MedLinkDE – MedDRA Entity Linking for German with Guided Chain of Thought Reasoning

Roman Christof[1,2], Farnaz Zeidi[2], Manuela Messelhäußer[3], Dirk Mentzer[3], Renate König[2], Liam Childs[2], Alexander Mehler[1]

[1]Text Technology Lab, Goethe-University Frankfurt,
[2]Host-Pathogen Interactions, Paul Ehrlich Institute,
[3]Safety of Biomedicines and Diagnostics, Paul Ehrlich Institute,
Correspondence: roman.christof@stud.uni-frankfurt.de

Abstract

In pharmacovigilance, effective automation of medical data structuring, especially linking entities to standardized terminologies such as MedDRA, is critical. This challenge is rarely addressed for German data. With MedLinkDE we address German MedDRA entity linking for adverse drug reactions in a two-step approach: (1) retrieval of medical terms with fine-tuned embedding models, followed (2) by guided chain-of-thought re-ranking using LLMs. To this end, we introduce RENOde, a German real-world MedDRA dataset consisting of reportings from patients and healthcare professionals. To overcome the challenges posed by the linguistic diversity of these reports, we generate synthetic data mapping the two reporting styles of patients and healthcare professionals. Our embedding models, fine-tuned on these synthetic, quasi-personalized datasets, show competitive performance with real datasets in terms of accuracy at high top-n recall, providing a robust basis for re-ranking. Our subsequent guided Chain of Thought (CoT) re-ranking, informed by MedDRA coding guidelines, improves entity linking accuracy by approximately 15% (Acc@1) compared to embedding-only strategies. In this way, our approach demonstrates the feasibility of entity linking in medical reports under the constraints of data scarcity by relying on synthetic data reflecting different informant roles of reporting persons.

1 Introduction

In the medical domain, particularly in pharmacovigilance, the volume of data continues to grow rapidly (European Medicines Agency (2024), U.S. Food and Drug Administration (2024)). A critical task in processing this vast amount of safety data is structuring and standardizing the reported information (de Oliveira et al., 2021). This includes mapping medical terms to standardized medical ontologies, such as MedDRA, SNOMED-CT and UMLS[1], to ensure consistency and interoperability. Given the increasing volume of data, automation or semi-automation is essential to reduce processing time and allow professionals to focus on critical decision-making tasks (Salvo et al., 2023).

Reaction As Reported	Kurzatmigkeit bei Anstrengung *Shortness of breath during exertion*
MedDRA PT Term/Code	Belastungsdyspnoe, 10013971 *Dyspnoea exertional*

Figure 1: MedDRA entity linking example with original German text and *English translation*.

This paper focuses on MedDRA[2] (Medical Dictionary for Regulatory Activities), a standardized medical terminology widely used in pharmacovigilance to encode ADEs (adverse drug events). Its hierarchical structure provides a consistent framework for reporting and analyzing ADEs across different regulatory bodies and healthcare institutions. Accurate entity linking to MedDRA is crucial for ensuring that ADE reports are correctly interpreted, utilized in pharmacovigilance workflows and downstream tasks such as signal detection (Bansal et al., 2024). In the context of MedDRA entity linking, the process involves mapping an ADE to a corresponding MedDRA term or code (see Figure 1). It presents a unique challenge due to variations in terminology, synonyms, linguistic complexity, and the vast number of MedDRA entries that must be accurately mapped (Schroll et al., 2012).

In this paper, we address the challenge of German MedDRA entity linking for ADEs by imple-

[1]The Unified Medical Language System (UMLS) includes SNOMED-CT and partially mappings to MedDRA, but does not contain the full MedDRA terminology. https://www.nlm.nih.gov/research/umls/index.html

[2]MedDRA® trademark is registered by IFPMA on behalf of International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). https://www.meddra.org/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31570–31582
November 4-9, 2025

menting a two-step entity linking process: (1) retrieval using fine-tuned embedding models, where cosine similarity is used for similarity-based matching, and (2) guided CoT re-ranking with a large language model (LLM) to refine final entity links. Our goal is to improve entity linking performance by leveraging synthetic data and advanced ranking techniques. Our contributions are as follows:

1. We introduce RENOde (*REaction NOrmalization de*), a novel real-world test dataset for German MedDRA entity linking.

2. We propose a synthetic data generation approach for fine-tuning embedding models, incorporating personae-based text generation to reflect different linguistic styles used by patients and healthcare professionals.

3. We explore LLM-based re-ranking using a guided CoT mechanism and compare it to LLM-based embedding models, simple prompting strategies, and classical CoT-based re-ranking.

By combining embedding-based retrieval and advanced LLM-based ranking techniques, we aim to improve the accuracy and efficiency of linking German MedDRA entities, ultimately contributing to better pharmacovigilance and ADE reporting. The article is organized as follows: Section 2 provides an overview of the biomedical entity linking research landscape and existing datasets. Section 3 describes the real and synthetic datasets developed for this study and details our retrieve-rerank approach incorporating guided CoT reasoning. Section 4 presents experimental results and a manual error analysis. Section 5 discusses the results, while Section 6 outlines potential directions for future research. We conclude by highlighting the main contributions of the paper (Section 7) and discussing its limitations (Section 8).

2 Related Work

2.1 Methods for Biomedical Entity Linking

Various dictionary-based approaches have been developed to normalize biomedical text, including MetaMap (Aronson and Lang, 2010), cTAKES (Savova et al., 2010) and MagiCoder (Aronson and Lang, 2010). These approaches struggle with synonyms and variations in expression (Leaman et al. (2013), Doğan et al. (2014)).

To address these limitations, ML-based approaches have been introduced. TaggerOne (Leaman and Lu, 2016) and MedDRA Tagger (Humbert-Droz et al., 2022) take a hybrid approach by integrating it with rule-based methods. Recent studies go further in this direction and use language models to extract conceptual embeddings and compute semantic similarity scores between entity mentions and their corresponding standardized names. For example, BioSyn (Sung et al., 2020) uses BioBERT, while MedLinker (Loureiro and Jorge, 2020) additionally uses SciBERT and ClinicalBERT as embedding models. CONORM-EN (Yazdani et al., 2023) extends this by applying a dual-transformer retrieval approach with Dynamic Context Refining. Zhang et al. (2022) introduce KrissBERT, a contextual mention encoder trained with contrastive learning and self-supervision, which retrieves candidate entities and re-ranks the top-K using cross-attention mechanisms. BioLORD-2023 (Remy et al., 2024), a fine-tuned cross-lingual BERT-based model, enhances biomedical entity linking by leveraging LLM-generated concept definitions, contrastive learning, and self-distillation. BERGAMOT (Sakhovskiy et al., 2024) integrates structural knowledge of the ontologies by employing Graph Neural Networks (GNNs) to model entity relations through graph-based embeddings. However, challenges remain for contextual embedding models in the case of ambiguous entity mentions and fine-grained differences (Kartchner et al. (2023), Garda et al. (2023)).

LLM-based re-ranking has emerged to account for these subtle differences. For example, Zhu et al. (2022) explore LLM-based re-ranking using prompt learning to account for variability and ambiguity in entity linking. Similarly, Shlyk et al. (2024) integrate retrieval-augmented entity linking (REAL) with LLMs, using GPT-3.5 to re-rank top-K candidates based on cosine similarity. Yan et al. (2025) propose a pre-training strategy that incorporates linearized knowledge graph (KG) triples into a generative LLM, improving entity linking accuracy by leveraging semantic relationships from UMLS. Rouhizadeh et al. (2025) use fine-tuned LLMs as dense retrievers and rerankers. Instead of using the LLM as an embedding model, we use it in an instruction-following manner.

With the introduction of CoT reasoning (Wei et al., 2022), some studies have combined LLM prompting with CoT to improve disambiguation and retrieval accuracy. Wang et al. (2023a) ex-

plore CoT-based in-context learning for biomedical entity linking, integrating ontology-driven prompting to improve entity understanding. Similarly, we inject MedDRA ontology information into the prompt to improve the accuracy of re-ranking retrieved entities. PromptLink (Xie et al., 2024) builds on the CoT framework by incorporating a self-consistency loop to further refine accuracy. In addition, Liu et al. (2024) introduce OneNet, which extends the CoT methodology with CoT Exemplar Pooling and an Adaptive CoT Selector to improve candidate ranking by effectively combining contextual cues and prior knowledge. We extend this CoT approach by guiding the CoT through LLM distilled human coding guidelines for MedDRA.

While most studies focus on linking biomedical entities in English, some studies focus on German data. Becker and Böckmann (2016) apply a dictionary-based method to match German-translated terms to UMLS and SNOMED CT. More recent work explores ML-based approaches. Mustafa et al. (2024) develop a German biomedical entity linking model using UMLS on Wikidata, comparing ScispaCy, SapBERT, M3, and Jina embeddings, and releasing a fine-tuned SapBERT-DE model. Similarly, Idrissi-Yaghir et al. (2024) study German clinical and biomedical language models, but framing entity linking as a multi-label classification task. Their approach includes continuous pre-training and fine-tuning of BERT-based models such as MedBERTde, GBERT-Clinical, and GeBERTa-Clinical.

Our approach also uses embedding models to compute the semantic similarity between reported reactions and MedDRA terms. However, we optimize embeddings via contrastive learning on synthetic data reflecting the perspectives of patients and healthcare professionals. In a second step, we apply a re-ranking strategy using an LLM that incorporates ontology knowledge and guidelines for human MedDRA coders to guide the model's decision making during CoT.

2.2 Datasets for Biomedical Entity Linking

Several human-annotated English biomedical datasets exist, including MedMentions (Mohan and Li, 2019), a UMLS-linked dataset from PubMed abstracts. SMM4H-2020 (Gonzalez-Hernandez et al., 2020) provides a manually curated dataset where adverse event (AE) mentions in tweets have been mapped to MedDRA. The Biomedical Entity Linking Benchmark (BELB) dataset (Garda et al., 2023) covers genes, diseases, chemicals, species, and cell lines, all mapped to UMLS. Additionally, CT-ADE (Yazdani et al., 2025) is a clinical trial dataset containing ADEs mapped to MedDRA. Given the challenges of collecting biomedical text, such as privacy concerns and the time-consuming nature of manual annotation (Kartchner et al., 2023), automated and semi-automated data generation approaches have emerged. WikiMed-DE (Wang et al., 2023b) is a silver-standard dataset designed for German biomedical entity linking, constructed by automatically mapping German Wikipedia articles to UMLS entities. Wang et al. (2024a) generate an HPO-based synthetic dataset for rare disease normalization and fine-tuned LLaMA 2-7B using template-based corpus generation to enhance entity linking. Likewise, Remy et al. (2024) train BioLORD-2023 with LLM-generated training data, incorporating GPT-3.5-generated concept definitions and contextual descriptions. Similarly, Yuan et al. (2022) introduce GenBioEL, a generative biomedical entity linking framework that employs knowledge base (KB)-guided pre-training and synonym-aware fine-tuning. Shlyk et al. (2024) propose REAL, a retrieval-augmented entity linking framework that generates synthetic entity variations, including misspellings, abbreviations, and paraphrases, by prompting GPT-3.5. Sasse et al. (2024) use LLaMA-2-13B to paraphrase UMLS disease concepts and mapped to MedDRA terms via fuzzy matching. All of these existing datasets have limitations for our specific tasks: They focus primarily on English data and ontologies, rarely including the MedDRA ontology, and when they do, the types of texts analyzed (mostly scientific publications) are usually different from our use case, i.e. patient and physician reports. Our REN-Ode dataset fills these gaps by providing German MedDRA-annotated data (reactions as reported in German and German MedDRA terms) specifically derived from patient and healthcare professional reports. Furthermore, we supplement these data with synthetically generated training data that reflect the perspective of the reporting individuals.

3 Materials and Methods

We start with introducing the datasets created for training and evaluation. We then describe the methods used, including the fine-tuning of the embedding models used for retrieval and our guided CoT re-ranking approach. The complete pipeline is il-

	Dataset	#Entries	Avg. Chars	#LLT Codes	#PT Codes
Train	RENOdeTrain-Real	15,047	61.35	3,921	2,165
	RENOdeTrain-Synth	22,221	30.53	3,341	2,023
Dev	RENOdeDev	200	28.90	171	144
Test	RENOdeTest	200	31.48	169	135
	Mantra-MedDRA-DE	276	16.39	/	231

Table 1: Comparison of test dataset statistics with unique codes and length in characters.

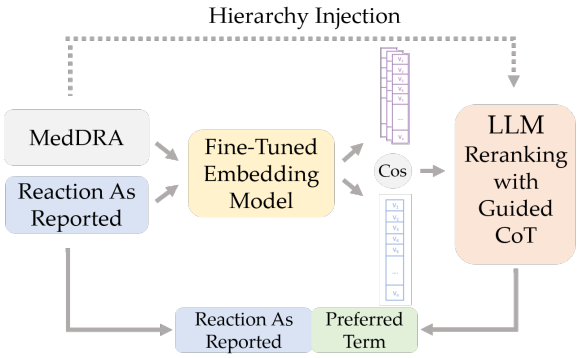

Figure 2: The MedLinkDE pipeline

.

lustrated in Figure 2.

3.1 A Novel Real-world Dataset: RENOde

Our real-world dataset RENOde (*REaction NOrmalization de*) consists of pairs of reported adverse events and their corresponding MedDRA terms at the Lowest Level Term (LLT) level. In some cases, an alternative MedDRA term is provided when multiple valid mappings exist. RENOde includes three dataset splits: (1) RENOdeTest serves as a test set without identical pairs (i.e., cases where the reported adverse event exactly matches the MedDRA term), (2) RENOdeDev is a development dataset also containing identical pairs, as well as (3) RENOdeTrain-Real, a large training dataset. RENOdeTest was double-checked, while the development and training datasets were annotated by a single annotator following the standard MedDRA guidelines. Note that due to database specifications in the training dataset, a reported reaction may include multiple MedDRA terms, but only one will be coded at a time. An overview of the number of entries, characters and entities is given in Table 1.

The dataset originates from the spontaneous reporting system of the Paul Ehrlich Institute[3], which consists of adverse event reports from both healthcare professionals and patients. These two sources differ significantly in their linguistic characteristics. Physician reports typically use precise medical terminology and structured phrasing, while patient reports often lack formal medical terminology and instead use simple, colloquial language and subjective descriptions. This contrast in reporting styles influences how adverse events are expressed. For example, a physician may report "*Schlafstörung/insomnia*", while a patient may describe it with the phrase "*konnte nicht schlafen/could not sleep*". In addition, the reports consist of isolated phrases without a broader context, which is common in pharmacovigilance databases.

Our dataset includes reports with a significant proportion related to COVID-19 vaccine cases, which introduces a potential bias towards specific adverse events. To our knowledge, this is the first German dataset for MedDRA-based entity linking in the context of pharmacovigilance databases.

In addition to RENOde, we use the publicly available Mantra-GSC dataset[4] (Kors et al., 2015), which contains Medline abstract titles, drug labels, and biomedical patent claims in multiple languages, including English, French, German, Spanish, and Dutch. For our purposes, we extract only the German subset. Since the annotations in Mantra-GSC are provided in UMLS, we use the *Concept Unique Identifier* (CUI) to map entities to their corresponding MedDRA PTs (Preferred Term) where possible. Since UMLS integrates several medical ontologies, not every CUI corresponds to a MedDRA term. As the proportion of German entries that can also be mapped to MedDRA in the overall dataset is small, we merge all three sources – Medline abstract titles, drug labels, and biomedical patent claims – into a unified test dataset that we call Mantra-MedDRA-DE. It has a total of 276 records (see Table 1).

3.2 Synthetic Dataset

A central contribution of our work is to generate and provide a synthetic dataset that reflects the two roles in the context of medical event reporting: patients and healthcare professionals. To this end,

[3] https://www.pei.de

[4] We use the Mantra GSC exclusively as a test dataset, in accordance with its CC BY-NC 4.0 license.

we generate RENOdeTrain-Synth to evaluate how well the models perform in scenarios where real data is limited or unavailable. This will allow us to compare model performance on real and synthetic datasets to assess the usefulness, reliability, and potential limitations of using LLM-generated text in MedDRA entity linking. To maintain consistency with real data, we follow the same MedDRA LLT distribution when generating synthetic phrases. To ensure that the generated data is consistent with a realistic pharmacovigilance scenario, we model two personae:

1. First, the *patient perspective*, where phrases are formulated in layman's terms, using non-technical language that reflects how patients typically describe symptoms, conditions or reactions. Regarding this group of persons, the focus is on subjective experiences, sensations and emotional expressions rather than clinical terminology.

2. Second, the *perspective of healthcare professionals*, where phrases are generated using formal medical language that reflects how professionals name, describe, and discuss diagnoses, symptoms, and treatments.

To enforce this binary distinction, we include a one-shot example in the prompt to demonstrate the expected language style for each persona (see Appendix A for the prompt). Additionally, we apply guided decoding in vLLM[5] to structure the output in JSON format for consistent extraction. We set the temperature to 0.8 to encourage diverse expressions. However, since guided decoding is not always perfectly reliable due to JSON format errors, we attempt generation five times for each MedDRA term. If extraction fails on all five attempts, we discard the term. This occurs in approximately 15% of the cases.

For synthetic data generation, we use Llama-3.1-Nemotron-70B-Instruct-HF in combination with vLLM for efficient and scalable generation. The synthetic data generation took approximately 14 hours and was executed (as well as the later fine tuning and re-ranking) on a cluster of four NVIDIA A100 GPUs (GA100, 20b5 rev a1, 80GB HBM2e). See Table 2 for an example of the synthetic data generated and Appendix B (Table 4) for a more detailed statistic of the RENOdeTrain-Synth dataset.

[5]An open-source inference engine designed for fast, efficient generation with LLMs: https://github.com/vllm-project/vllm

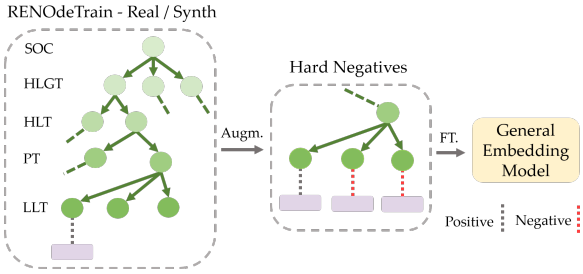

Figure 3: Fine-tuning approach for embedding model. Abbrevations: SOC=System Organ Class; HLGT=High Level Group Term; HLT=High Level Term; PT=Preferred Term; LLT=Lowest Level Term; Augm=Augmented; FT=Fine-Tuned

3.3 Retrieving with Embedding Models

3.3.1 Setup

Both reported reactions and all MedDRA terms at the LLT level are embedded using a pretrained model, specifically the M3 embedding model[6] (Chen et al., 2024). Cosine similarity is then computed between the embeddings to identify the most relevant matches. To optimize retrieval efficiency, we employ k-nearest neighbors search.

3.3.2 Fine-Tuning Embedding Model

We use BGE-M3 for fine-tuning due to its strong Acc@100 performance (see section 4 for a comparison of different pre-trained models). Our primary goal at this stage is to maximize Acc@100 to ensure a solid foundation for the next step, re-ranking, rather than focusing solely on Acc@1.

Fine-tuning is conducted using both real-world and synthetic data. We apply contrastive fine-tuning with hard negative examples, where for each given positive MedDRA term, we randomly select another LLT under the same PT as a hard negative, as shown in Figure 3, and assign the pair a similarity score of 0.9 to indicate that, although incorrect, it is still semantically close. We focus exclusively on hard negatives because general embedding models often struggle with fine-grained distinctions between terms that have very similar meaning. The fine-tuning is done using the BGE fine-tuning script[7] with the following default configuration: 2 epochs, batch size of 2, learning rate of 1e-5 and 10% warm-up ratio.

[6]https://huggingface.co/BAAI/bge-m3

[7]https://github.com/FlagOpen/FlagEmbedding/tree/master/examples/finetune/embedder

MedDRA LLT	Health Professional	Patient
Beweglichkeit vermindert (*Mobility decreased*)	Eingeschränkte Gelenkbeweglichkeit (*Restricted joint mobility*)	Ich kann mich kaum bewegen (*I can hardly move*)
Schwellung der Zunge (*Swollen tongue*)	Linguale Ödeme (*Lingual edema*)	Zunge fühlt sich geschwollen an (*My tongue feels swollen*)
Bauschmerzen (*Abdominal pain*)	persistierende Unterbauchschmerzen (*Persistent lower abdominal pain*)	Bauch tut weh (*My tummy hurts*)

Table 2: Example of reactions as reported in synthetic data for health professional and patient perspective with *English translations*.

3.4 Re-ranking with Guided CoT Reasoning

To improve the accuracy of MedDRA term selection, we implement a guided CoT reasoning approach that incorporates established coding guidelines. We base our reasoning on the "ICH-Endorsed Guide for MedDRA Users"[8], which provides best practices for term selection, coding conventions, and handling ambiguous cases. To increase efficiency and applicability within our model, we also use Llama-3.1-Nemotron-70B-Instruct-HF to generate an abbreviated version of the lengthy original guidelines (62 pages), which is then refined by a human to ensure that key decision-making principles are explicitly incorporated into the re-ranking process. By distilling complex rules into a structured format, the model can prioritize medically and contextually appropriate MedDRA terms, especially when multiple candidate terms have high similarity scores.

Our guided CoT prompting framework consists of four components, as illustrated in Figure 4:

1. First, we provide a general instruction to rerank a given list of terms based on their similarity to a target reaction term.

2. Second, we include a condensed version of the MedDRA coding guidelines, as described above, to serve as structured reasoning steps to guide the model's decision making. All seven reasoning steps that are integrated in the prompt framework are shown in Appendix C.

3. Third, we present three examples to illustrate the desired behavior.

4. Fourth, we present the initial candidate pool consisting of the top 100 MedDRA LLTs retrieved via embedding similarity. For each of these candidates, we inject hierarchical information from the MedDRA ontology, including relations across the LLT, PT, HLT (High Level Terms), HLGT (High Level Group Terms), and SOC (System Organ Classes) levels. Finally, we output a structured JSON object.

In contrast to generic CoT prompting, our method introduces a structured and domain-informed order, resembling a constraint satisfaction process in which the system must reason while adhering to MedDRA-specific constraints. To evaluate the effectiveness of our guided CoT re-ranking approach, we compare it to alternative methods, including LLM-based embedding re-ranking, simple prompting strategies, and classical CoT approaches. This comparative analysis allows us to evaluate whether guided reasoning improves term selection accuracy over more general LLM-based methods. For all re-ranking strategies we use Qwen2.5-14B-Instruct[9] (Yang et al., 2024).

4 Experiments and Evaluation

To evaluate MedLinkDE, we consider models that are relevant in terms of size, currency, and language coverage, with a focus on German and multilingual models including German. For general-purpose embedding models, we include models that have been widely used in previous studies (Mustafa et al. (2024), Rouhizadeh et al. (2025)) and demonstrate robust performance across multiple tasks, specifically BGE-M3 and Multilingual-E5 (Wang et al., 2024b). We also select SapBERT-DE, BioLORD-2023-M, and BERGAMOT because these models are specifically designed for biomedical entity linking while also supporting German or multilingual biomedical text processing. As an additional baseline, we use a rule-based method based on normalized indel similarity using fuzzy string matching to compare reactions as reported and MedDRA terms. We implement this method using RapidFuzz[10]. To

[8] https://admin.meddra.org/sites/default/files/guidance/file/001006_termselptc_r4_24_mar2024.pdf

[9] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[10] https://rapidfuzz.github.io/RapidFuzz/

Model	RENODde200						Mantra-MedDRA-DE					
	@1	@5	@10	@20	@60	@100	@1	@5	@10	@20	@60	@100
Rule-Based	11.39	11.39	11.39	11.39	11.39	11.39	60.07	61.15	61.15	61.15	61.15	61.15
BGE-M3	43.07	64.36	75.25	80.20	**90.59**	**91.09**	67.99	80.22	83.45	86.33	92.09	92.81
Multilingual-E5 (large)	37.63	64.36	71.78	75.74	84.16	90.09	68.71	80.58	83.81	87.77	91.01	92.45
SapBERT-DE	31.68	51.49	58.42	65.84	77.23	80.20	55.76	79.14	85.97	89.21	91.01	93.17
BioLORD-2023-M	**51.49**	**68.81**	**76.73**	**81.68**	87.13	88.61	**83.09**	**88.49**	**91.73**	**93.53**	**95.32**	**96.04**
BERGAMOT	22.28	46.04	52.97	60.40	68.81	71.78	45.32	68.71	78.42	83.09	89.57	92.09
Our (BGE-M3 ft. real data)	**59.90**	**86.14**	**90.10**	**91.58**	95.05	96.04	**78.78**	85.61	90.29	92.07	94.60	95.68
Our (BGE-M3 ft. synth. data)	48.51	70.79	77.72	88.12	94.06	95.54	74.82	**85.97**	91.01	**93.53**	95.32	96.40
Our (BGE-M3 ft. real & synth. data)*	58.91	85.15	89.11	91.09	**95.54**	**97.52**	76.26	**85.97**	**91.37**	92.45	**95.68**	**96.76**
Our (BGE-M3 ft. synth. data) + Guided CoT Prompt	71.78	89.11	91.01	91.59	91.59	91.59	**83.09**	90.65	93.17	93.53	93.53	93.53
Our (BGE-M3 ft. real. data) + Guided CoT Prompt*	**74.75**	89.60	90.59	90.59	91.09	91.09	**83.09**	91.73	93.53	93.53	93.88	93.88
Our (BGE-M3 ft. real & synth. data) +												
LLM Embedding	23.67	43.07	55.45	68.31	89.10	**97.52**	51.80	67.27	73.02	80.94	93.17	**96.76**
Simple Prompt	72.77	90.59	92.57	**94.55**	**95.54**	95.54	80.58	90.29	91.27	92.45	93.88	93.88
Standard CoT Prompt	71.29	89.60	91.58	92.08	93.07	93.56	77.70	87.41	91.37	92.45	93.88	93.88
Guided Prompt	73.76	**92.57**	**94.06**	**94.55**	95.05	95.05	82.73	**92.81**	**94.60**	**94.96**	**94.96**	94.96
Guided CoT Prompt	74.26	91.09	92.57	92.57	92.57	92.57	82.37	90.65	93.53	94.24	94.24	94.24

Table 3: Evaluation of fine-tuned embedding models and re-ranking approaches on RENOdeTest and Mantra datasets with Accuracy@n at PT level. Our best performing MedLinkDE approach marked with *: BGE-M3 ft. real data + Guided CoT prompt re-ranking. Abbreviations: ft=fine-tuned; synth=synthetic.

General Instruction

Based on the input reaction, analyze each term in the provided list to assess its similarity to the reaction. Then rank the terms from most to least relevant. All input reactions must occur in the ranked result. Then output the result in a structured JSON format.

Guided Reasoning Steps

1. Always choose the most specific LLT (Lowest Level Term) …
2. Apply medical judgment when selecting terms …
3. Verify the MedDRA hierarchy …

…

Examples

Example 1
Input reaction: ...
Provided terms: ['...', '...', ...]
Output: {"reaction": "....", "meddra_terms_reranked": {"1": "...", "2": ".", ... }}

Example 2
...

Generate Answer

Input reaction: {reaction}
Provided terms: {meddraterms_with_hierarchy}
Output: Provide the ranking of the terms in JSON format. Ensure the terms are ranked by relevance to the reaction. Exclude the hierarchy from the result list.

Figure 4: Guided prompt framework (for the guided CoT prompt we add 'Think step by step' to the general instruction).

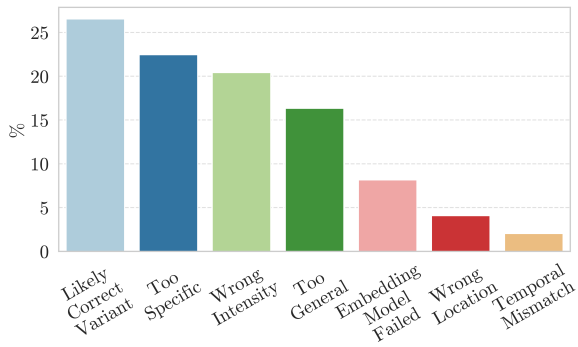

Figure 5: Manual error analysis for guided CoT prompt re-ranking on RENOdeTest.

evaluate the performance at the MedDRA PT level, we map the predicted LLT to the corresponding PT and then compare it to the test dataset. Table 3 shows the results for the baseline models, our fine-tuned embedding models, and comparisons between different re-ranking strategies: For the RENOde dataset, fine-tuning on real data generally outperforms synthetic data at lower Acc@n, though differences diminish at higher Acc@n levels. Combining real and synthetic data shows slightly higher results at Acc@100 with around 1.5%. All re-ranking methods (except LLM embedding) show significant improvement (up to 15%) in Acc@1 using a guided CoT prompt. For the Mantra dataset, fine-tuning with real or synthetic data improves performance by about 10% compared to pre-trained embeddings, with real data performing better at lower Acc@n, while synthetic data slightly outperforms BioLORD-2023-M at Acc@100 by 0.36%. Combining the datasets yields a marginal improvement at higher Acc@n. Unlike the RENOde dataset, re-ranking yields a less pronounced improvement for Mantra. The guided prompt without CoT slightly outperforms the guided CoT prompt. At Acc@1 the performance is slightly below that of BioLORD-2023-M, but at Acc@5 it exceeds

it by about 3%. For both RENODe and Mantra-MedDRA-DE, the guided CoT reranking approach achieves competitive results even when using only synthetic embedding model outputs, compared to using only real data. We performed an error analysis of the guided CoT prompt re-ranking results on the RENOde200 dataset, as shown in Figure 5. It shows that most errors occur when the model selects a MedDRA term that is very similar to the gold standard term, which could also be considered partially correct. In most other cases, the model selects terms that contain overly specific information that is not reflected in the reported reaction. In addition, the model often selects terms with the wrong intensity - either higher or lower - or terms that are too general compared to the reported reaction. Occasionally, re-ranking fails because the correct MedDRA term is missing from the initial list of 100 terms. A more detailed description of the error types can be found in Appendix D.

5 Discussion

Our results demonstrate that MedDRA entity linking can be significantly improved by combining customized embedding fine-tuning with guided (CoT) re-ranking. In particular, our embedding model, fine-tuned on approximately 22k synthetic training pairs, outperforms general biomedical models such as BioLORD-2023-M, which relies on over 100 million concept-definition pairs, for higher Acc@n. This highlights the importance of incorporating the reporter perspective – such as that of patients and clinicians – into synthetic data generation to better account for linguistic variability, rather than relying solely on formal ontology definitions. Our approach was particularly effective on the RENODeTest dataset, which consists of adverse reaction reports from patients and healthcare professionals. Our findings indicate significant potential for scaling this approach to all 60k+ LLTs within MedDRA. With respect to the re-ranking component, the incorporation of re-ranking significantly mitigated the challenges of embedding models with lower Acc@n scores. Our results illustrate that human coding guidelines can be effectively integrated into prompts, either in a guided fashion or supplemented by CoT reasoning. Explicitly embedding these guidelines in prompts effectively guides the reasoning processes of LLMs, seems promising for improving semantic alignment and decision accuracy. The improvements achieved through our re-ranking are substantially larger (around 10%) on RENODeTest compared to the publicly available Mantra-MedDRA-DE dataset. We attribute this to the fact that Mantra-MedDRA-DE primarily consists of short, simple phrases, which make the entity linking task relatively easier. In such cases, embedding-based models like BioLORD already perform reasonably well. In contrast, many real-world clinical texts consist of noisier and more ambiguous phrases under which our guided re-ranking approach demonstrates clearer advantages. Our findings also highlight that existing benchmarks may underestimate the true complexity of MedDRA entity linking tasks.

We conducted a power analysis using the statsmodels Python package[11] to assess the statistical strength of our evaluation. Using the current test dataset, RENOdeTest, which contains 200 entries, we estimated the power of McNemar's test for comparing guided CoT prompt reranking on top-1 accuracy results from two embedding models: one trained exclusively on synthetic data (71.78% accuracy) and the other on real data (74.75% accuracy). The resulting power is 12.3%. To achieve the threshold of 80% power, approximately 2,000 entries would be required. These results suggest a promising direction in using embedding models trained on persona-based synthetic data combined with guided CoT reranking, but also highlight the need for further validation using larger datasets.

6 Future Work

Future work should explore the optimal number of retrieved results to consider during re-ranking. Currently, 100 retrieved candidates are considered, but to determine the ideal number, the optimal number of candidates should be investigated. A low number of candidates could lead to a necessary failure of re-ranking if the correct entity is not in the initial retrieved entity list, while too many candidates could degrade model performance due to information overload. In addition, future work should investigate scenarios where reported reactions need to be mapped to multiple MedDRA terms. The guided CoT approach is particularly beneficial in these cases, as the guidelines and prompt framework can be specifically tailored. Currently, such complex cases are not addressed, which represents a valuable direction for further research. Finally, additional research efforts could evaluate, compare,

[11] https://www.statsmodels.org

and benchmark alternative LLMs for their effectiveness in the re-ranking task, and also examine dedicated reasoning models such as DeepSeek-R1 (DeepSeek-AI et al., 2025).

7 Conclusion

In this paper, we have presented an approach that aims to improve entity linking in the field of German pharmacovigilance using the MedDRA ontology. To this end, we have developed a new dataset, RENOde, which is based on an approach to generating synthetic data using persona-sensitive prompts. Our results show that synthetic data with this specific role-based perspectivity enable fine-tuning of embedding models that achieve competitive Acc@n for high n, comparable to models trained on real data. This shows a new perspective for their effective use for downstream re-ranking tasks, especially in the context of data-poor environments. For re-ranking, we have developed a guided CoT approach that takes into account human MedDRA coding guidelines and the ontology context for the retrieved entities. This combined approach of MedLinkDE effectively improves the accuracy of the re-ranked results and thus the performance of German medical entity linking.

8 Limitations

Our study has several limitations: First, the test dataset includes a limited number of terms with a bias toward the COVID-19 vaccine case, which may limit the evaluation. Second, the synthetic data generation uses only a small subset of all MedDRA terms. Third, the current implementation focuses on German. Fourth, the approach depends on existing entity coding guidelines, which may change over time and vary between institutions. Therefore, scenarios where guidelines need to be generated entirely synthetically require further investigation. Fifth, cases where multiple MedDRA terms apply to a single reported reaction were not considered in the current evaluation. In addition, explicit testing of negations within reported reactions was not performed.

Code and Data Availability

The code is available at `https://github.com/RomanChristof/MedLinkDE` under the MIT License. The RENOde datasets are not publicly accessible due to privacy and license restrictions. What is available, however, is the complete procedure for generating this synthetic data, which generates analogous data given input terms of MedDRA.

Ethics Statement

This study used data that is derived from sources that are not freely accessible to the public. It does not include any direct personally identifiable information or patient-specific details. It only contains reported adverse reactions and corresponding MedDRA terms. The dataset is biased towards COVID-19 vaccine cases and therefore represents only a limited number of adverse events.

Acknowledgments

We would like to acknowledge the collaboration between the Paul Ehrlich Institute (departments of FoG3, headed by Dr. Renate König, and Pharmacovigilance, headed by Dr. Dirk Mentzer) and the TTLab (Text Technology Lab) headed by Prof. Alexander Mehler at the Goethe University in Frankfurt.

References

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.

Dipika Bansal, Beema T Yoosuf, and Muhammed KT Favas. 2024. Role of medical coding in signal detection. In *Signal Analysis in Pharmacovigilance*, pages 95–110. CRC Press.

Matthias Becker and Britta Böckmann. 2016. Extraction of umls® concepts using apache ctakes™ for german language. In *Health Informatics Meets eHealth*, pages 71–76. IOS Press.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Jezer Machado de Oliveira, Cristiano André da Costa, and Rodolfo Stoffel Antunes. 2021. Data structuring of electronic health records: a systematic review. *Health and Technology*, page 1219–1235.

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, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 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*, 47:1–10.

European Medicines Agency. 2024. 2023 annual report on eudravigilance for the european parliament, the council and the commission. Reporting period: 1 January to 31 December 2023.

Samuele Garda, Leon Weber-Genzel, Robert Martin, and Ulf Leser. 2023. Belb: a biomedical entity linking benchmark. *Bioinformatics*, 39(11):btad698.

Graciela Gonzalez-Hernandez, Ari Z. Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, Anne-Lyse Minard, Elena Tutubalina, Zulfat Miftahutdinov, and Ilseyar Alimova, editors. 2020. *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, Barcelona, Spain (Online).

Marie Humbert-Droz, Jessica Corley, Suzanne Tamang, and Olivier Gevaert. 2022. Development and validation of meddra tagger: a tool for extraction and structuring medical information from clinical notes. *medRxiv*.

Ahmad Idrissi-Yaghir, Amin Dada, Henning Schäfer, Kamyar Arzideh, Giulia Baldini, Jan Trienes, Max Hasin, Jeanette Bewersdorff, Cynthia S Schmidt, Marie Bauer, et al. 2024. Comprehensive study on german language models for clinical and biomedical text understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3654–3665, Torino, Italia. ELRA and ICCL.

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

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. OneNet: A fine-tuning free framework for few-shot entity linking via large language model prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13634–13651, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Loureiro and Alípio Mário Jorge. 2020. Medlinker: Medical entity linking with neural representations and dictionary matching. In *Advances in Information Retrieval*, pages 230–237. Springer International Publishing.

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

Faizan E Mustafa, Corina Dima, Juan Ochoa, and Steffen Staab. 2024. Leveraging Wikidata for biomedical entity linking in a low-resource setting: A case study for German. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 202–207, Mexico City, Mexico. Association for Computational Linguistics.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Hossein Rouhizadeh, Anthony Yazdani, Boya Zhang, David Vicente Alvarez, Matthias Hueser, Alexandre Vanobberghen, Rui Yang, Irene Li, Andreas Walter, and Douglas Teodoro. 2025. Large language models struggle to encode medical concepts-a multilingual benchmarking and comparative analysis. *medRxiv*.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643. Association for Computational Linguistics.

Francesco Salvo, Joelle Micallef, Amir Lahouegue, Laurent Chouchana, Louis Létinier, Jean-Luc Faillie, and Antoine Pariente. 2023. Will the future of pharmacovigilance be more automated? *Expert Opinion on Drug Safety*, 22(7):541–548.

Kuleen Sasse, Shinjitha Vadlakonda, Richard E. Kennedy, and John D. Osborne. 2024. Disease entity recognition and normalization is improved with large language model derived synthetic normalized mentions. *Preprint*, arXiv:2410.07951.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Jeppe Bennekou Schroll, Emma Maund, and Peter C Gøtzsche. 2012. Challenges in coding adverse events in clinical trials: a systematic review. *PloS one*, 7(7):e41174.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. REAL: A retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389, Bangkok, Thailand. 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, Online. Association for Computational Linguistics.

U.S. Food and Drug Administration. 2024. Fda adverse event reporting system (faers). Accessed: March 19, 2024.

Andy Wang, Cong Liu, Jingye Yang, and Chunhua Weng. 2024a. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association*, 31(9):2076–2083.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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

Yi Wang, Corina Dima, and Steffen Staab. 2023b. Wikimed-de: Constructing a silver-standard dataset for german biomedical entity linking using wikipedia and wikidata. In *Wikidata@ ISWC*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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 2024, pages 2589–2593. ACM.

Xi Yan, Cedric Möller, and Ricardo Usbeck. 2025. Biomedical entity linking with triple-aware pretraining. In *Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data 2025*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Anthony Yazdani, Alban Bornet, Philipp Khlebnikov, Boya Zhang, Hossein Rouhizadeh, Poorya Amini, and Douglas Teodoro. 2025. An evaluation benchmark for adverse drug event prediction from clinical trial results. *Scientific Data*, 12(1).

Anthony Yazdani, Hossein Rouhizadeh, Alban Bornet, and Douglas Teodoro. 2023. Conorm: Context-aware entity normalization for adverse drug event detection. *medRxiv*, pages 2023–09.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. 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. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tiantian Zhu, Yang Qin, Qingcai Chen, Baotian Hu, and Yang Xiang. 2022. Enhancing entity representations with prompt learning for biomedical entity linking. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4036–4042. International Joint Conferences on Artificial Intelligence Organization.

A Synthetic Data Generation Prompt

The following prompt is used to generate synthetic training data in German for different roles (healthcare professional and patient), based on a given MedDRA term:

```
Erstelle fünf Variationen, wie ein {personae}
diesen MedDRA-Begriff in einer kurzen Phrase
berichten könnte.

Beispiel:
{example}

Gib das Ergebnis für diese MedDRA-Begriffe
{meddraterm} im JSON-Format zurück:


pesonae = ["Arzt", "Patient"]
```

One shot example for patient:

```
MedDRA term: Insomnia
{
    "medra_term": "Insomnia",
    "reactionasreported": {
        "1": "schlaflos",
        "2": "kann nicht schlafen",
      "3": "Schwierigkeiten beim Einschlafen",
        "4": "kann nicht einschlafen",
        "5": "kann nachts kaum schlafen"
    }
}
```

One-shot example for healthcare professional:

```
MedDRA term: Insomnia
{
    "medra_term": "Insomnia",
    "reactionasreported": {
        "1": "Dauerhafte
                Insomnie-Symptomatik",
      "2": "Patient klagt über anhaltende
        Schlaflosigkeit",
        "3": "Persistierende Insomnie",
        "4": "Primäre Insomnie",
        "5": "Chronische Ein- und
        Durchschlafstörung"
    }
}
```

B Synthetic Dataset Statistics

Table 4 shows in more detail the statistics for the synthetic RENOde dataset.

C Guided Reasoning Steps

The following seven guided reasoning steps are integrated in the prompt framework:

1. Always choose the most specific LLT (Lowest Level Term): Choose the MedDRA LLT that most accurately represents the reported information. Example: For 'Abscess on face,' the LLT 'Abscess on the face' is more specific than 'Abscess.'

2. Apply medical judgment when selecting terms: If no exact match is found, choose an existing LLT that best covers the concept.

3. Verify the MedDRA hierarchy: Check the hierarchy above the LLT (e.g., PT, HLT, HLGT, SOC) to ensure the correct meaning. This is especially important for medication errors and product quality issues.

4. Consider all reported information: Choose terms for all reported adverse reactions (ARs/AEs), medication errors, medical history, etc. Example: For 'Abdominal pain, elevated serum amylase and lipase,' select separate terms rather than just 'Pancreatitis.

Personae	#Entries	Avg. Chars	#LLT Codes	#PT Codes
Doctor	7,237	32.98	1,403	1,013
Patient	14,984	28,08	2993	1,860
Total	22,221	30.53	3,341	2,023

Table 4: Comparison of personae statistics in RENOdeTrain-Synth.

Error Type	Description	Example
Embedding Model Failed	The correct MedDRA term is not present in the Acc@100 output from the first step (the embedding model), making a correct re-ranking result impossible.	/
Wrong Location	The linked entity refers to an incorrect anatomical location.	*Injektionsstelle* instead of *Arm*
Wrong Intensity	The entity does not match the severity level of the original term.	*Hypohidrose* instead of *Anhidrose*
Temporal Mismatch	The linked entity represents an incorrect timeframe.	*Verzögert* instead of *Verfrüht*
Too Specific	The entity is overly detailed	*Alopecia Areata* instead of *Apolezie*
Too General	The entity is too broad and lacks necessary specificity.	*Kopfschmerzen* instead of *Kopfschmerzen im Zusammenhang mit einem Verfahren*.
Likely Correct Variant	The linked entity is a plausible but not the gold standard term.	*Parese* instead of *Beweglichkeit vermindert*

Table 5: Error Taxonomy with examples.

5. Diagnoses with/without signs and symptoms: For definitive diagnoses without symptoms, select only the diagnosis. For preliminary diagnoses, prioritize the diagnosis and symptoms.

6. Treat death and outcomes as terms: Death is treated as an outcome, not an AR/AE; choose the appropriate terms if reported.

7. Combine or split terms as needed: Combine terms when meaningful ('Retinopathy due to diabetes' -> 'Diabetic retinopathy'). Split reports into multiple terms if it provides more clinical information (e.g., 'Diarrhea and vomiting' coded separately).

D Error Taxonomy

The following Table 5 presents the error taxonomy used in the manual error analysis: