IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2024, accepted 25 September 2024, date of publication 2 October 2024, date of current version 18 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3472500

RESEARCH ARTICLE

Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization

AKHILA ABDULNAZAR[1,2], ROLAND ROLLER[3], STEFAN SCHULZ[1], AND MARKUS KREUZTHALER[1]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, 8036 Graz, Austria
[2]CBmed GmbH–Center for Biomarker Research in Medicine, 8010 Graz, Austria
[3]German Research Center for Artificial Intelligence (DFKI), 10559 Berlin, Germany

Corresponding author: Markus Kreuzthaler (markus.kreuzthaler@medunigraz.at)

ABSTRACT Most clinical information is only available as free text. Large language models (LLMs) are increasingly applied to clinical data to streamline communication, enhance the accuracy of clinical documentation, and ultimately improve healthcare delivery. This study focuses on a corpus of anonymized clinical narratives in German. On the one hand it evaluates the use of ChatGPT for text cleansing, i.e., the automatic rephrasing of raw text into a more readable and standardized form, and on the other hand for retrieval-augmented generation (RAG). In both tasks, the final goal was medical concept normalization (MCN), i.e., the annotation of text segments with codes from a controlled vocabulary using natural language processing. We found that ChatGPT (GPT-4) significantly improves precision and recall compared to simple dictionary matching. For all scenarios, the importance of the underlying terminological basis was also demonstrated. Maximum F1 scores of 0.607, 0.735 and 0.754 (i.e, for top 1, 5 and 10 matches) were achieved through a pipeline including document cleansing, bi-encoder-based term matching based on a large domain dictionary linked to SNOMED CT, and finally re-ranking using RAG.

INDEX TERMS ChatGPT, medical concept normalization, retrieval augmented generation, text cleansing.

I. INTRODUCTION

Electronic health records include both structured data and unstructured narrative content. Structured data are easy to analyze, particularly when coded by standardized semantic identifiers. In contrast, clinical narratives exhibit the whole range of phenomena that emerge when humans take notes in a hurry. To bridge the semantic gap between unstructured data and semantically explicit, coded information, natural language processing (NLP) methods such as named entity recognition and medical concept normalization (MCN) have been used [1]. Traditionally, the two tasks have been tackled separately, with most approaches focusing on either entity recognition [2], [3] or normalization independently [4], [5]. However, optimizing this process relies heavily on annotated data [6]. Current trends in large language models (LLMs) have shown strong performance across various NLP tasks without requiring extensive parameter tuning or training [7]. This suggests their potential and versatility for better few-shot and transfer learning abilities, indicating that they may ultimately serve as a comprehensive framework for various NLP tasks [7], [8], [9], [10], [11]. In this article, we explore how the potential of LLMs can be harnessed to enhance performance in MCN.

MCN links words and phrases (entity mentions) to standardized and language-independent codes in controlled vocabularies and ontologies [12]. MCN is crucial for information extraction and heavily relies on the coverage of language resources, particularly terminology systems, often further specialized as controlled vocabularies, thesauri, statistical classifications and ontologies. English has a big advantage over all other languages because it is by far the language best

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

covered by these systems. Although achieving acceptable MCN results is still challenging for English-language clinical texts, it is largely more difficult for clinical texts in other languages.

Over the past decade, deep learning and language models have revolutionized NLP, enabling machines to understand and generate human language with unprecedented accuracy. These advancements began with the development of pre-trained word embeddings from large non-annotated text corpora, which have shown their usefulness for MCN, unsurprisingly with a strong bias towards English texts [13], [14]. ELMo introduced contextual word embeddings, advancing the cutting-edge for several major NLP benchmarks [15]. The Generative Pre-trained Transformer (GPT) further minimized task-specific parameters by allowing simple fine-tuning for downstream tasks [16]. Unlike earlier models such as ELMo and GPT, which used unidirectional language models, BERT introduced masked language models for pre-training bidirectional representations, significantly improving performance, as evidenced for eleven NLP tasks [17], [18]. One of the currently most popular LLM, ChatGPT [19], incorporates generative techniques to produce contextually relevant text. Not specifically trained on medical data, ChatGPT has showcased its versatility in various research and healthcare applications [20], including diagnosis support, treatment optimization, and medical question-answering. However, the proprietary nature of many LLMs, particularly GPT and their opaque, ''black box'' character has raised concerns about transparency, accountability, and potential biases regarding their deployment in the context of health care [21].

This paper reports on the combination of BERT and generative models for MCN in German texts, supported by two German medical terminology resources linked to the clinical terminology standard SNOMED CT, an ontology with more than 350,000 units of meaning (concepts) [22], [23], [24]. For MCN, we use a bi-encoder model specifically pre-trained to understand biomedical terminology [25], [26]. On the one hand, ChatGPT's, GPT-4 [27] architecture is used in a preprocessing step to make raw clinical narratives more uniform and interpretable, thus not only easier for human understanding but also for MCN. On the other hand, GPT-4 is used to optimize the selection of candidates for concept matching using retrieval-augmented generation (RAG).

Our investigation is structured as follows. First, we report on how we created a German-language annotated corpus for MCN, using SNOMED CT as the annotation vocabulary, adhering to the annotation guidelines of the n2c2 (National NLP Clinical Challenge) normalization task [28]. Then we expose how GPT-4 was prompted to perform text cleansing. This resulted in two datasets: (i) raw data and (ii) cleansed data. Simple dictionary matching and bi-encoder-based matching were then employed for MCN. Additionally, the RAG capability of GPT-4 was implemented to re-rank the best match from the mapped list.

II. RELATED WORK

A. MEDICAL CONCEPT NORMALIZATION (MCN)

Conventional MCN includes dictionary lookup, deep learning, retrieval, and ranking methods [29], [30], [31], [32]. Deep learning models such as convolutional neural networks and recurrent neural networks with pre-trained word embeddings had shown significant improvements in MCN accuracy, surpassing the previous state-of-the-art [33]. One method involves encoding terminology labels and synonyms into a vector space that uses text and graph embeddings to represent text sequences as vectors. Using a semantic proximity measure, e.g. cosine similarity, the nearest terminology item can be found – and, in consequence, the most appropriate terminology concept – for a given input, resulting in improved classification accuracy across benchmark datasets [34]. BERT models have demonstrated superior performance for MCN compared to other architectures [35], as they excel in managing multilingual data and better capture contextual information [1], [18], [36], [37], [38]. In the 2019 n2c2/UMass Lowell task on MCN, the most accurate approach involved a deep learning architecture with a pre-trained SciBERT layer [28]. Self-alignment pretraining for biomedical entity representation (SapBERT), is a pretraining scheme designed for learning representations of biomedical entities. It outperforms existing models in MCN tasks and achieves cutting-edge results across various datasets without the need to fine-tune the labeled data of the task [25]. Fine-tuning SapBERT established a new benchmark for MCN, including cross-lingual normalization [39], [40], a task also tackled by the modular xMEN [41] system, which uses unsupervised candidate generation and supervised cross-encoders for re-ranking, surpassing previous state-of-the-art performance on diverse benchmarks.

These approaches offer advantages such as improved classification accuracy, scalability to millions of target concepts, and efficient accommodation of growing lexicon sizes [34]. However, they require manual mapping of training data, show difficulty in mapping unseen concepts, and require the retraining of models whenever new content is added. Limitations also arise from the dynamic and fragmented nature of clinical language, a genre replete with spelling errors, jargon expressions, and shorthand expressions, which require constant contextual corrections and disambiguation [42]. Kartchner et al. [43] suggested that LLM-based normalization can enhance the performance of existing models and improve the quality and accuracy of LLM-generated text.

B. LLMS IN THE CLINICAL DOMAIN

Great expectations are associated with the integration of LLM into clinical data management workflows. These models have shown excellent language understanding skills across many domains and performed well in tasks such as summarizing [44]. Agrawal et al. [45] described how ChatGPT optimizes content retrieval and saves time for healthcare

professionals. The provision of concise patient summaries facilitated rapid access to essential information [46]. ChatGPT also demonstrated the potential to generate diagnostic reports or recommendations based on past clinical data, aiding in identifying patterns and connections not immediately apparent to clinicians [47]. Together with genetic information and biomarkers, ChatGPT was shown to offer tailored treatment recommendations and predict individual responses to therapies [48]. In telemedicine, it has facilitated virtual patient-physician interactions, assisted in triaging, and provided remote guidance for home care [49]. LLMs have also shown remarkable potential in biomedical applications, particularly in named entity recognition, by leveraging strategic prompting and integration of external resources. While BERT excels in precision, GPT surpasses in recall and F-score, making it more comprehensive in identifying relevant entities [11], [50]. Nevertheless, seamless integration of LLMs into existing clinical workflows and systems is crucial for their effective use in healthcare.

Ethical issues, privacy concerns, and technical limitations have constantly been discussed [21], particularly in the current context where the leading LLMs are proprietary, and the performance of open models that can be run on premises has lagged. There is a broad consensus that integrating LLMs in healthcare poses risks, including bias in training data, incorrect content, lack of explanations, and reduced need for human expertise. Privacy breaches, legal disputes, interpretability challenges, and misinformation are also concerns [51]. Mitigation requires accurate benchmarking, the use of explainable AI methodologies, and rigorous certification before deployment as medical products.

1) TEXT CLEANSING

Language models have been shown to help cleanse the typical hastily written clinical jargon by correcting spelling, standardizing terms, and improving text clarity [52], [53], across document types such as discharge summaries, radiology reports and other clinical narratives [54] and reaching a satisfactory quality level [55]. Even highly elliptical texts overloaded with short forms can be organized in a coherent manner [56]. For instance, a physician can instruct a language model to include specific elements and to briefly explain some ideas, allowing it to rapidly generate a formal discharge summary. Preliminary studies suggest that ChatGPT could improve the quality of discharge summaries [57], potentially reducing the risk of miscommunication and improving patient care [58]. However, while these generated summaries may appear well-structured, their accuracy and reliability must be rigorously evaluated to ensure they meet clinical standards.

2) RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG is a common practice to address the limitations of models that may not contain all necessary information or have become outdated. This technique combines retrieval with text generation models to incorporate additional information dynamically at runtime [59], [60]. The retrieval component fetches relevant information from a database in response to a query, while the generative language model uses this information to craft contextually relevant responses. Thus, the model's output generation capabilities are enhanced without requiring costly re-training cycles.

Advanced RAG additionally incorporates sophisticated pre-retrieval and post-retrieval processes. One critical post-retrieval aspect in advanced RAG is ''Re-Rank'', which reorders the retrieved documents by relevance [61]. It employs algorithms that adjust the ordering based on criteria such as document diversity or relevance to the query. Re-ranking aims to present the most pertinent information to the LLM, thereby improving the quality and relevance of the generated responses [62]. The use of re-ranking in ChatGPT's RAG approach can lead to more accurate and contextually relevant responses, as it allows the model to consider the latest information from knowledge bases and adjust its responses accordingly. This can be particularly beneficial in the clinical domain, where the accuracy and relevance of responses are critical for effective communication and decision-making. A recent study has explored LLMs for few-shot information extraction tasks and introduced a novel paradigm to enhance their effectiveness [63]. Using prompting strategies and an adaptive filter-then-rerank approach, the system achieved notable improvements (averaging 2.4% F1 gain) over existing methodologies, showcasing the potential of LLMs to tackle challenging information extraction tasks.

Besides this broad scope of LLM applications presented since 2022, to the best of our knowledge, no work has prompted a conversational LLM with the specific goal of optimizing narratives for MCN. Following the suggestions by [43], our goal is to leverage LLMs to enhance MCN. By combining the text cleansing capability of ChatGPT with RAG, our objective is to improve the accuracy and efficiency of MCN.

III. METHODOLOGY AND DATA

That pre-processing of clinical narratives using an LLM improves the precision of the MCN in various settings is a central hypothesis of our work, investigating a scenario where SapBERT is used for similarity search. To validate this hypothesis, we created an annotated corpus for MCN using SNOMED CT as an annotation vocabulary. Then, we used GPT-4 [27] for cleansing the narratives and as a re-ranker for the concepts retrieved by SapBERT. The results were compared to a simple dictionary lookup baseline. Since the proposed methodology does not include any pretraining, it should be considered as an unsupervised way of MCN using the advantages of LLMs. The diagrammatic representation of the study is shown in Figure 1.

A. DATASET CREATION

We created a corpus of clinical texts in German, representative of the clinical information system of KAGes, an Austrian network of public hospitals. Ten discharge

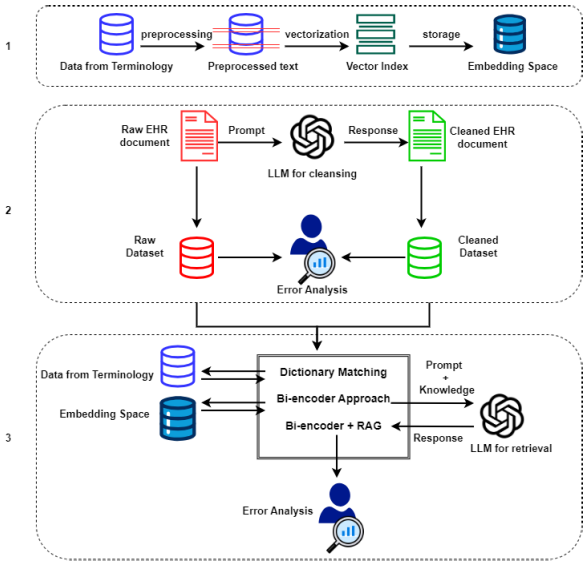

FIGURE 1. Diagrammatic representation of the proposed methodology, (1) generation of the embedding space base from two German custom terminologies, (2) cleansing of the clinical narratives using GPT-4 and its error analysis, (3) MCN using different approaches such as dictionary lookup and bi-encoders. Re-ranking the retrieved concepts by the bi-encoder approach using GPT-4 and error analysis.

summaries from different clinical departments had first been manually anonymized. Then, their content was moderately alienated so that it could be safely assumed not to denote any particular patient. This was done by the third author, a medical expert, in a way that these synthetic documents appeared maximally authentic both in content and structure, following an approach described in [64]. The assignment of SNOMED CT [23] codes was carried out collaboratively by two experts, one experienced in clinical practice and the other in biomedical sciences, following the n2c2 Annotation Guidelines [28]. The output consisted of 660 annotated surface terms.

B. PROMPT FOR CLEANSING

Following the initial phase of corpus construction, a data cleansing procedure was initiated to refine the quality of corpus content. The ten narratives underwent cleansing using two types of GPT-4 prompts. A first general prompt aimed at standardization by expanding acronyms and correcting spelling errors to ensure clarity and consistency. This design was inspired by practical guidelines such as those provided by Google Gemini [65], which emphasize the importance of clarity in the prompts. A second, more specific prompt included these aims and additionally aimed at the enrichment of branded drug names by their corresponding substance names, which can then be aligned with SNOMED CT codes (SNOMED CT does not contain any brand names). This approach was informed by best practices that advocate for detailed and specific instructions to improve accuracy and relevance. Upon cleansing, terms aligned with the annotated corpus were extracted from the narratives. This process yielded two distinct datasets: one from the raw narratives and the other one from the cleansed ones. Each dataset was designed to meet specific analytical objectives; the second prompt was preferred for its detailed instructions and the potential to improve the overall completeness of the dataset. The prompts are as follows, accompanied by an example:

System: You are an expert in the clinical domain.

Prompt: Standardize and transform the given German clinical narrative into standardized language without abbreviations and spelling errors in German. Any abbreviations should be expanded into the corresponding long form, any existent spelling errors should be corrected. The corresponding substance should be added in round brackets after the given drug name for any drug or pharmaceutical name found in the clinical narrative. All input information should be considered and be represented in the output. No additional explanations should be added other than the transformed clinical narrative text. The clinical narrative is found below:

Example:

Raw text snippet[1]*: "Anamn. besteht eine dilat. CMP, eine incipiente KHK sowie eine intermitt. VHFA."*

Cleansed text snippet: "In der Anamnese besteht eine dilatative Kardiomyopathie, eine beginnende koronare Herzkrankheit sowie ein intermittierendes Vorhofflimmern."

C. TERMINOLOGY BASE

To standardize clinical terms, MCN links information to codes from terminology systems, which act as semantic identifiers to domain terms in one or more languages. The Metathesaurus of the Unified Medical Language System (UMLS) [66] links lexical and conceptual information between approximately 200 different biomedical terminology systems, encompassing MeSH, SNOMED CT, ICD-10, MedDRA, and RxNORM. UMLS was created and is maintained by the U.S. National Library of Medicine, which explains its strong focus on English. To a minor extent, it includes lexical content from other languages, such as Spanish or German. Due to the terminology cross-links, each concept can be enhanced by (quasi-)synonymous terms from different terminologies, including non-English ones.

SNOMED CT translation to German is still in an early stage. Therefore, we used two custom German term collections linked to SNOMED CT codes, in order to build two embedding spaces. The first custom terminology, UMLS_DE, was created on the fly by extracting all German terminological units from the UMLS Metathesaurus for which a connection to some SNOMED CT code could be established via a common identifier (CUI). Thus, approximately 79,000 medical terms for 41,000 SNOMED CT concepts were harvested. The second custom terminology, IT_DE, consists of a German Interface Terminology built semi-automatically [67] during the past ten years. The

[1]Raw: "H/O dil CMP, beginning CHF and intermitt AFib"; Cleansed: "History of dilatative cardiomyopathy, beginning congestive heart failure and intermittent atrial fibrillation."

extract used for this study links approximately 2.5 million manually harvested and automatically combined clinical terms to 278,000 SNOMED CT concepts. Both custom terminologies can be described as term collections annotated with SNOMED CT codes without claiming to be SNOMED CT translations in a proper sense.

D. PRE-PROCESSING

A series of terminology pre-processing steps were done to enhance the readiness of the data, including the removal of special characters, conversion to lowercase, and vectorization using a cross-lingual SapBERT trained with UMLS 2020AB (all languages) and XML-RoBERTa (large) [68] as the underlying model. Subsequently, the processed data is indexed via FAISS, an open-source library to perform fast similarity searches in high-dimensional vector spaces [69], facilitating efficient retrieval and utilization in subsequent analyses and applications. The entries in both custom terminologies (UMLS_DE and IT_DE) are then represented as 768-dimensional embeddings and FAISS-indexed and create the corresponding embedding space. The cosine similarity function is used for vector similarity matching.

E. MCN–MEDICAL CONCEPT NORMALIZATION

The following presents different normalization techniques used in this work, starting with a baseline method of dictionary mapping, followed by a bi-encoder approach, and further improving it with an RAG approach.

1) BASELINE APPROACH

A baseline for MCN was created using simple dictionary matching to identify correspondences with SNOMED CT codes by matching the text in the clinical corpus against the custom terminologies UMLS_DE and IT_DE. The process included pre-processing the clinical text as detailed in Section (III-D) and ensuring exact matches to identify the correspondences.

2) BI-ENCODER APPROACH

The input clinical terms underwent the pre-processing steps detailed in Section (III-D), followed by using cross-lingual SapBERT to convert the clinical terms into vectors. Then, these pre-processed and vectorized terms were submitted to a cosine similarity search within the embedding space using FAISS. A threshold of 0.9 had been set for similarity matching to ensure precision. The score was chosen based on the average similarity scores observed between known synonyms and the preferred terms in SNOMED CT, aiming to reduce ambiguity in the terms identified. This process retrieved the top ten candidate concepts from the embedding space.

3) RETRIEVAL-AUGMENTED GENERATION (RAG) APPROACH

In this approach, the ten candidate concepts retrieved by the bi-encoder approach (in Section (III-E2)), were presented to the GPT-4 model, along with their input terms and contexts. A prompt was created that instructs the model to re-rank these candidates based on contextual relevance. We used two prompts in experiments across four examples, selecting the most effective prompt based on its ranking performance. The following refers to the prompt used for this process.

System: As an expert ranker of related words tailored to specific contexts, your role is pivotal in identifying the most pertinent terms within a given context.

Prompt: Upon receiving an input text along with its context and a predefined list of 10 terms, your task is to re-rank these terms based on their contextual significance, prioritizing the most suitable choices at the top. Importantly, ensure that the re-ranked list includes only the terms provided in the input list, without filtering or adding any additional terms. Guideline: Re-rank the terms based on their contextual relevance, optimizing their alignment with the provided context. Output Always present the re-ranked list without any additional explanations in the format [term: id].

TABLE 1. Summary of ChatGPT improvements at narrative and surface term level.

Evaluation metric	Surface terms
Average number of words changed per narrative	7.72%
Average number of lines changed per narrative	67.85%
Missing annotated spans	0.45%
Changed annotated spans after cleansing	45%
Changes with potential good influence:	
Well-formed	31.0%
Synonym	2.0%
Changes with potential neutral influence:	
Synonym	3.0%
Misspell	1.0%
Changes with potential bad influence:	
Wrong	3.0%
Misspell	1.0%
Incomplete	1.0%
Hypernym	2.5%
Hyponym	0.5%

IV. RESULTS

Table 1 details the evaluation of the cleansing approach, including systematic analysis and medical content evaluation by an expert. The raw and cleansed narratives were thoroughly examined, resulting in a 7.72% reduction in word count and a 67.85% decrease in line numbers. A detailed comparison of surface terms between the raw and cleansed texts revealed that 0.45% (3 out of 660 annotated terms) were missing after cleansing, necessitating the addition of surface terms from the raw narratives to ensure completeness. Furthermore, after cleansing, 45% of the annotated spans were found to be changed. Changes were categorized according to their potential influence: well-formed and beneficial synonyms, suggesting a good impact (33.0%). Neutral changes include neutral synonyms and misspellings (4.0%). Potentially negative changes (8.0%) include wrong entries, misspellings, incomplete information, hypernyms,

TABLE 2. Concept normalization using different approaches, referred to under the column 'Method' based on SNOMED CT codes, using different terminologies as in column 'Source', (i) a custom terminology for German extracted from the UMLS Metathesaurus (UMLS_DE) and (ii) the German Interface Terminology for SNOMED CT (IT_DE). Column 'Data' shows the different types of data used, (i.e., raw and cleansed). The performance of these approaches is evaluated for precision (P), recall (R) and F1 score for top one, five and ten matches.

Method	Source	Data	P@1	R@1	F1@1	P@5	R@5	F1@5	P@10	R@10	F1@10
Dictionary matching	UMLS_DE	raw text	0.137	0.115	0.122						
		cleansed text (GPT-4)	0.195	0.173	0.178						
	IT_DE	raw text	0.330	0.284	0.296						
		cleansed text (GPT-4)	0.339	0.282	0.297						
Bi-encoder	UMLS_DE	raw text	0.252	0.232	0.232	0.343	0.323	0.325	0.366	0.341	0.346
		cleansed text (GPT-4)	0.306	0.285	0.286	0.440	0.435	0.432	0.458	0.453	0.450
	IT_DE	raw text	0.593	0.523	0.542	0.673	0.629	0.641	0.702	0.656	0.669
		cleansed text (GPT-4)	**0.618**	**0.552**	**0.568**	0.764	0.724	0.735	**0.782**	**0.745**	**0.754**
RAG	UMLS_DE	raw text	0.253	0.244	0.241	0.351	0.332	0.333			
		cleansed text (GPT-4)	0.321	0.300	0.298	0.444	0.443	0.436			
	IT_DE	raw text	0.615	0.558	0.572	0.686	0.641	0.653			
		cleansed text (GPT-4)	**0.646**	**0.595**	**0.607**	**0.768**	**0.724**	**0.735**			

and hyponyms. These findings highlight the importance of the cleansing process in reducing narrative length and improving the clarity of terms. However, they also reveal potential risks associated with content loss or alteration. To address this, we ensure transparency and reliability through manual evaluation by a domain expert. AI in clinical settings should assist, not replace, human decision-making, with rigorous validation and testing to catch errors and ensure reliability.

Subsequently, the datasets underwent MCN using three distinct approaches: dictionary matching, the bi-encoder method, and the RAG for re-ranking. Table 2 details the performance of both raw and cleaned corpora across the two custom terminologies (i.e, (UMLS_DE) and (IT_DE)) and approaches. Cleansing text with GPT-4 often improves precision and recall, leading to higher F1 scores when compared to using raw text in all three approaches. Moving from dictionary matching to the bi-encoder approach showed a significant 91.25% increase in the F1 score, indicating a substantial performance improvement using the text cleaned by GPT-4 and the IT_DE terminology. Notably, the unsupervised bi-encoder method using GPT-4-cleaned text achieved an F1 score of 0.568 for top 1 matches, 0.735 for top 5 matches and 0.754 for top 10 matches. Furthermore, transitioning from the bi-encoder to the RAG resulted in a smaller but notable 6.87% gain in F1 score for the top 1 matches, improving from 0.568 to 0.607, demonstrating continued enhancement in performance within the same setting for data and terminology. The re-ranked terms have been cross-checked using an algorithm to ensure that no changes were made to the list of terms other than their order, thereby ensuring that no hallucinations occurred. The IT_DE terminology consistently shows higher performance metrics compared to UMLS_DE. This observation, combined with the performance improvements, highlights the efficacy of these methods in enhancing the efficiency in MCN systems.

A. ERROR ANALYSIS

We performed a qualitative error analysis on the top ten candidates retrieved by the best-performing scenario, i.e. bi-encoder on the cleansed corpus using the custom terminology IT_DE. Typical errors are described in the following. We distinguish between document-cleansing errors and concept-matching errors. Out of the 660 terms, for the top 10 matches, 25.6% (169/660) were wrong matches, of which 35.5% (60/169) were due to document cleansing.

1) DOCUMENT CLEANSING ERRORS

These errors affected the text in a way that the meaning of the cleansed text was altered. E.g., the meaning (in English) of ''Abdomen: soft abdominal wall, no tenderness, intense bowel sounds, no flank pain'' can be unambiguously obtained from the following passage in the German raw text: ''Abd. BD weich, kein DS, DG's rege, NL frei.''. Cleansing transformed this to ''Abdomen: Bauchdecke weich, kein Druckschmerz, Darmgeräusche rege, Leber nicht vergrößert'', where ''NL frei'' (''no flank pain'', annotated in the gold standard with ''300447004 Kidney non-tender (situation)''), was erroneously rephrased to ''Leber nicht vergrößert'' (''liver not enlarged'').

A similar phenomenon characterises the next example, which means in English:''The cardiac activity is rhythmical, with normal sinus rhythm'' and appears in raw text as ''HA rhythmisch, nc.''. A cardiologist understands from the context that ''nc.'' is the abbreviation of ''normocard'', which justifies the annotation of the whole passage with ''64730000 Normal sinus rhythm (finding)''. Cleansing transformed this passage to ''Herzaktivität ist rhythmisch, normal korrigiert'', which ended up being annotated by the system with ''263699000 Cardiac activity (observable entity)''. This is not wrong, like in the previous example, but not specific enough. The (non-existing) expansion of ''nc.'' to ''normal korrigiert'' was fortunately not annotated by the

system, but clearly shows the tendency toward hallucination when the language model encounters unknown expressions in uncommon contexts. When analyzing the document cleansing errors, 42% were due to incorrect expansion of short forms, while 20% were caused by adding incorrect substances to drug names.

2) MAPPING ERRORS

These were frequent errors, independent of the cleansing process. So had ''Mäßig inhomogene Parenchymstruktur der Leber wie bei Steatose/ LPS.'' (which means ''Moderately heterogeneous parenchymal structure of the liver as in steatosis / Liver Parenchymal Steatosis'') the gold standard annotation ''197321007 Steatosis of liver (disorder)'', because the annotator concluded from the context that ''steatosis'' here means ''steatosis of the liver''. However, the system mapped it to the parent concept ''1187537008 Steatosis (disorder), along with ''127879008 Structure of parenchyma of liver (body structure)''. This reveals a fundamental problem of compositional terminologies, *viz.* that different sequences of pre-coordinated concepts can be derived from the same atomic elements.

V. DISCUSSION

A. IMPACT OF DOCUMENT CLEANSING

Similar to the study by Ayre et al. [53], our cleansing process resulted in substantial reductions in word and line counts, cf. Table 1. This suggests that redundant, unnecessary, or irrelevant content was removed, which was likely to enhance text clarity and conciseness. Despite variations in the extent of reduction among different narratives, there was a consistent trend toward a more concise and structured information presentation. However, in some cases, the word number increased, suggesting restructuring for clarity improvement. A typical case is the expansion of acronyms. Across all methods and for both terminology sources, a notable performance improvement occurred whenever text was cleansed by GPT-4. This suggests that this task successfully approximated the medical terms to more common, normalized terms corresponding to the terminology collections used, leading to improved normalization results.

B. EFFECTIVENESS OF DICTIONARY MATCHING

Dictionary matching, in isolation, achieved the lowest performance, especially when applied to raw text. This underscores the limitations of a simple matching approach, particularly with clinical texts known for their special jargon that is often not covered by standard terminologies.

C. BI-ENCODER METHOD PERFORMANCE

This method consistently outperformed the baseline across all metrics and for both terminology sources, underscoring its MCN effectiveness. Furthermore, the result demonstrates the significant advantage when applied to the (large) interface terminology (IT_DE) over the much smaller UMLS Metathesaurus extract (UMLS_DE). This finding strongly supports the critical importance of good terminology coverage in MCN.

D. IMPACT OF GPT-4 re-ranker

The GPT-4 re-ranker, in addition to the bi-encoder method, further enhanced performance, particularly on already cleansed text. As per the study by Ma et al. [63], LLMs as re-rankers in challenging information extraction tasks exhibited an F1 gain around 2.4%. This observation underscores the effective refinement of candidate concepts generated by the bi-encoder, leading to more precise normalization outcomes. To the best of our knowledge, LLMs as re-rankers have not been previously utilized for MCN tasks. Our investigation thus demonstrates a notable improvement, achieving a 6.87% F1 gain compared to the bi-encoder approach alone. A previous attempt to perform MCN directly by prompting GPT-4 for finding the right SNOMED code for a given expression had been immediately abandoned due to its propensity to the suggestion of completely hallucinated codes, indicating that GPT-4 lacks sufficient medical terminology knowledge. This highlights the necessity for a framework, such as bi-encoders, that first retrieves the correct concepts from the terminology and then improves accuracy by re-ranking.

E. THE IMPORTANCE OF TERMINOLOGIES

The result provides interesting insight when comparing UMLS_DE with IT_DE, two very different term collections with links to SNOMED CT codes. The former is of limited size (79,000 biomedical terms for 41,000 concepts) and only linked to SNOMED CT indirectly, whereas the latter is huge (2.5 million biomedical terms for 278,000 concepts) and provides good coverage of the numerous varieties of clinical jargon. This explains the almost doubling of the F-values across experiments. However, the high terminological coverage also explains that IT_DE benefitted comparatively less from document cleansing. It is, however, not surprising that document cleansing is the more beneficial, the scarcer is the coverage by existing terminologies. This is an important message for other languages, for which UMLS terminology extracts similar to our UMLS_DE could easily be obtained, but for which resources like IT_DE do not exist.

F. THE INTERPRETATION OF THE BEST-PERFORMING SCENARIO

The combination of IT_DE with text cleansing, bi-encoder and re-ranking yields an optimal F1@1 value of 0.607. This is all the more remarkable when compared to results of an earlier human text annotation exercise with SNOMED CT, where only low inter-annotator agreement values (Krippendorff's Alpha approx. 40%) had been achieved on a mix of clinical texts [70], and which had been interpreted as a result of SNOMED CT's huge size and of the unclear meaning of semantically related concepts.

G. PATIENT SAFETY

While challenges such as dealing with medical terminology, correctly resolving brand names, and preventing hallucinations remain, the proposed method faces risks such as loss or alteration of content at a relatively low level, given the black-box nature of LLMs. Their potential impact on patient safety must be assessed individually for each implementation scenario, which is, however, beyond the scope of this study. This means, particularly, the adaptation of the weighting between false positives and false negatives to the respective scenario. For instance, false positives may be more acceptable in cohort-building use cases for patient recruitment than in decision support scenarios, where a hallucinated drug or disease code can put the patient at risk. In contrast, when LLMs are used to improve content retrieval in medical records, a high recall is to be aimed at and false positives are more readily tolerated.

H. DATA PROTECTION

For our study, we chose GPT-4 was as the currently best-performing language model. However, the use of AI in healthcare requires strict compliance with data protection and security regulations, which makes GPT-4's proprietary nature a severe obstacle. Open source models such as Llama 3 should be preferred, but they have so far had significantly lower performance, especially for languages other than English. So it remains to wait for open source models with better performance. The deployment of commercial LLMs on trusted environments, such as Azure Cloud, could be an alternative. The extent to which this actually ensures the protection of highly sensitive patient data will be the subject of future discussions, as will all the trends that are currently observed in the extremely dynamic landscape of language models.

VI. CONCLUSION

Our study addressed medical concept normalization (MCN) of clinical narratives, i.e., the automated annotation of clinically relevant text passages with codes from the terminology SNOMED CT. The results support the transformative benefits of integrating LLMs such as ChatGPT into a concept normalization workflow. More precisely, ChatGPT significantly enhanced the performance of MCN, particularly by applying text cleansing and retrieval augmented generation (RAG) to clinical narratives in German language. Our investigation reveals a notable 6.87% increase in F1 score when using a bi-encoder with a re-ranker, compared to traditional bi-encoder approaches, underscoring ChatGPT's superiority in enhancing MCN tasks.

The benefits of document cleansing are especially pronounced in scenarios with limited terminological coverage. The impressive ability of ChatGPT to transform raw text into text with more standardized medical terms is an important message, particularly when using MCN for languages with limited support for terminological resources. Nevertheless, our experiments on German texts reveal the importance of a good terminological basis, even in times of LLMs, when comparing an extract of UMLS Metathesaurus linked to SNOMED codes with a German-specific interface terminology where F1 values roughly doubled.

Looking ahead, adopting LLMs holds promise for streamlining healthcare workflows, reducing documentation errors, and ultimately improving patient care outcomes. Future research should focus on refining these models, expanding their integration into diverse healthcare settings, and evaluating their impact in the real world on clinical practice.

Continuous attention must be paid to the settings (e.g., cloud platforms) in which commercial LLMs can be safely used for routine patient data. Likewise, the black-box nature of LLMs with the unpredictability of hallucinations requires careful monitoring and mitigation, e.g., by using additional resources for checking the plausibility of medical term changes proposed by the LLM.

Finally, clinical corpora in languages other than English, annotated with codes from standard terminologies at high granularity, are urgently needed as a source of additional ground truth data covering the full spectrum of clinical specialties. This would solve the problems of studies with a limited sample size, such as ours. Nevertheless, we can claim that this study has indicated directions for future research on larger corpora, such as those being created in the ongoing German GeMTeX project [71], which marks an important new step in resource creation for medical concept normalization.

ACKNOWLEDGMENT

This work was approved by the IRB of the Medical University of Graz (30-496 ex 17/18).

REFERENCES

[1] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "BERN2: An advanced neural biomedical named entity recognition and normalization tool," *Bioinformatics*, vol. 38, no. 20, pp. 4837–4839, Oct. 2022.

[2] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decis. Anal. J.*, vol. 10, Mar. 2024, Art. no. 100426.

[3] M. Y. Landolsi, L. Ben Romdhane, and L. Hlaoua, "Hybrid medical named entity recognition using document structure and surrounding context," *J. Supercomput.*, vol. 80, no. 4, pp. 5011–5041, Mar. 2024.

[4] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[6] B. Santana, R. Campos, E. Amorim, A. Jorge, P. Silvano, and S. Nunes, "A survey on narrative extraction from textual data," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8393–8435, Aug. 2023.

[7] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, and Y. Zhang, "A large language model for electronic health records," *NPJ Digit. Med.*, vol. 5, no. 1, p. 194, 2022.

[8] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, G. Lipori, D. A. Mitchell, N. S. Ospina, M. M. Ahmed, W. R. Hogan, E. A. Shenkman, Y. Guo, J. Bian, and Y. Wu, "A study of generative large language model for medical research and healthcare," *npj Digit. Med.*, vol. 6, no. 1, p. 210, Nov. 2023.

[9] M. Javaid, A. Haleem, and R. P. Singh, "ChatGPT for healthcare services: An emerging stage for an innovative perspective," *BenchCouncil Trans. Benchmarks, Standards Evaluations*, vol. 3, no. 1, Feb. 2023, Art. no. 100105.

[10] T. Dave, S. A. Athaluri, and S. Singh, "ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations," *Frontiers Artif. Intell.*, vol. 6, May 2023, Art. no. 1169595.

[11] Á. García-Barragán, A. G. Calatayud, O. Solarte-Pabón, M. Provencio, E. Menasalvas, and V. Robles, "GPT for medical entity recognition in Spanish," *Multimedia Tools Appl.*, pp. 1–20, Apr. 2024.

[12] S. Schulz, P. Daumke, M. Romacker, and P. López-García, "Representing oncology in datasets: Standard or custom biomedical terminology?" *Informat. Med. Unlocked*, vol. 15, Jan. 2019, Art. no. 100186.

[13] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang, and D. Huang, "CNN-based ranking for biomedical entity normalization," *BMC Bioinf.*, vol. 18, no. 11, pp. 79–86, Oct. 2017.

[14] Y. Luo, G. Song, P. Li, and Z. Qi, "Multi-task medical concept normalization using multi-view convolutional neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2018, pp. 2227–2237.

[16] A. Radford. (2018). *Improving Language Understanding With Unsupervised Learning*. [Online]. Available: https://openai.com/research/language-unsupervised

[17] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, pp. 1–16.

[18] Z. Ji, Q. Wei, and H. Xu, "BERT-based ranking for biomedical entity normalization," in *Proc. AMIA Summits Transl. Sci.*, 2020, p. 269.

[19] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology—A recent scoping review," *Diagnostic Pathol.*, vol. 19, no. 1, p. 43, Feb. 2024.

[20] R. K. Sinha, A. D. Roy, N. Kumar, and H. Mondal, "Applicability of ChatGPT in assisting to solve higher order problems in pathology," *Cureus*, vol. 15, no. 2, Feb. 2023.

[21] A. Alsadhan, F. Al-Anezi, A. Almohanna, N. Alnaim, H. Alzahrani, R. Shinawi, H. Aboalsamh, A. Bakhshwain, M. Alenazy, W. Arif, S. Alyousef, S. Alhamidi, A. Alghamdi, N. AlShrayfi, N. B. Rubaian, T. Alanzi, A. AlSahli, R. Alturki, and N. Herzallah, "The opportunities and challenges of adopting ChatGPT in medical research," *Frontiers Med.*, vol. 10, Dec. 2023, Art. no. 1259640.

[22] SNOMED International. *SNOMED CT Starter Guide*. Accessed: Apr. 18, 2024. [Online]. Available: https://confluence.ihtsdotools.org/display/DOCSTART

[23] E. Chang and J. Mostafa, "The use of SNOMED CT, 2013–2020: A literature review," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 9, pp. 2017–2026, Aug. 2021.

[24] S. Schulz, W. Del-Pinto, L. Han, M. Kreuzthaler, S. Aghaei, and G. Nenadic, "Towards principles of ontology-based annotation of clinical narratives," in *Proc. ICBO*, 2023, pp. 1–12. [Online]. Available: https://ceur-ws.org/Vol-3603/

[25] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 4228–4238.

[26] A. Abdulnazar, M. Kreuzthaler, R. Roller, and S. Schulz, "SapBERT-based medical concept normalization using SNOMED CT," in *Caring is Sharing-Exploiting the Value in Data for Health and Innovation*. IOS Press, 2023.

[27] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, N. Zaman, P. Sarker, A. G. Lee, and A. Tavakkoli, "GPT-4: A new era of artificial intelligence in medicine," *Irish J. Med. Sci.*, vol. 192, no. 6, pp. 3197–3200, Dec. 2023.

[28] Y.-F. Luo, S. Henry, Y. Wang, F. Shen, O. Uzuner, and A. Rumshisky, "The 2019 n2c2/UMass lowell shared task on clinical concept normalization," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, p. 1529, Oct. 2020.

[29] D. Xu, M. Gopale, J. Zhang, K. Brown, E. Begoli, and S. Bethard, "Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1510–1519, Oct. 2020.

[30] L. Pape-Haugaard, "Clinical concept normalization on medical records using word embeddings and heuristics," *Stud. Health Technol. Inform.*, vol. 270, pp. 93–99, Jan. 2020.

[31] L. Chen, W. Fu, Y. Gu, Z. Sun, H. Li, E. Li, L. Jiang, Y. Gao, and Y. Huang, "Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 10, pp. 1576–1584, Oct. 2020.

[32] K. S. Kalyan and S. Sangeetha, "Target concept guided medical concept normalization in noisy user-generated texts," in *Proc. Deep Learn. Inside Out (DeeLIO), 1st Workshop Knowl. Extraction Integr. Deep Learn. Architectures*, 2020, pp. 64–73.

[33] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, "Medical concept normalization for online user-generated texts," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2017, pp. 462–469.

[34] N. Pattisapu, S. Patil, G. Palshikar, and V. Varma, "Medical concept normalization by encoding target knowledge," in *Proc. Mach. Learn. Health Workshop*, 2020, pp. 246–259.

[35] Z. Miftahutdinov, A. Kadurin, R. Kudrin, and E. Tutubalina, "Medical concept normalization in clinical trials with drug and disease representation learning," *Bioinformatics*, vol. 37, no. 21, pp. 3856–3864, Nov. 2021.

[36] H. Cho, D. Choi, and H. Lee, "Re-ranking system with BERT for biomedical concept normalization," *IEEE Access*, vol. 9, pp. 121253–121262, 2021.

[37] P. Wajsbürt, A. Sarfati, and X. Tannier, "Medical concept normalization in French using multilingual terminologies and contextual embeddings," *J. Biomed. Informat.*, vol. 114, Feb. 2021, Art. no. 103684.

[38] M. Sung, H. Jeon, J. Lee, and J. Kang, "Biomedical entity representations with synonym marginalization," 2020, *arXiv:2005.00239*.

[39] D. Xu and T. Miller, "A simple neural vector space model for medical concept normalization using concept embeddings," *J. Biomed. Informat.*, vol. 130, Jun. 2022, Art. no. 104080.

[40] M. Schwarz, K. Chapman, and B. Häussler, "Multilingual medical entity recognition and cross-lingual zero-shot linking with Facebook AI similarity search," in *Proc. IberLEF@ SEPLN*, 2022, pp. 1–11.

[41] F. Borchert, I. Llorca, R. Roller, B. Arnrich, and M.-P. Schapranow, "XMEN: A modular toolkit for cross-lingual medical entity normalization," 2023, *arXiv:2310.11275*.

[42] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, and E. Fosler-Lussier, "Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 3, pp. 516–532, Mar. 2021.

[43] D. Kartchner, J. Deng, S. Lohiya, T. Kopparthi, P. Bathala, D. Domingo-Fernández, and C. Mitchell, "A comprehensive evaluation of biomedical entity linking models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 14462–14478.

[44] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[45] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 1–26.

[46] Y. Liu, S. Ju, and J. Wang, "Exploring the potential of ChatGPT in medical dialogue summarization: A study on consistency with human preferences," *BMC Med. Informat. Decis. Making*, vol. 24, no. 1, p. 75, Mar. 2024.

[47] Z. Zhou, "Evaluation of ChatGPT's capabilities in medical report generation," *Cureus*, vol. 15, no. 4, Apr. 2023, Art. no. e37589.

[48] N. Shrestha, Z. Shen, B. Zaidat, A. H. Duey, J. E. Tang, W. Ahmed, T. Hoang, M. R. Mejia, R. Rajjoub, J. S. Markowitz, J. S. Kim, and S. K. Cho, "Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: A comparison study," *Spine*, vol. 49, no. 9, pp. 640–651, May 2024.

[49] R. K. Garg, V. L. Urs, A. A. Agrawal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, "Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review," *Health Promotion Perspect.*, vol. 13, no. 3, pp. 183–191, Sep. 2023.

[50] S. J. Jung, H. Kim, and K. S. Jang, "LLM based biological named entity recognition from scientific literature," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2024, pp. 433–435.

[51] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023.
[52] S. Biswas, "ChatGPT and the future of medical writing," *Radiology*, vol. 307, no. 2, Apr. 2023, Art. no. e223312.
[53] J. Ayre, O. Mac, K. McCaffery, B. R. McKay, M. Liu, Y. Shi, A. Rezwan, and A. G. Dunn, "New frontiers in health literacy: Using ChatGPT to simplify health information for people in the community," *J. Gen. Internal Med.*, vol. 39, no. 4, pp. 573–577, Mar. 2024.
[54] T. Brown, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
[55] S. R. Ali, T. D. Dobbs, H. A. Hutchings, and I. S. Whitaker, "Using ChatGPT to write patient clinic letters," *Lancet Digit. Health*, vol. 5, no. 4, pp. e179–e181, Apr. 2023.
[56] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios," *J. Med. Syst.*, vol. 47, no. 1, p. 33, Mar. 2023.
[57] S. B. Patel and K. Lam, "ChatGPT: The future of discharge summaries?" *Lancet Digit. Health*, vol. 5, no. 3, pp. 107–108, Mar. 2023.
[58] H. L. Walker, S. Ghani, C. Kuemmerli, C. A. Nebiker, B. P. Müller, D. A. Raptis, and S. M. Staubli, "Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument," *J. Med. Internet Res.*, vol. 25, Jun. 2023, Art. no. e47479.
[59] E. Y. Song, S. Kim, H. Lee, J. Kim, and J. Thorne, "Re3val: Reinforced and re-ranked generative retrieval," in *Proc. EACL*, 2024, pp. 393–409.
[60] Y. Guo, W. Qiu, G. Leroy, S. Wang, and T. Cohen, "Retrieval augmentation of large language models for lay language generation," *J. Biomed. Informat.*, vol. 149, Jan. 2024, Art. no. 104580.
[61] M. Eibich, S. Nagpal, and A. Fred-Ojala, "ARAGOG: Advanced RAG output grading," 2024, *arXiv:2404.01037*.
[62] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1316–1331, Nov. 2023.
[63] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 10572–10601.
[64] L. Modersohn, S. Schulz, C. Lohr, and U. Hahn, "GRASCCO—The first publicly shareable, multiply-alienated German clinical text corpus," *Stud. Health Technol. Inform.*, vol. 296, pp. 72–76, Jan. 2022.
[65] *Gemini for Google Workspace Prompting Guide*. Accessed: Jun. 26, 2024. [Online]. Available: https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf
[66] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. 267–270, Jan. 2004.
[67] D. H. Nik, Z. Kasác, Z. Goda, A. Semlitsch, and S. Schulz, "Building an experimental German user interface terminology linked to SNOMED CT," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, 2019.
[68] A. Conneau, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.
[69] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.
[70] J. A. Miñarro-Giménez, R. Cornet, M. C. Jaulent, H. Dewenter, S. Thun, K. R. Gøeg, D. Karlsson, and S. Schulz, "Quantitative analysis of manual annotation of clinical text samples," *Int. J. Med. Informat.*, vol. 123, pp. 37–48, Mar. 2019.
[71] F. A. Meineke, L. Modersohn, M. Loeffler, and M. Boeker, "Announcement of the German medical text corpus project (GeMTeX)," *Stud. Health Technol. Inform.*, vol. 302, pp. 835–836, May 2023.

AKHILA ABDULNAZAR received the bachelor's degree in electronics and biomedical engineering from Cochin University of Science and Technology, India, and the master's degree in applied electronics and instrumentation from Kerala Technological University, India. She is currently pursuing the Ph.D. degree with the Medical University of Graz, Austria, specializing in standardizing clinical text data using computational semantics, NLP, and LLMs to enhance healthcare interoperability. Prior to her current roles, she was a Software Engineer with Siemens and a Research Intern with Robert Bosch, specializing in AI/ML applications. Her research interests include data visualization for complex biomedical data; empowering healthcare professionals through NLP, LLM, and image classifiers; and real-time signal classifiers.

ROLAND ROLLER received the degree in computer science and computational linguistics from the University of Trier and Saarland University, and the Ph.D. degree in information extraction from biomedical literature from the University of Sheffield, in 2015. He is a Senior Researcher and the Project Manager with the Speech and Language Technology Group, German Research Center for Artificial Intelligence (DFKI). Currently, he is working on topics related to information extraction, clinical decision support, anonymization, and chatbots.

STEFAN SCHULZ is a Full Professor of medical informatics with the Medical University of Graz. He trained as a Medical Doctor with a doctorate in theoretical medicine, he transitioned to medical informatics after two years of clinical practice. An expert in biomedical terminologies and ontologies, he contributes to standards development with SNOMED International. He also holds a part-time position with German NLP Company Averbis. His research group has been pivotal in several EU projects and has hosted international events. With over 250 peer-reviewed publications, he is a highly published researcher and has received multiple awards for his contributions to the field. His research interests include biomedical terminologies, ontologies, electronic health records, and medical language.

MARKUS KREUZTHALER is an Assistant Professor with the Institute for Medical Informatics, Statistics, and Documentation, Medical University of Graz, specializing in computational semantics for health. He focuses on extracting and representing relevant information from clinical texts using natural language processing, particularly machine learning methods. His current national and international research projects involve creating extended structured and standardized clinical patient profiles to support secondary use cases scenarios from clinical real-world data, utilizing international standards, such as SNOMED CT.