.

Various methods have been applied to biomedical EL. Alias-matching approaches align mentions with entities based on lexicographical properties. This includes models like SciSpacy (Neumann et al., 2019), MetaMap (Aronson and Lang, 2010) or SapBERT (Liu et al., 2021). Contextualized EL learns vector representations (Vaswani et al., 2023) of entities by leveraging the contextual information from the mention and try to match it to an entity description. This includes models like KRISSBERT (Zhang et al., 2022), ClusterEL (Angell et al., 2021) or ArboEL (Agarwal et al., 2022). These representations are typically employed to generate a list of high-quality candidates for a mention, which are then refined using a re-ranker (Wu et al., 2020).

Large language models (LLMs) have recently shown promise in various biomedical Natural Language Processing (NLP) (Tian et al., 2023) applications. However, for information extraction tasks like entity recognition, their performance still lags significantly behind the previously cited models (Jahan et al., 2024).

Prior work has leveraged LLMs as external tools

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 301–312

to enhance data quality for improved entity normalization. For instance, Borchert et al. (2024) simplifies entity mentions by rephrasing long or complex terms into more general or concise forms before candidate generation, while Chen employs LLMs to generate high-quality augmented data, improving the performance of traditional biomedical entity linking models. Garda and Leser (2024) integrates LLM-powered homonym disambiguation, which resolves ambiguous entity names by appending clarifying descriptors, and candidate sharing, which enhances training by leveraging contextually related mentions within the same document. However, no prior work has fully integrated LLMs as a core component of the entity linking pipeline itself.

This is primarily due to a lack of domain-specific knowledge and tendencies toward hallucination. However, with sufficient information, LLMs could potentially leverage their general understanding of text to perform the normalization. In this work, we explore the use of LLMs for biomedical entity disambiguation task.

This paper contributes the following :

- LLM is introduced as a **entity disambiguator** for named entity disambiguation, building on top of the existing methods for candidate generation.
- LLM entity disambiguation achieved consistent improvements for alias-matching EL models. It outperformed previous SOTA for several datasets with a remarkable gain of 16 points in GNormPlus.
- The method can be seamlessly integrated into existing EL models without requiring additional training.

2 Methodolody

2.1 Entity Linking model for candidate generation

The first step of our approach involves generating candidates using EL models. While multiple models are available, we will concentrate on the top-performing model in each category: SapBERT for alias-matching EL and ArboEL for contextualized EL.

SapBERT (Liu et al., 2021) selects top candidates by computing cosine similarity between the test mention and all aliases in the database, using enhanced embedding representations. ArboEL (Agarwal et al., 2022) first trains a bi-encoder to improve embedding representations using mention–mention coreference signals. This is followed by a cross-encoder training that scores each candidate by concatenating the test mention with the candidate's context. The top-k candidates are then selected based on these scores.

2.2 Disambiguation step with LLMs

2.2.1 The mention

To fully leverage the capabilities of the LLM, it's crucial to provide the mention along with its surrounding context, ensuring the model has access to the maximum amount of relevant information. We formatted the mention for the LLM as follows:

$$c_{left} \text{ [ENTITY START] } m \text{ [ENTITY END] } c_{right}$$

where c_{left} and c_{right} represent the left and right contexts of the mention m.

2.2.2 The entity

Similarly, to ensure maximum reliability, we provided the entity along with its relevant data to the LLM in the following dictionary format:

```
entity_data = {
    "cui": entity.cui,
    "name": entity.name,
    "types": entity.types,
    "aliases": entity.aliases,
    "definition": entity.definition,
}
```

- cui: Concept Unique Identifier
- name: Entity name
- types: Entity Types
- aliases: Entity aliases
- definition: Entity definition

2.2.3 In-context Learning

In-context learning (Dong et al., 2024) is an approach in natural language processing (NLP) where LLMs make predictions based on contexts enriched with a few examples. It has been shown that incorporating such examples in the prompt can enhance model performance. To optimize the selection of examples, we employed a RAG-like system (Lewis et al., 2021) that retrieves the most relevant examples from the training set to include in the prompt.

Specifically, the LLM predicts the response to the query x by conditioning on k training examples $\{(x_i, y_i)\}_{i=1}^{k}$ in the prompt :

$$\hat{y} = \text{LLM}(x|x_1, y_1, \cdots, x_k, y_k)$$

To find the most relevant examples for a given query, we first embed all mentions along with their surrounding context from the training set using a sentence embedding model (Gao et al., 2022). We then create an index using Faiss (Douze et al., 2024) for faster and more efficient retrieval. During inference, we embed the query (the mention along with its context) and retrieve the most similar mentions from the corpus for the LLM to process.

3 Results

The performance of LLMs as entity disambiguators was evaluated across multiple biomedical datasets. Prior work has typically relied on supervised learning using BERT-base models for entity linking tasks. In a comprehensive evaluation, Kartchner et al. (2023) identified SapBERT as the most effective alias-matching model, and ArboEL as the strongest contextualized model. These results were also later confirmed by Bathala et al. (2025) in their benchmarking package. Based on these evidences, our assessment is conducted by applying LLMs to these two leading approaches.

Performance was evaluated using five different models : Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023), Mistral-Nemo-Instruct-2407 (Jiang et al., 2024), Qwen2.5-7B-Instruct (Bai et al., 2023), GPT-4o-mini and GPT-4o (team, 2024).

All LLMs were prompted using a greedy-decoding strategy; temperature was set to 0.

A comprehensive description of the experimental setup is provided in Appendix A.1 to ensure reproducibility of the experiments. This includes the exact details of the prompt used to obtain the correct CUI and the top-k candidates (Appendix A.1.1), along with the analysis of the impact of the number of candidates (Appendix A.1.3) and number of examples (Appendix A.1.4) to include in the prompt.

3.1 Datasets

The experiments were carried out on five prevalent biomedical datasets coming from BigBio (Fries et al., 2022) : NCBI-Disease (Dogan et al., 2014), GNormPlus (Wei et al., 2015), NLM Chem (Dogan et al., 2021a), NLM Gene (Dogan et al., 2021b), and Medmentions-ST21PV (Mohan and Li, 2019).

Details on the datasets are provided in Table 1.

The corresponding ontologies are MEDIC (Davis et al., 2019), Entrez (Maglott et al., 2005), MeSH (Lipscomb, 2000), and UMLS (Bodenreider, 2004).

Dataset	Train Mentions	Test Mentions	Ontology
NCBI-Disease	6,881	960	MEDIC
GNormPlus	6,252	3,223	Entrez
NLM-Chem	37,999	11,660	MeSH
NLM-Gene	15,553	2,729	Entrez
MM-ST21PV	203,282	40,143	UMLS

Table 1: Datasets used for evaluation.

3.2 With Alias-Matching EL

The impact of LLMs as entity disambiguators is illustrated for alias-matching EL methods using SapBERT. All results are detailed in Table 2.

When the base model performs well, with accuracy and recall@20 being close, the LLM's impact is minimal. For instance, on the NCBI-Disease dataset, accuracy improves slightly, while recall@5 remains unchanged.

However, when the initial model struggles, integrating an LLM can significantly boost performance. For instance, SapBERT fails to effectively differentiate aliases in gene-centric datasets like GNormPlus and NLM-Gene, as indicated by the large gap between its accuracy and recall@20. Using GPT-4o as a disambiguator improved SapBERT's accuracy on GNormPlus from 19.1% to 74.8% and recall@5 from 56.6% to 83.8%, surpassing the previous SOTA by 16% in accuracy and 19% in recall@5.

SapBERT struggles to distinguish between multiple gene entities sharing the same alias. However, since candidates with the same alias are ranked closely, the correct entity often appears shortly after in the list. By leveraging contextual information, the LLM can distinguish the correct entity more efficiently addressing this limitation.

GPT-4o is the top-performing disambiguator for alias matching EL models, outperforming smaller 7B-12B models. However, it is noteworthy that these "small" open-source LLMs still provide significant improvements and even surpass the previous state-of-the-art (SOTA) in many scenarios.

The reported SOTA numbers are based on the results we obtained from re-running the models ourselves using the BioEL package with default parameters (Bathala et al., 2025), not the ones originally reported by (Kartchner et al., 2023).

	Base: SapBERT			Base model + Llama3		Base model + Mistral		Base model + Qwen2.5		Base model + GPT-4o-mini		Base model + GPT-4o		SOTA	
	@1	@5	@20	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
NCBI-Disease	0.752	0.899	0.924	0.781 ↑	0.835 ↓	0.782 ↑	0.852 ↓	**0.791** ↑	**0.899**	0.783 ↑	0.887 ↓	0.790 ↑	0.869 ↓	0.771	0.820
GNormPlus	0.191	0.566	0.862	0.444 ↑	0.780 ↑	0.482 ↑	0.770 ↑	0.365 ↑	0.772 ↑	0.386 ↑	0.761 ↑	**0.748** ↑	**0.838** ↑	0.585	0.647
NLM-Chem	0.754	0.876	0.889	0.842 ↑	0.876	0.833 ↑	0.879 ↑	0.849 ↑	0.876	0.849 ↑	0.879 ↑	**0.859** ↑	**0.879** ↑	0.790	0.856
NLM-Gene	0.072	0.344	0.824	0.314 ↑	0.659 ↑	0.324 ↑	0.571 ↑	0.251 ↑	0.599 ↑	0.251 ↑	0.556 ↑	0.497 ↑	0.668 ↑	**0.559**	**0.751**
MM-ST21PV	0.594	0.771	0.794	0.667 ↑	0.776 ↑	0.664 ↑	0.787 ↑	0.675 ↑	0.785 ↑	N/A	N/A	N/A	N/A	**0.685**	**0.798**

Table 2: Comparison Accuracy (Recall@1) and Recall@5 between initial base model SapBERT and after applying different LLM disambiguators. N/A : Not Available due to prohibitive cost considerations. SOTA model is ArboEL.
x : Best result / x : Beat SOTA / ↑ : Improvement over base / ↓ : Degradation from base

	Base: ArboEL			Base model + Llama3		Base model + Mistral		Base model + Qwen2.5		Base model + GPT-4o-mini		Base model + GPT-4o	
	@1	@5	@20	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
NCBI-Disease	0.771	0.820	0.838	0.760 ↓	0.816 ↓	0.749 ↓	0.821 ↑	0.767 ↓	0.822 ↑	0.764 ↓	0.833 ↑	0.758 ↓	**0.837** ↑
GNormPlus	0.585	0.647	0.659	0.580 ↓	0.649 ↑	0.585	0.646 ↓	0.600 ↑	0.647	0.600 ↑	0.652 ↑	**0.618** ↑	**0.654** ↑
NLM-Chem	0.790	**0.856**	0.863	0.815 ↑	0.849 ↓	0.806 ↑	0.847 ↓	0.821 ↑	0.849 ↓	**0.830** ↑	0.853 ↓	0.815 ↑	0.852 ↓
NLM-Gene	**0.559**	**0.751**	0.778	0.549 ↓	0.700 ↓	0.549 ↓	0.714 ↓	0.550 ↓	0.726 ↓	0.538 ↓	0.736 ↓	0.537 ↓	0.726 ↓
MM-ST21PV	0.685	**0.798**	0.811	0.649 ↓	0.735 ↓	0.641 ↓	0.767 ↓	**0.686** ↑	0.790 ↓	N/A	N/A	N/A	N/A

Table 3: Comparison Accuracy (Recall@1) and Recall@5 between initial base model ArboEL and after applying different LLM disambiguators. N/A : Not Available due to prohibitive cost considerations. ArboEL is SOTA model.
x : Best result / x : Beat SOTA / ↑ : Improvement over base / ↓ : Degradation from base

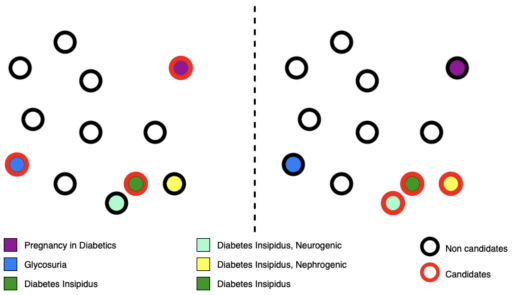

Figure 2: Embedding Space of Candidates from Alias Matching (left) and Contextualized EL model (right)

3.3 With Contextualized EL

The impact of LLMs as entity disambiguators was demonstrated for contextualized EL methods using ArboEL. The results are detailed in Table 3.

When applied to contextualized EL, the results show minimal improvement for the best model and even degrade for others. This happens because the base model already handles mentions effectively when the correct CUIs are included in the candidate list, as it is also considering the contextual information of the mention.

3.4 Performance disparity

The performance gap between contextualized and alias-matching EL models stems from their differing approaches. Alias-matching methods often achieve higher recall at large recall levels by capturing a broad range of potential matches based on surface similarities (e.g., aliases or synonyms), without filtering out even when the candidates are totally different.

Context-aware models, while generally more precise, tend to offer less improvement in recall@k at larger k values. These models are designed to sharply distinguish between correct and incorrect entities, creating a large gap between top-ranked candidates and those deemed contextually different. While this helps in pinpointing the correct entity, it often discards candidates that deviate from the expected answer. As a result, if the correct answer is not initially ranked high due to contextual ambiguity, it may be pushed far down the ranking for appearing too dissimilar.

Consequently, context-aware models may miss the correct candidate in large pools, where alias-matching methods would still retain it, even if ranked lower.

Figure 2 provides an example of such scenario :

while alias-matching EL produced a diverse set of top-ranked candidates, (e.g., "Pregnancy in Diabetics," "Glycosuria," "Diabetes Insipidus") contextualized EL generated candidates that were more closely related and grouped together in the rankings. (e.g., "Diabetes Insipidus," "Diabetes Insipidus, Neurogenic," "Diabetes Insipidus, Nephrogenic")

Further results on error analysis, performance disparity in low-data slices and data leakage risks are available in Appendix A.2

4 Conclusion and Future Work

Entity linking is crucial in knowledge-driven NLP, particularly in scientific and biomedical domains, where accurately mapping textual mentions to specific concepts is essential for extracting meaningful insights and advancing research. This study introduced a novel approach that utilized LLMs as entity disambiguators for biomedical EL. The proposed methodology demonstrated a significant performance improvement when integrated with an alias-matching-based EL model, as evidenced by experiments conducted on five standard biomedical EL datasets. Notably, this approach requires no fine-tuning, making it highly adaptable as LLMs continue to evolve in robustness and scalability. A promising direction for future work is the development of a retrieval mechanism to identify additional high-quality candidates when the correct entity is absent from the initial candidate set due to omissions by the underlying EL model.

5 Limitations

This method has demonstrated substantial effectiveness in enhancing the performance of alias-matching EL models. However, its application to contextualized EL models can occasionally result in performance degradation, leading to inconsistencies in outcomes. Furthermore, leveraging LLMs for entity disambiguation introduces a significant trade-off in inference time, which poses potential scalability challenges for datasets containing millions of mentions. The approach is also highly dependent on the candidate generation step; if the correct Concept Unique Identifier (CUI) is absent from the candidate set, there is no mechanism to retrieve it. Additionally, the use of proprietary LLMs can be prohibitively expensive. Finally, LLMs remain susceptible to errors, a critical concern in biomedical applications where precision is essential and any degree of randomness is unacceptable.

Acknowledgments

Funding support provided by National Science Foundation CAREER grant 1944247, National Institute of Health grant R35GM152245, and Chan Zuckerberg Initiative grant 253558 to C.S.M.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. pages 4644–4658.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based inference for biomedical entity linking. *Preprint*, arXiv:2010.11253.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner, and Cassie S. Mitchell. 2025. BioEL: A comprehensive python package for biomedical entity linking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1709–1721, Albuquerque, New Mexico. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Florian Borchert, Ignacio Llorca, and Matthieu-P Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with llm-based text simplification. *Database*, 2024:baae067.

Haihua Chen.

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2019. The comparative toxicogenomics database: update 2019. *Nucleic acids research*, 47(D1):D948–D954.

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, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, and more Deepseek team. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Rezarta Dogan, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, Rob Guzman, Preeti Kochar, Stella Koppel, Dorothy Trinh, Keiko Sekiya, Janice Ward, Deborah Whitman, Susan Schmidt, and Zhiyong lu. 2021a. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. *Scientific Data*, 8.

Rezarta Dogan, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong lu. 2021b. Nlm-gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of Biomedical Informatics*, 118:103779.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Jason A. Fries, Troy Mayfield, Kenneth Brimacombe, Michael M. Bronstein, David Silver, David Wehner, et al. 2022. Bigbio: A large-scale biomedical corpus. *Journal of Biomedical Informatics*, 135:104037.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

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

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478, Singapore. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik Choon Tan, and Jaewoo Kang. 2016. Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLOS ONE*, 11.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. 2005. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *Preprint*, arXiv:1902.09476.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenAI team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

S Tian, Q Jin, L Yeganova, et al. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. arxiv. *arXiv preprint arXiv:2306.10070*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Preprint*, arXiv:1802.00400.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong lu. 2015. Gnormplus: An integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015:918710.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval.

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

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

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.

A Appendix

The appendix is organized into three parts:

- The detailed method implementation is presented in Appendix A.1.
- Additional results on the evaluation are presented in Appendix A.2.
- Discussion on LLM for candidate generation A.3

A.1 Detailed Experimental Setup

This section outlines the experimental setup.

A.1.1 Prompts

Figures 3 and 4 illustrate the prompts used for the accuracy and recall tasks, respectively.

Elements enclosed in *{italics}* represent arguments passed to the prompt, which are detailed below:

- `number_candidates`: The number of candidates provided to the model.
- `mention`: The specific mention that needs to be correctly linked.
- `context`: The surrounding context in which the mention appears.
- `candidates`: The relevant data for all candidates provided to the model.
- `example_answers`: An example of valid answers.

A.1.2 Context window size

A 64-word context window was used, evenly split around each mention. This setup offered the best balance between performance and runtime.

A.1.3 Impact of the number of candidates in the prompt

We evaluated candidate set sizes ranging from 5 to 64 and found that using 10 to 20 candidates yielded the best performance. A set of only 5 candidates resulted in the exclusion of many correct options, whereas increasing the number to 50 introduced an excess of information, reducing precision as the LLM struggled to focus on relevant details, consistent with findings by (Liu et al., 2023).

Ultimately, the effectiveness of candidate selection depends on the ranking of correct entities. If a substantial portion of correct Concept Unique Identifiers (CUIs) are positioned beyond the top 20, increasing the number of candidates can enhance performance. However, in most cases—particularly for contextualized entity linking (EL) models—the proportion of correct candidates ranked below the top 20 is relatively low.

Figure 6 and 7 show the accuracy versus runtime for varying numbers of candidates in the prompt on GNormPlus and NLM-Gene dataset.

In the majority of configurations (dataset - CG Model - LLM), using 10 and 20 candidates yields the best performance.

A.1.4 Impact of the number of examples in the prompt

We also experimented with different numbers of examples in the prompt to evaluate their impact on performance.

Figure 8 and 9 presents the accuracy versus running time for varying numbers of examples k in the prompt on NCBI-Disease and GNormPlus dataset.

The optimal number of examples to include in the prompt varies depending on the configuration (dataset - CG Model - LLM). For instance, GNormPlus performance consistently improved as k increased, while NCBI-Disease performed best with k=3.

A.1.5 Use of separate prompts for accuracy and recall

The reported results in Table 2 and 3 for accuracy (recall@1) were all calculated independently, meaning the LLM was asked to return only the single most likely CUI for each query. Indeed, we noticed that the score for accuracy was higher when the LLM is tasked with retrieving a single candidate compared to ranking several candidates.

A.2 Additional results

This section presents additional results, including error analysis, data leakage risk assessment, a study of prompting strategies, statistical significance testing, and runtime evaluation.

A.2.1 Error Analysis

Table 4 presents the distribution of errors between Candidate Generation (CG) and Named Entity Disambiguation (NED) for each model. CG failure is defined as the absence of the correct CUI within the top-k generated candidates (in this case, k = 20). NED failure is defined as an instance where the

System Instruction
You are a professional data annotator and curator.
Your task is to identify the correct entity for a given mention based on the provided context and the descriptions of {*number_candidates*} candidate entities.

User
Here are a few examples:
{*topk_examples*}

This is the specific mention that needs to be linked to the correct entity:
{*mention*}

This is the context where the mention appears:
{*context*}

These are the candidate entities to choose from:
{*candidates*}

You MUST PROVIDE an ANSWER among the candidates
Use step-by-step reasoning but do not add provide any explanations to me. I only want the final answer.

Return the results in JSON format with just the CUI.
For instance {*example_answers*} are valid answers.

Figure 3: Prompt for the accuracy task, outputting only the best candidate CUI.

System Instruction
You are a professional data annotator and curator.
Your task is to rank the candidate entities from best to worst for a given mention based on the provided context and the descriptions of each candidate entities.

User
Here are a few examples:
{*topk_examples*}

This is the specific mention that needs to be linked to the correct entity:
{*mention*}

This is the context where the mention appears:
{*context*}

These are the candidate entities to choose from:
{*candidates*}

Rank the top {*recall_k*} candidate entities from best to worst.
Use step-by-step reasoning but do not add provide any explanations to me. I only want the final answer.

Return the results in JSON format as a list of CUIs ["CUI1", "CUI2", "CUI3", ...].
For instance {*example_answers*} are valid answers.

Figure 4: Prompt for the recall task, outputting only the best candidates CUI.

	NCBI-Disease		**GNormPlus**		**NLM-Chem**		**NLM-Gene**		**MM-ST21PV**	
	CG	ED	CG	ED	CG	NED	CG	ED	CG	ED
SapBERT	0.307	0.693	0.167	0.833	0.451	0.549	0.186	0.814	0.508	0.492
ArboEL	0.695	0.305	0.823	0.177	0.654	0.345	0.503	0.497	0.598	0.402

Table 4: Failure stage of the entity linking models.

top-ranked candidate is incorrect, when the correct CUI is indeed present in the list of plausible candidates. SapBERT is clearly better at identifying the correct alias than finding the correct candidate, as evidenced by its lower CG error and higher NED error. In this scenario, leveraging an LLM for the disambiguation step can significantly enhance performance, as demonstrated by the big improvement in gene-centric datasets.

A performance plot in low-data slices (training overlap, long-tail entities, no alias match, few aliases, zero-shot) before and after LLM disambiguation for NCBI-Disease, GNormPlus, NLM-Chem and NLM-Gene is shown in Figure 10.

ArboEL is observed to be more sensitive to training overlap and long-tail entities compared to SapBERT. ArboEL benefits greatly from prior exposure to similar examples in the training set across all datasets, while its performance deteriorates in zero-shot settings where either mentions or entities have not been encountered before (e.g., NLM-Chem, NCBI-Disease, NLM-Gene). When the mention does not match any alias of the correct CUI (no alias match), performance degradation is observed only in NCBI-Disease and NLM-Chem. Similarly, when the correct CUI has fewer than five aliases (<5 aliases), the effect is primarily noticeable in NCBI-Disease. Incorporating LLM-based disambiguation appears to mitigate these performance disparities across different data slices, leading to a more uniform performance across all datasets. This trend is consistent for both GPT-4o

System Instruction
You are a professional data annotator and curator.
Your task is to identify the correct entity for a given mention based on the provided context and the descriptions of {*number_candidates*} candidate entities.

User
Here are a few examples:
{*topk_examples*}

This is the specific mention that needs to be linked to the correct entity:
{*mention*}

This is the context where the mention appears:
{*context*}

These are the candidate entities to choose from:
{*candidates*}

You MUST PROVIDE an ANSWER among the candidates

Use step-by-step reasoning :
- Analyze the mention within the provided context to capture its intended meaning
- Evaluate each candidate entity based on its name, aliases, definition, and other attributes
- Pay special attention to semantic similarity whether the entity aligns with the mention's usage in context
- Discard entities that are inconsistent with the mention's meaning in context
- Among the remaining candidates, check the most precise match by considering finer details such as common usage and specific associations
- If multiple candidates remain viable, prioritize based on confidence alignment with context

Once you have reached a final decision, return the answer in the following JSON format.
For instance {*example_answers*} are valid answers.

Figure 5: Prompt for the accuracy task, with reasoning.

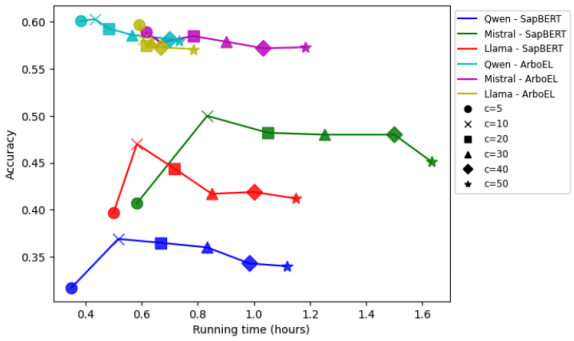

Figure 6: Accuracy vs running time for varying number of candidates in the prompt. Dataset : GNormPlus

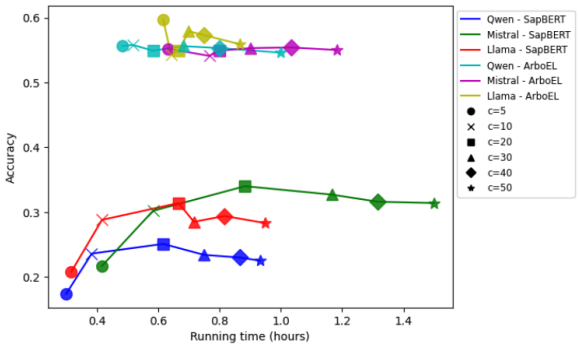

Figure 7: Accuracy vs running time for varying number of candidates in the prompt. Dataset : NLM-Gene

and Qwen, indicating that LLM-based disambiguation helps align slice-specific performance with the overall model performance.

A.2.2 Risk of data leakage

To provide more relevant context during inference, we selected the top-k most similar examples based on document-level similarity. While this approach improves contextual alignment, it also introduces the risk of data leakage (cases where the prompt includes the test mention). Such leakage can simplify the disambiguation task for the LLM.

To assess the extent and effect of this phenomenon, we analyzed how often it occurred and how it influenced model performance in Table 5. Our findings confirms that LLM indeed performs better when the prompt includes the test mention. In gene-centric datasets, the performance gain is more pronounced but these situations is very rare. In contrast, for datasets like NCBI-Disease, where such cases are more common, the performance difference is less significant.

We chose to retain these examples in the evaluation because such scenarios can plausibly arise in real-world applications.

A.2.3 Various prompts

The performance of three different prompts was evaluated using Qwen2.5 7B as base model :

- 1) A prompt designed to generate only the final candidate CUI without any explanation, enabling faster inference. (Figure 3)
- 2) A step-by-step reasoning prompt that first provides detailed reasoning before generating the final result, similar to chain-of-thought (Wei et al., 2023). (Figure 5)
- 3) A variation of the first prompt that utilizes the reasoning model DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025).

The results are presented in Figure 11. Among the evaluated prompting strategies, the simplest

	NCBI-Disease		GNormPlus		NLM-Chem		NLM-Gene	
	✓	✗	✓	✗	✓	✗	✓	✗
SapBERT + GPT-4o	0.842	0.767	0.862	0.742	0.880	0.853	0.582	0.492
SapBERT + Qwen2.5	0.838	0.773	0.661	0.360	0.871	0.836	0.363	0.245
Exact Test Mention present in few-shot	22.9%	77.1%	1.1%	98.9%	3.2%	96.8%	1.6%	98.4%

Table 5: Performance difference and frequency between cases with and without exact matches between prompt examples and the test mention.
✓: The test mention is in the few-shot examples / ✗ : The test mention is not in the few-shot examples

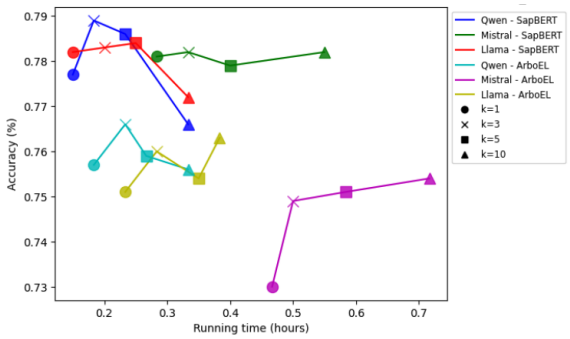

Figure 8: Accuracy vs Runtime for varying number of examples in the prompt. Dataset : NCBI-Disease

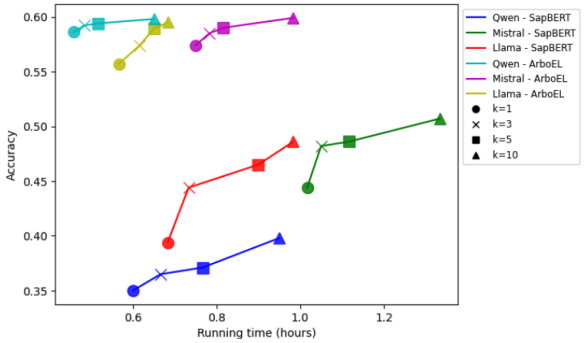

Figure 9: Accuracy vs Runtime for varying number of examples in the prompt. Dataset : GNormPlus

prompt—one that directly outputs the CUI without additional details—demonstrated the best performance across all datasets for this model. This approach not only proved to be significantly faster than the other two methods but also achieved higher accuracy. We hypothesize that this superiority stems from the nature of the task: the model can directly extract and infer the correct CUI from the given information without the need for multi-step reasoning. By avoiding unnecessary intermediate steps, it reduces the risk of error propagation and minimizes computational overhead.

A.2.4 Statistical significance tests

In Table 6 , we used McNemar's paired test to compute the p-values for the differences in accuracy between the base model and with LLMs across all evaluated datasets. The p-values determine whether the performance differences within each LLM are statistically significant across different datasets.

SapBERT combined with an LLM consistently show extremely low p-values across all datasets, indicating that the observed positive difference is statistically significant for all LLMs when applied on SapBERT.

These findings are consistent with the results presented in Section 3.

	SapBERT + Llama3	SapBERT + Mistral	SapBERT + Qwen2.5
NCBI-Disease	$9.14 \cdot 10^{-4}$	$4.51 \cdot 10^{-3}$	$5.38 \cdot 10^{-6}$
GNormPlus	$4.63 \cdot 10^{-175}$	$1.22 \cdot 10^{-217}$	$4.64 \cdot 10^{-132}$
NLM-Chem	$2.76 \cdot 10^{-212}$	$2.46 \cdot 10^{-158}$	$3.02 \cdot 10^{-188}$
NLM-Gene	$2.95 \cdot 10^{-152}$	$9.84 \cdot 10^{-163}$	$5.81 \cdot 10^{-102}$
MM-ST21PV	0.0	$8.30 \cdot 10^{-288}$	0.0

Table 6: P-value of Accuracy for all evaluated datasets and different LLMs using base model SapBERT.

A.2.5 Runtime

Figure 12 show the running time of various LLMs across different datasets, using SapBERT-generated candidates for Accuracy task.

Mistral, as the largest model, has the highest computational demand and requires the longest runtime. Running it on the MM-ST21PV dataset (31,827 evaluated mentions) took nearly a full day, underscoring the significant time requirements of this approach. As the dataset size increases, this can become prohibitively long, making this approach challenging for larger-scale applications.

However, this remains faster than ArboEL, which required 20 days to train on MM-ST21PV.

All experiments were run on a single Nvidia A40 GPU using vllm framework (Kwon et al., 2023).

A.3 LLM for candidate generation

Current LLMs face several limitations for direct candidate generation abilities. First, they lack explicit access to ontology databases, often leading

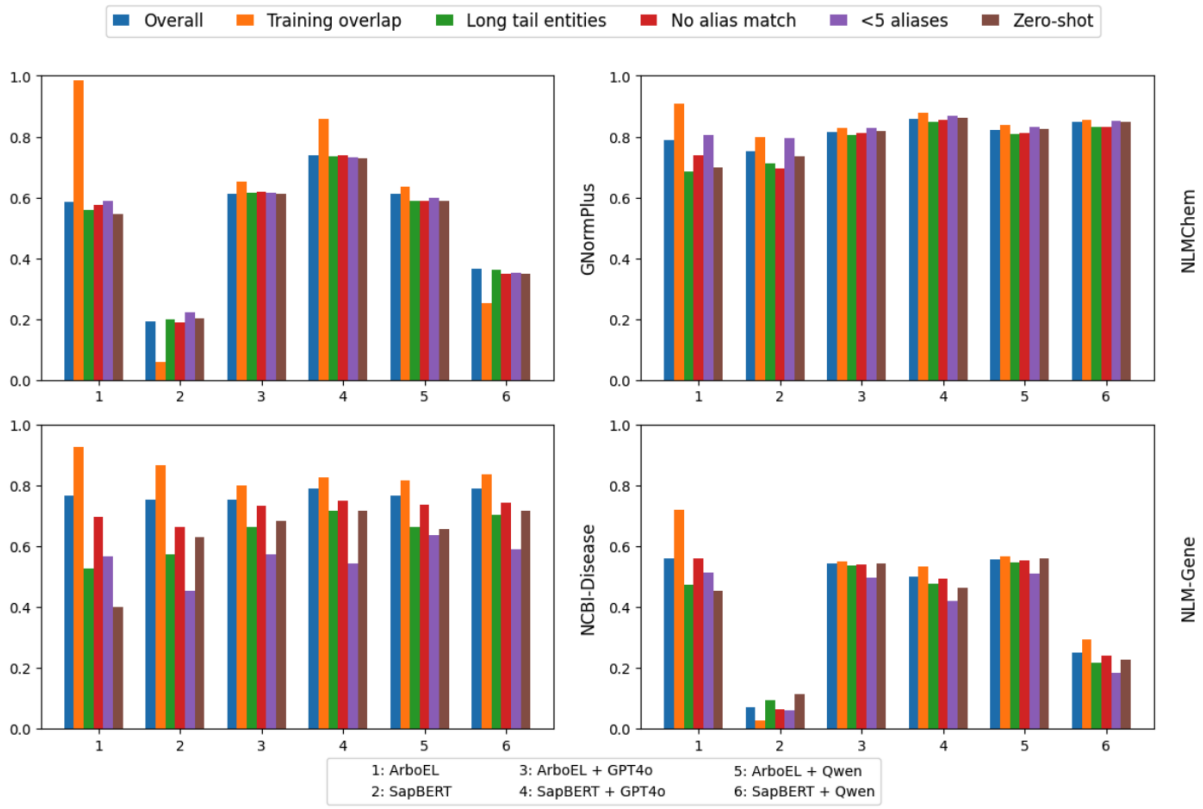

Figure 10: Performance across different data slices: training overlap, long-tail entities, no alias matches, few aliases, and zero-shot cases—both before and after LLM disambiguation.

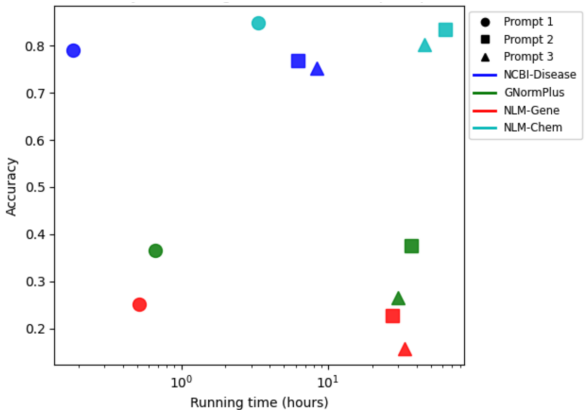

Figure 11: Accuracy vs. Runtime. Comparison of three prompting strategies across four datasets

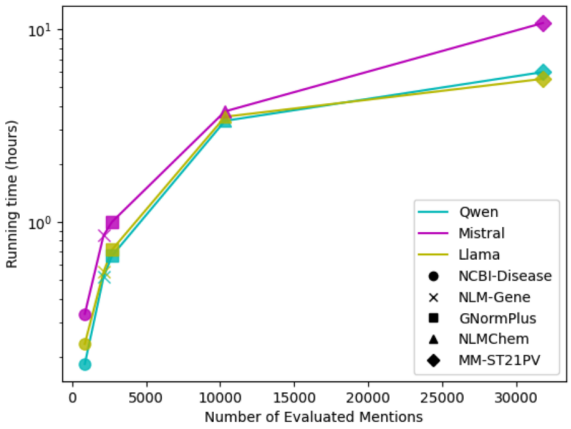

Figure 12: Runtime vs Number of Mentions for Accuracy task across Different LLMs. Base model = SapBERT

to hallucinations when generating candidates. Second, ontologies such as UMLS for MM-ST21PV contain over two million concepts, making it infeasible to include all candidates in the prompt due to context window constraints and computational overhead. These challenges make LLMs currently unreliable for direct large-scale candidate generation. Training models specifically for this task such as BioBART (Yuan et al., 2022a) or BioGenEL (Yuan et al., 2022b) could address the issue, but this lies beyond the scope of our work.