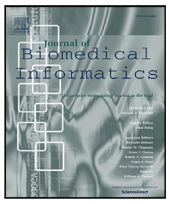
Journal of Biomedical Informatics 165 (2025) 104812

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Enhancing data quality in medical concept normalization through large language models☆

Haihua Chen [a], Ruochi Li [c], Ana Cleveland [b], Junhua Ding [a,*]

[a] The Anuradha & Vikas Sinha Department of Data Science, University of North Texas, Denton, 76203, TX, USA
[b] Department of Information Science, University of North Texas, Denton, 76203, TX, USA
[c] Department of Computer Science, North Carolina State University, Raleigh, 27695, NC, USA

ARTICLE INFO

Keywords:
Medical concept normalization
Machine learning
Data quality
Data augmentation
Large language model
ChatGPT

ABSTRACT

Objective: Medical concept normalization (MCN) aims to map informal medical terms to formal medical concepts, a critical task in building machine learning systems for medical applications. However, most existing studies on MCN primarily focus on models and algorithms, often overlooking the vital role of data quality. This research evaluates MCN performance across varying data quality scenarios and investigates how to leverage these evaluation results to enhance data quality, ultimately improving MCN performance through the use of large language models (LLMs). The effectiveness of the proposed approach is demonstrated through a case study.

Methods: We begin by conducting a data quality evaluation of a dataset used for MCN. Based on these findings, we employ ChatGPT-based zero-shot prompting for data augmentation. The quality of the generated data is then assessed across the dimensions of correctness and comprehensiveness. A series of experiments is performed to analyze the impact of data quality on MCN model performance. These results guide us in implementing LLM-based few-shot prompting to further enhance data quality and improve model performance.

Results: Duplication of data items within a dataset can lead to inaccurate evaluation results. Data augmentation techniques such as zero-shot and few-shot learning with ChatGPT can introduce duplicated data items, particularly those in the mean region of a dataset's distribution. As such, data augmentation strategies must be carefully designed, incorporating context information and training data to avoid these issues. Additionally, we found that including augmented data in the testing set is necessary to fairly evaluate the effectiveness of data augmentation strategies.

Conclusion: While LLMs can generate high-quality data for MCN, the success of data augmentation depends heavily on the strategy employed. Our study found that few-shot learning, with prompts that incorporate appropriate context and a small, representative set of original data, is an effective approach. The methods developed in this research, including the data quality evaluation framework, LLM-based data augmentation strategies, and procedures for data quality enhancement, provide valuable insights for data augmentation and evaluation in similar deep learning applications.

Availability: https://github.com/RichardLRC/mcn-data-quality-llm/tree/main/evaluation

1. Introduction

Medical Concept Normalization (MCN), also referred to as medical entity linking, is a natural language processing (NLP) task that seeks to map informal medical terms or phrases, often found on social media or other online platforms, to formal medical concepts in standardized medical databases [1,2]. The informal terms are typically sourced from platforms like X, Reddit, and AskaPatient, while the target medical databases include the Unified Medical Language System (UMLS) [3], the Systematized Nomenclature of Medicine — Clinical Terms (SNOMED CT) [4], Medical Subject Headings (MeSH) [5], among others [6,7]. Some examples of MCN tasks are shown in Fig. 1.

MCN was first introduced in the late 1980s in response to the rapid expansion of medical literature, with the goal of enhancing search efficiency by mapping users' search queries to relevant MeSH terms [8]. The significance of MCN grew in the 2000s as the widespread adoption of social media and web applications led to a surge in informal

☆ This research is partially support by NSF grants #2231519, #2244259 and #2225229.
* Corresponding author.
E-mail addresses: haihua.chen@unt.edu (H. Chen), rli14@ncsu.edu (R. Li), ana.cleveland@unt.edu (A. Cleveland), junhua.ding@unt.edu (J. Ding).

https://doi.org/10.1016/j.jbi.2025.104812
Received 7 September 2024; Received in revised form 26 January 2025; Accepted 13 March 2025
Available online 1 April 2025

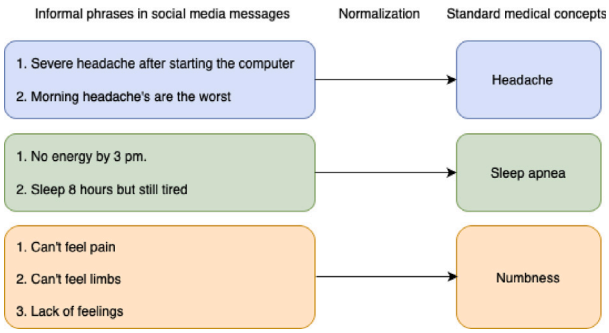


Fig. 1. MCN Examples. Phrases in the left are informal phrases, the arrows indicate MCN task, phrases in the right are the corresponding medical concepts of the informal phrases. Each color represents a MCN task. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

expressions of medical terms. MCN plays a critical role in improving the efficiency, accuracy, and effectiveness of various healthcare applications, including electronic health record (EHR) management, clinical decision support (CDS) systems, health information exchange (HIE), precision medicine (PM), clinical trial retrieval, and automated patient messaging systems. Ultimately, MCN has a direct impact on both patient care and healthcare administration.

As a crucial task in both NLP and healthcare, MCN has garnered significant attention over the past decade. During this time, more than 20 prominent MCN datasets have been released, encompassing various languages, data sources, scales, and purposes. Notable examples include NCBI [9], Cadec [10], AskApatient [11], TwADR-L/S [11], SMM4H 2017 ADR [12], PsyTAR [13], MCN [14], MedRed [15], COMETA [16], WikiMed [17], PubMedDS [17], and others. Leveraging these datasets, researchers have developed a range of machine learning approaches for MCN, including shallow learning models [18], deep learning models [19,20], pre-trained language models [21–24], transfer learning models [25,26], and graph neural networks [27].

Recent studies evaluating large language models (LLMs), such as Llama2 and GPT-3, for rare disease concept normalization have shown promising results [28]. However, performance improvements vary across datasets, and the current results are far from sufficient for practical applications. The MCN task is still suffering from several challenges: (1) There is a shortage of high-quality data, and generating such data for MCN is expensive and labor-intensive. (2) Many existing MCN datasets are either of low quality or lack rigorous validation, casting doubt on model performance. Despite the critical importance of data quality, the production of high-quality datasets for MCN is often overlooked due to the high costs and the difficulty in articulating scientific contributions from data creation efforts. However, high-quality data is the cornerstone of modern data-centric AI, including deep learning and generative AI, and directly influences MCN performance [29,30]. The well-known computing adage "garbage in, garbage out" is especially relevant to data-driven AI in the context of medical concept normalization [26], and under-valuing data quality in this field can have disproportionately negative effects on vulnerable populations and contexts [31–33].

Experts in data-centric AI emphasize that systematically evaluating the quality of datasets is crucial for developing high-performance AI systems [34–37]. Such evaluations provide valuable insights for data enhancement and system performance improvement [26,30,31,38]. For instance, Chen et al. [26] investigated the data quality issues and problematic validation process of an over-claimed DL-based MCN system [11]. Based on the investigation results, they proposed different strategies for performance improvement of the MCN system that was built on the low-quality datasets [26]. Similarly, Budach et al. [30] explored empirically the relationship between six data quality dimensions and the performance of fifteen widely used ML algorithms covering the tasks of classification, regression, and clustering, finding that completeness, feature accuracy, label accuracy (correctness) have a high effect, and class balance has a moderate effect on text classification tasks. Other research has demonstrated that the influence of different data quality dimensions on machine learning varies across tasks and scenarios, suggesting that tailored techniques should be employed to enhance data quality [39,40].

Unlike existing studies that primarily focus on model development, this study emphasizes the data quality aspect of medical concept normalization (MCN). Our goal is to explore how large language models (LLMs), particularly ChatGPT, can be applied to improve data quality in MCN. Specifically, we aim to address the following research questions:

- RQ1: What is the most effective strategy for enhancing data quality in MCN?
- RQ2: How can the data quality of generated data in MCN be fairly evaluated?
- RQ3: How can adequate data be developed for model training and evaluation in MCN?
- RQ4: How does data quality impact model performance in MCN?

To address the research questions, this paper first employs ChatGPT-based zero-shot prompting for data augmentation in MCN. We then evaluate the quality of the data generated by ChatGPT along the dimensions of correctness and comprehensiveness. Subsequently, we conduct a series of experiments to analyze the impact of data quality on MCN model performance. Based on these evaluations and analyses, we implement ChatGPT-based few-shot prompting to further enhance both data quality and model performance. The key contributions of this research are summarized as follows:

1. We apply different strategies for data augmentation in MCN using LLMs and evaluate the effectiveness of the strategies on two MCN datasets. Based on the research result, we propose an approach for selecting an effective data augmentation strategy in MCN.
2. We propose two quality metrics: correctness and comprehensiveness, for evaluating the data quality of the augmented data and develop experiments to quantitatively evaluate them.
3. We conduct a series of experiments to investigate the impact of data quality to MCN performance. Based on the results, we propose and test an approach for producing adequate data in MCN using LLMs.

The LLM-based data augmentation process, data quality evaluation methods, and performance improvement strategies discussed in this study can be valuable for machine learning researchers and practitioners in building high-performance systems beyond MCN. The code and datasets used in this research are available on GitHub.[1]

[1] https://github.com/RichardLRC/mcn-data-quality-llm/tree/main/evaluation.

Statement of significance

Summary	Description
Problem	Existing studies on medical concept normalization primarily focus on models and algorithms, often overlooking the vital role of data quality.
What is already known	Although more than 20 prominent medical concept normalization datasets have been released, encompassing various languages, data sources, scales, and purposes, many of them are either of low quality or lack rigorous validation, casting doubt on model performance.
What this paper adds	This study evaluates medical concept normalization performance across varying data quality scenarios and investigates how to leverage these evaluation results to enhance data quality, ultimately improving medical concept normalization performance using large language models (LLMs).
Who would benefit from the knowledge in this paper	Machine learning researchers and practitioners who would like to build high- performance systems in medical concept normalization, other healthcare applications, and beyond.

2. Related work

In this article, we focus on improving data quality using large language models (LLMs) to enhance medical concept normalization (MCN). This research is closely related to several key areas, including medical concept normalization, existing deep neural network and pre-trained language model algorithms for MCN, and data quality evaluation. Additionally, we explore various techniques, including LLMs, for augmenting training datasets.

2.1. Medical concept normalization

Medical concept normalization (MCN) is a fundamental but challenging task in medical domain. There is an increasing attention on MCN in the last decade. Table 1 provides a summary of the major datasets, state-of-the-art algorithms, and their corresponding performance as reported in the existing literature.

Machine learning datasets for medical concept normalization (MCN) have primarily been developed using three approaches: expert annotation [9,14,41], semi-automatic or automatic recognition [17], and patient self-annotation. These corpora are sourced from social media platforms (e.g., Twitter and Reddit), medical forums (e.g., AskaPatient), Wikipedia texts, and biomedical literature (e.g., PubMed abstracts), among others. The annotations typically include mentions of concepts such as drugs, adverse events, symptoms, and diseases, which are linked to their corresponding entries in controlled vocabularies like SNOMED Clinical Terms, AMT, and MedDRA [10].

However, the quality of annotated datasets varies, as it is influenced by several factors, including the data source, annotation guidelines, multi-stage annotation processes, measures of inter-annotator agreement, and the expertise of clinical terminologists. Numerous MCN datasets have been reported to exhibit data quality issues. For instance, both [1,26] identified problems in the AskAPatient and TwADR-L datasets, such as: (1) significant overlap between training and test data, and (2) a limited portion of medical concepts being mapped to informal phrases on platforms like Twitter or AskaPatient, indicating a lack of comprehensiveness and class balance. These data quality issues can introduce bias into models and produce misleading evaluation results. Alarmingly, several recent studies have used these datasets while overlooking these data quality concerns [42].

Different deep learning models, such as convolutional neural network (CNN), recurrent neural network (RNN), bidirectional long short-term memory (Bi-LSTM), and pre-trained language models (PLM) have been applied for the MCN task, as shown in 1. Most advanced algorithms for MCN are based on PLMs, such as BERT and its variants; different neural networks are usually embedded to enhance the training. For example, Li et al. [43] framed named entity recognition (NER) as a word-word relation classification task (called W^2NER), which defined a multi-granularity 2D neural network for better refining the grid representations based on BERT and BiLSTM, then a co-predictor is used to sufficiently reason the word-word relations. W^2NER achieved SOTA performance on Cadec and several other NER datasets. However, medical NER is still different than the MCN task, therefore, W^2NER was not widely implemented as a baseline model in the MCN task. In additional, W^2NER was not evaluated on social media language. Liu et al. [42] proposed SapBERT, a pretraining scheme that self-aligns the representation space of biomedical entities. It offers an elegant one-model-for-all solution to the problem of medical entity linking (MEL, or MCN), achieving SOTA performance on six widely used MCN datasets, including NCBI, MedMentions, AskAPatient, COMETA, and others. A recent study comprehensively evaluated nine recent SOTA MCN models along five axes: accuracy, speed, ease of use, generalization, and adaptability to new ontologies and datasets [44], finding that ArboEL [45] and SapBERT [42] achieved the best performance; however, ArboEL was the most difficult to adapt and reproduce. Therefore, in the paper, we implement SapBERT [42] as our fundamental model for medical concept normalization.

2.2. Approaches for data quality evaluation

The quality of training data significantly influences the efficiency, accuracy, and complexity of machine learning (ML) tasks [46,47]. A lack of high-quality training data has become a major challenge for the effective use of ML, particularly in deep learning applications [26,48,49]. Despite this, both ML researchers and practitioners tend to focus heavily on models and algorithms while undervaluing the importance of data quality [32]. Experts argue that systematically evaluating data quality across intelligently designed dimensions (metrics) and developing strategies to address quality gaps can reduce the need for iterative debugging in the ML pipeline, ultimately improving model performance with less effort from data scientists [26,39,47,48,50,51].

Recently, a survey paper provides valuable guidance for evaluating dataset quality in the field of machine learning by introducing a comprehensive quality evaluation process, which includes a framework for dataset quality evaluation with dimensions and metrics, computation methods for quality metrics, and assessment models [52]. The approaches for data quality evaluation can be divided into two categories: (1) quantitative methods; and (2) qualitative methods. Statistical analysis, experimental study, and empirical evaluation were commonly used quantitative methods. A set of data quality dimensions fit for the purpose of building specific ML applications are identified, and a group of experiments are usually designed to validate the data quality on the pre-selected dimensions for the experimental study. Table 2 summarizes the recent approaches and empirical studies on data quality evaluation using quantitative methods. Quality dimensions, such as relevance, duplication, accuracy, completeness, class balance, and others are evaluated; different ML and DL algorithms, such as transfer learning (TL), reinforcement learning (RL), deep neural embeddings (DNE), active learning (AL), and others, are selected for experimental study for different purposes and tasks, such as intrusion detection, legal text classification, medical concept normalization, and others [26,30,39,50,51,53–55]. These studies also demonstrate that data quality can be quantitatively evaluated, and the evaluation results can guide practitioners to develop more reliable and higher performance machine learning systems.

Table 1
Summary of existing studies on medical concept normalization using different datasets (selected).

Datasets	Url	Purpose	Descriptions	Algorithms	Performance	Year
NCBI (2014)	Link	Disease name entity recognition and normalization	793 PubMed abstracts, 6892 disease mentions, 790 unique disease concepts	SapBERT ResCNN KEBLM	0.9230[a] 0.9240[a] 0.9350[a]	2021 2021 2023
Cadec (2015)	Link	Mapping medical forum posts (AskaPatient) to medical controlled vocabularies	393,618 Wikipedia texts, 1,067,083 medical mentions 57,739 unique UMLS CUIs	BertMCN CODER BioLORD KnowCAGE	0.8995[c] 0.7619[b] 0.6300[a] 0.8710[a]	2021 2021 2024 2024
AskAPatient (2016)	Link	Mapping medical forum posts (AskaPatient) to SNOMED-CT and Australian Medicines Terminology	3749 phrases 1036 medical concepts 156,652 records for training 7926 records for validation 8662 records for testing	MTA-CharCNN SapBERT ULMFit BERT BioBART CODER	0.8465[a] 0.8764[a] 0.7817[a] 0.8491[a] 0.8713[a] 0.7011[a]	2019 2021 2021 2021 2022 2022
TwADR-L (2016)	Link	Normalization of drugs and adverse drug reactions to SIDER 4.1 drug profile databases in English Tweets	1436 distinct twitter phrases 2220 medical concepts 48,057 records for training 1256 records for validation 1427 are used for testing	MTA-CharCNN ULMFit BERT SapBERT BertMCN CODER	0.4646[a] 0.3986[a] 0.4171[a] 0.4513[a] 0.4832[a] 0.3146[a]	2019 2021 2021 2021 2021 2022
SMM4H (2016–2024)	Link	Normalization of AE mentions in English tweets	17,385 tweets for training, 915 tweets for validation, 10,984 tweets fro testing	BioSyn SapBERT BioLORD KnowCAGE	0.3310[a] 0.4340[a] 0.4770[a] 0.8720[a]	2020 2021 2024 2024
PsyTAR (2019)	Link	Patient posts of effectiveness and ADEs associated with psychiatric medications	891 drugs reviews 4813 ARDs, 590 WDs, 1219 SSIs, 792 DIs 916 UMLS concepts	SapBERT CODER Roberta BioLORD	0.7171[b] 0.7291[b] 0.8242[c] 0.6630[a]	2021 2021 2021 2024
MCN (2019)	Link	Mapping medical problem, treatment, and test entities to medical controlled vocabularies	100 discharge summaries 10,919 concept mentions 3792 unique concepts	SapBERT+T NN classifier SciBERT	0.6936[a] 0.8526[a] 0.8700[a]	2022 2023 2023
MedRed (2020)	Link	Normalization of symptom/disease & drug entities in Reddit posts	1980 Reddit posts 974 drug entities 3511 symptom entities	SapBERT ResCNN BioBART	0.5040[a] 0.5500[a] 0.7178[a]	2021 2021 2022
COMETA (2020)	Link	Normalization of SNOMED-CT entities in Reddit posts	100K Reddit posts 19,911 medical entities 4003 specific concepts	ResCNN BioBART KEBLM	0.8010[a] 0.8177[a] 0.8080[a]	2021 2022 2023

Notes: Adverse Events (AEs), Adverse Drug Events (ADEs) Adverse Drug Reaction (ADR), Withdrawal Symptoms (WDs), Sign/Symptoms/Illness (SSIs), Drug Indications (DIs).
In the performance column,
[a] Means accuracy@1.
[b] Means accuracy@3.
[c] Means accuracy@5.

2.3. Data augmentation techniques for quality improvement

Various methods have been employed to generate high-quality data for machine learning tasks with insufficient training data, including generative adversarial networks (GANs) [56], simulation [57], semi-supervised learning [39,58,59], bootstrapping [60], supervised contrastive learning [61], and large language models (LLMs) [62]. For instance, Han et al. [60] introduced an iterative bootstrapping framework for question-answer (QA) data augmentation, which iteratively generates large-scale, high-quality QA data based on an initial seed set of supervised examples. Similarly, Wu et al. [61] proposed a supervised contrastive learning model for text classification, which leverages data quality augmentation. Their approach dynamically trains on screened, high-quality datasets that contain beneficial information for model training, and further augments the selected data using key words with tag information.

Among current data augmentation techniques, large language models (LLMs), such as GPT and its variants, which are trained on large-scale datasets using complex transformer-based architectures, have demonstrated superior performance compared to other methods [62]. LLMs offer several advantages, including natural language generation, contextual understanding, scalability, flexibility, error reduction, and the ability to augment sparse data. These strengths have made LLMs widely adopted for data augmentation in various text classification tasks. Fig. 2 presents different methods in LLM-based data augmentation.

The most popular method in LLM-based data augmentation is direct prompting, as it requires fewer computational resources and is easier to implement. Zero-shot prompting (ZSL) and few-shot prompting (FSL) fall under this category. ZSL leverages LLMs to generate data without any prior examples from the training data, relying solely on specific instructions or labels to guide the generation process [63]. In contrast, FSL provides the LLM with a small set of examples in the prompts, along with task instructions, to guide the model in producing the desired outputs [64]. Recently, LLM-based data augmentation has been applied and evaluated in clinical, biomedical, and healthcare domains, showing promising performance [65,66].

However, ensuring the quality of augmented data, particularly in high-stakes domains such as healthcare, is arguably the most critical and challenging aspect of LLM-based data augmentation [66]. While augmented data typically enhances data diversity, improving a model's generalizability and preventing overfitting, it can also introduce noise and errors, potentially degrading model performance rather than improving it [67]. Our investigation identifies three common methods for controlling the quality of augmented data: (1) selecting augmented data based on the probability scores assigned by the language model, (2)

Table 2
Summary of the recent approaches and empirical studies on data quality evaluation.

Dimension(s)	Evaluation method	Techniques	Findings	Reference
Duplication, overlap in training and test	Experimental study	ML and DL	Data duplication and overlap in a dataset had different performance impacts on the pre-trained models and the classic ML model	Tran et al. [51]
Consistent representation, completeness, accuracy, uniqueness, class balance	Empirical study	Classification, clustering, and regression algorithms	Data quality has a direct impact on machine learning performance, but the impact of different quality dimensions on classification, clustering, and regression tasks are different	Budach et al. [30]
Comprehensiveness, class balance	Experimental study	ML and DL	The insufficient amount of data and class imbalance are the two major data quality issues for legal argument classification.	Chen et al. [39]
Duplication, overlap in training and test	Experimental study	TL	A rigorous evaluation of data quality is necessary for guiding the quality improvement of machine learning	Chen et al. [26]
Data valuation	Experimental study	RL	The proposed meta learning framework can rank the data values for the training dataset efficiently and effectively	Yoon et al. [53]
Relevance	Empirical evaluation	DNE	Relevance can be evaluated from different perspectives, such as the quantity of relevant data and the degree of semantic similarity	Liu et al. [50]
Data bias	Experimental study	AL	The proposed generic formula for Data Quality Index (DQI) can help dataset creators create datasets free of unwanted biases	Mishra et al. [54]
Variety, veracity	Empirical study	DL	The impact of the volume and quality of training data to the performance of deep learning and the importance of the data quality evaluation	Ding et al. [55]

Notes: transfer learning (TL), reinforcement learning (RL), deep neural embeddings (DNE), active learning (AL).

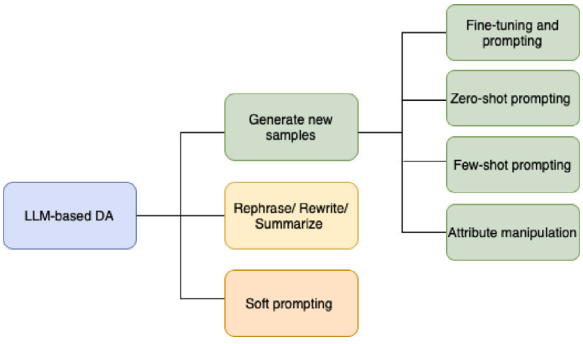


Fig. 2. Taxonomy of data augmentation based large language models, such as GPT.

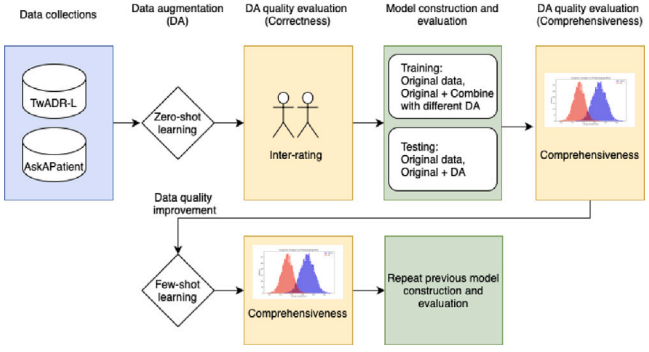


Fig. 3. The workflow of our research. Blue color means data collections, yellow color means data quality evaluation, green color means different training and testing strategies for model construction and evaluation. We performance zero-shot learning and few-shot learning for data quality improvement (data augmentation) in different phases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

selecting data by measuring text similarity, and (3) involving domain experts in the data selection process.

Nevertheless, high-quality data cannot be defined solely by correctness. Even when the generated data is correctly labeled, it does not necessarily lead to improvements in model training. A more precise definition and evaluation of augmented data quality is needed, along with a comprehensive analysis of how different data quality dimensions impact specific downstream applications [66,68–71].

3. Research design and methodology

The research design, as illustrated in Fig. 3, begins with data augmentation using ChatGPT-based zero-shot prompting. After this initial generation, human experts manually verify the correctness of the augmented data to ensure baseline quality. Next, we experiment with different combinations of the original and augmented datasets, using performance analysis to inform iterative improvements in data quality. The following subsections provide detailed descriptions of the data collection process, LLM-based data augmentation techniques, algorithms, baseline models, evaluation metrics, and experimental settings.

Table 3
Summary of the data collections TwADR-L and AskAPatient.

Item	TwADR-L	AskAPatient
Medical concepts	2220	1036
Phrases	1436	3749
Training pairs	48,057	156,652
Validation pairs	1256	7926
Test pairs	1427	8662
Concept mapped to phrases	273	1036
Concept–Concept pairs	24%	38.81%
Training ∩ validation	735	4891
Training ∩ test	861	5224
Duplication in training	17,567	112,537

Notes: (1) Training ∩ Validation and Training ∩ Test are used to check the overlaps among the training datasets, their corresponding validation datasets, and test datasets. An overlap existing in two datasets means the same record exists in the two datasets. For example, if a record in a training dataset is "Hunger– don't want to eat", and there is precisely the same record in a test dataset, then the record is considered as an overlapped record in the two datasets. (2) Duplication means two records are exactly the same; in other words, the same phrase is mapped to the same medical concept.

3.1. Data collections

Two widely used MCN datasets, TwADR-L and AskAPatient, are selected for experiments in this research. The detail information is illustrated in Table 3.

3.1.1. TwADR-L

The TwADR-L dataset, unveiled by Limsopatham and Collier [11], is a specialized aggregation of Twitter posts, specifically curated for medical concept normalization. It emerged from the detailed annotation of Twitter utterances, focusing on the final three months of tweets prior to the dataset's assembly. Aimed at supporting research into adverse drug reactions (ADRs), the dataset includes 1436 unique Twitter expressions, each linked to one or more out of 2220 medical concepts defined within it. The dataset is organized into ten folds, each comprising subsets for training, validation, and testing. In total, the TwADR-L contains 50,740 entries, with each entry pairing an informal phrase to its respective medical concept. Of these, 48,057 entries are designated for training, 1256 for validation, and 1427 for testing purposes.

3.1.2. AskAPatient

The AskAPatient dataset acts as a conduit between informal medical discussions on social media and established clinical ontologies, namely SNOMED-CT and the Australian Medicines Terminology (AMT) [11]. This collection encompasses 3749 phrases from social media, each correlated with one or more of the 1036 medical concepts delineated within SNOMED-CT and AMT [26], facilitating a comprehensive mapping between non-clinical vernacular and professional medical terminologies. Similar to the organizational structure of the TwADR-L dataset, AskAPatient is divided into ten folds as well, where each including designated training, validation, and testing datasets. Overall, AskAPatient boasts 173,240 records, with 156,652 allocated for training purposes, 7926 for validation, and 8662 for testing.

3.1.3. Quality issues of the datasets

According to the studies from [1,26], both of the datasets are suffering from several data quality issues. The first issue is redundancy: both of the datasets have over 50% of their records being duplicates [26], where the same phrases are repeatedly mapped to identical medical concepts. This issue extends across all data partitions; within each fold, a significant overlap exists between the training, validation, and testing datasets. Notably, in the testing datasets, over 60% of the entries duplicate those in the corresponding training datasets, raising concerns about the potential impact on the effectiveness of these datasets in evaluating models.

The second issue is lacking comprehensiveness. In both datasets, a specific category of entry, termed "concept to concept", is identified where the informal phrase directly matches the medical concept. In the testing datasets, such entries constitute 24% in TwADR-L and 38.81% in AskAPatient, further complicating the datasets' utility for precise model assessment.

Moreover, the TwADR-L dataset contains 1436 distinct Twitter phrases, each of them is supposed to be mapped to one or more medical concepts of the 2220 medical concepts; however, only 273 have been linked to their respective informal Twitter phrases. This discrepancy indicates that a significant portion of the medical concepts remain unassociated with informal expressions, highlighting a challenge with the dataset's breadth of coverage. This limitation points to areas where the dataset's comprehensiveness could be improved, as most medical concepts are not directly connected to the collected Twitter phrases.

3.2. Data augmentation with LLMs

In this section, we explore augmentation strategies for both datasets through the application of a Large Language Model (LLM), specifically ChatGPT.

3.2.1. Zero-shot prompt engineering

Addressing the quality issues identified in the Data Collection section, we found significant duplication within both datasets. Specifically, the TwADR-L dataset exhibits a pronounced lack of informal phrases for a majority of its medical concepts. To address these concerns and enhance the datasets' utility, we propose an augmentation strategy aimed at enriching the datasets with a broader array of informal phrases corresponding to each medical concept. This approach seeks to ameliorate the identified deficiencies, thereby increasing the datasets' comprehensiveness and relevance for research purposes. To generate a wide-ranging set of informal expressions corresponding to the formal medical terms identified in the datasets, we utilized the OpenAI API. This process involved the API generating 100 informal phrases for each medical concept, derived from common usages in social texts. The guiding prompt for this generation was: "Please generate 100 informal phrases from social text which can be mapped to the medical concept [medical concept]". This methodology resulted in the accumulation of 10,360 informal phrases for the AskAPatient dataset and 22,200 for the TwADR-L dataset.

3.2.2. Few-shot prompt engineering

To enhance the relevance and fidelity of the data further in alignment with the original datasets, we pursued a few-shot learning tactic as follows.

- We filtered out medical concepts devoid of informal phrases and instances where phrases directly mirrored the concepts. This filtering process selected 924 concepts from AskAPatient and 263 from TwADR-L for augmentation.
- The objective was to create 20 novel phrases for each concept, using prompts enriched with pre-existing examples for guidance.
- In cases where a concept was linked to less than 10 phrases, all available examples were incorporated into the prompt. If more than 10 phrases were associated, a random selection of 10 was used.
- The generation prompt was structured as: "For the given medical concept: [Medical Concept], produce 20 related informal phrases.

This approach refined our data generation process, ensuring a closer match to the quality and context of the initial datasets.

3.2.3. Quality dimensions of data augmentation

Correctness refers to the fact that a record in a dataset is accurate and valid, and they are correctly labeled if they are labeled records. Inaccurate or invalid data lead to data noises, and incorrectly labeled data lead to label noises. Therefore, a correct dataset should contain minimal label noises and data noises. In this study, correctness is calculated by the following formula:

$$Correctness = \frac{\#\ correctly\ labeled\ records}{Total\ \#\ records\ generated}$$

To evaluate correctness, we randomly sample 5 informal phrases for each medical concept. Each record is evaluated by two graduate students, who have the background in health informatics, independently to assure the quality. The students are requested to label each record as correct or incorrect. We calculate the agreement score using Cohen's Kappa value between the two students. Cohen's kappa is widely used statistic to measure inter-rater agreements between two annotators [72]. The value of Cohen's kappa ranges between −1 and 1. Generally, a kappa of 0.8 or above is considered stable [72].

Comprehensiveness in this study means that ChatGPT generated data should contain all representative samples from the initial dataset or be semantically as similar (or close) as the initial dataset. Specifically, we introduce the embedding similarity analysis between ChatGPT generated data and the initial data for measuring the comprehensiveness. More specifically, we utilize the BERT model (bert-base uncased model) to calculate the cosine similarities between the embeddings of informal phrases conveying identical medical concepts across the datasets. By employing bootstrap techniques, we generated distributions that reflect the semantic variances between the original and GPT-generated datasets. Comprehensiveness analysis allows us to gauge the semantic differences accurately, which provides guidance for refining our data generation strategies for better semantic consistency between the datasets. We implement the following steps for the comprehensiveness evaluation:

1. Data selection: We refine the original datasets to identify medical concepts each represented by over N (as the threshold) informal phrases, and then extract all informal phrases from both the original and GPT-generated datasets. We eliminate concept-to-concept mappings within each dataset to maintain focus on phrase-level analysis. In a bid to explore the impact of data uniqueness on our findings, we conduct parallel experiments: one with duplicate entries removed to ensure dataset uniqueness and another without duplicate removal. This dual approach allows us to assess the influence of data redundancy on the cosine similarity distribution across the datasets.
2. Bootstrap iteration: For each iteration of the bootstrap process, we randomly select one medical concept from each subset with replacement, ensuring the selection is proportional to the dataset size. This approach facilitates a balanced representation of concepts across iterations.
3. Similarity calculation: Within each selected medical concept, we compute the pairwise cosine similarity for all informal phrases. The average of these similarity scores for a concept provided a measure of its semantic coherence. After computing the averages for all selected concepts, we calculate the overall average similarity score for each bootstrap iteration. This step is pivotal in quantifying the semantic similarity at the dataset level.
4. Distribution generation: Repeating the bootstrap process 5000 times, we create a distribution of overall average similarity scores. This extensive repetition ensures the robustness of our analysis, offering a detailed view of the semantic landscape of our datasets.

3.3. Medical concept normalization algorithm

SapBERT, self-aligning pretrained BERT [42], served as the backbone model in this research. Pre-trained transformer-based models have demonstrated significant advancements in Natural Language Processing (NLP), particularly for specialized domains such as biomedical text processing [73]. Unlike general-purpose language models such as BERT, which may lack domain-specific coverage, SapBERT builds upon biomedical domain-specific backbone models through a self-alignment fine-tuning process. These backbone models, including Bio-BERT [74], ClinicalBERT [75], UMLSBERT [76], and PubMedBERT [77], are pretrained on biomedical corpora or clinical texts, providing rich domain-specific features essential for biomedical tasks. The flexibility of SapBERT allows it to leverage any of these domain-specific backbones, aligning their embeddings using the Unified Medical Language System (UMLS), which enhances the representation of biomedical terms through self-supervised learning. This approach ensures adaptability across various biomedical contexts, making SapBERT highly effective for medical concept normalization tasks.

Building upon this foundation, SapBERT introduces a self-alignment mechanism that optimizes the representation space of biomedical entities. This process clusters synonymous terms closely while pushing non-synonymous terms apart, effectively modeling semantic relationships.

Formal Definition: Given a biomedical term x and its categorical label y, SapBERT's objective is to learn a function $f(\cdot;\theta) : X \rightarrow \mathbb{R}^d$ that maps terms to a d-dimensional embedding space, where θ are the model parameters. For terms x_i and x_j, the model maximizes the cosine similarity $\langle f(x_i), f(x_j)\rangle$ for synonymous pairs while minimizing it for non-synonymous pairs, thereby improving clustering and semantic grouping.

Online Hard Pairs Mining: SapBERT employs an Online Hard Pairs Mining technique, which selects the most challenging positive and negative pairs within a mini-batch during training. For an anchor x_a, a positive match x_p (sharing the same concept), and a negative match x_n (a different concept), a triplet (x_a, x_p, x_n) is formed. Only triplets where the negative is closer to the anchor than the positive by a margin λ are retained. This focuses training on hard-to-classify pairs, improving discrimination capacity and overall embedding quality.

3.4. Baselines

In addition to SapBERT [42], We also implemented the following baselines in this study:

- **Deep neural network (DNN)** [1], includes convolutional neural network (CNN) and recurrent neural network (RNN). CNN is implemented with an input layer, followed by a convolutional layer with multiple filters, a pooling layer, and a final softmax classifier. A 300-dimensional embedding is used to encode each word in the informal phrase, and the output is a CUI representing the corresponding medical concept. For RNN, an unrolled RNN architecture is implemented with input, hidden, and output layers. Gated recurrent unit (GRU) is used to handle the vanishing gradient problem and to efficiently learn long-range dependencies.
- **Multi-task Attentional Character-level Convolution Neural Network (MTA-CharCNN)** [78], contains three components: (1) the main task for medical concept normalization, which takes a text sequence as input and the corresponding target concept category as output; (2) the auxiliary task, which aims to generate character-level domain-related importance weights of the input text sequence; (3) the joint learning of two tasks, which aims to learn all the parameters jointly by minimizing the overall loss function.

- **BERT + fine-tuning** [26], only fine-tunes BERT language without fine-tuning the classifier model. The output of the final transformer layer of the BERT language model is then used as the feature sequences to be fed to the classifier for the medical concept normalization. In this model, BERT-Base-Uncased is used, fine-tuning is conducted on the AskAPatient and TwADR-L, respectively.
- **BioBART** [79] is a biomedical auto-regressive generative language model, which is pretrained on the biomedical corpora (PubMed abstracts). BioBART adopts BART (Bidirectional and Auto-Regressive Transformers), a generative pretrained language model which achieves SOTA results on different NLG tasks in the general domain [80]. BioBART achieves outstanding performance on multiple MCN datasets, include MedMentions, BC5CDR, NCBI, COMETA, and AskAPatient.
- **CODER** [81], which stands for contrastive learning on knowledge graphs for cross-lingual medical term representation, leverages a contrastive learning framework on a medical knowledge graph, specifically the Unified Medical Language System (UMLS). CODER is designed to generate close vector representations for different terms that represent the same or similar medical concepts, with support across multiple languages. It is trained by contrasting positive and negative term pairs, incorporating relational knowledge from the knowledge graph into the embeddings. This relational knowledge is essential for medical term normalization, helping to capture semantic connections between terms that share related concepts or treatments. CODER shows superior performance in zero-shot term normalization, semantic similarity, and relation classification tasks across various benchmarks, outperforming several SOTA biomedical embeddings include Cadec and PsyTar.
- **BioBERT** [74], a domain-specific adaptation of BERT, is pretrained on biomedical corpora such as PubMed and PMC to handle domain-specific vocabulary and context. It retains BERT's architecture and employs WordPiece tokenization for handling out-of-vocabulary terms. Fine-tuning on tasks like Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA) demonstrates its superior performance, achieving notable improvements over state-of-the-art models. In this research, BioBERT is utilized as the backbone model, further fine-tuned on SapBERT to enhance medical concept normalization performance.
- **UMLSBERT** [76], a domain-specific adaptation of BERT, incorporates structured clinical knowledge from the Unified Medical Language System (UMLS) Metathesaurus. It enhances contextual embeddings by linking words sharing the same concept and leveraging semantic type embeddings to create clinically meaningful representations. Pre-trained on the MIMIC-III dataset, UMLSBERT outperforms BioBERT and Bio_ClinicalBERT in clinical Named Entity Recognition (NER) and natural language inference tasks. This research utilizes UMLSBERT as the backbone model, further fine-tuned on SapBERT to enhance medical concept normalization performance.
- **ClinicalBERT** [75], a domain-specific adaptation of BERT, is pretrained on clinical notes from the MIMIC-III dataset to capture the linguistic characteristics of clinical narratives. ClinicalBERT improves performance on tasks such as Named Entity Recognition (NER) and natural language inference (NLI) by fine-tuning general-domain BERT and BioBERT models with clinical data. It is particularly effective in modeling domain-specific terminology and context, outperforming general-domain models in non-de-identification tasks. In this research, ClinicalBERT serves as the backbone model, further fine-tuned on SapBERT for enhanced medical concept normalization.
- **PubMedBERT** [77] a biomedical domain-specific BERT model, is pre-trained from scratch using over 14 million PubMed abstracts, comprising 3.1 billion words. This dataset covers a wide range of biomedical topics, ensuring domain-relevant vocabulary and contextual representations. Unlike models that adapt general-domain BERT through continual pretraining, PubMedBERT demonstrates superior performance across biomedical NLP tasks, including Named Entity Recognition (NER) and relation extraction, due to its domain-specific vocabulary and pretraining corpus. In this research, PubMedBERT is utilized as a backbone model, further fine-tuned on SapBERT to enhance medical concept normalization.

3.5. Performance evaluation metrics

In this study, we evaluated the performance of our models on two datasets, AskAPatient and TwADR-L, utilizing top-N (N = 1 or 5) accuracy metrics, which measures the proportion of times the model's most confident prediction (i.e., the highest ranked prediction) matches the correct medical concept exactly. Top-N accuracy is calculated with the following formula:

$$\text{Top-N Accuracy} = \frac{\text{Number of correct top-N predictions}}{\text{Total number of predictions}}$$

where N equals to 1 or 5. Top-N accuracy is particularly useful in settings where multiple plausible predictions may be acceptable, as is often the case in medical applications.

3.6. Experiment setting

Our experiment settings were summarized in Table 4, we design experiments to test the impact of different training data, data augmentation methods, duplication, concept–concept mappings, data augmentation size, and testing data on the model performance regarding TwADR-L and AskAPatient datasets, respectively.

4. Experimental results: Data quality to MCN performance

In this section, we first present the initial experiment results on the two original datasets, TwADR-L and AskAPatient, by implementing six SOTA models for medical concept normalization. We then describe the quality evaluation results of the augmented data with ChatGPT and also analyze the impact of data augmentation of the model performance of MCN. In addition, we discuss the lessons we learn from the data quality evaluate, which provides us guidance for data quality improvement.

4.1. Initial results

The initial experimental results on the two original datasets using the baseline models are presented in Table 5. Among these, MTA-CharCNN and SapBERT achieve the best performance on *TwADR-L* and *AskAPatient*, respectively. When considering both datasets, SapBERT demonstrates superior performance compared to other baselines, aligning with findings from prior studies [44,82,83]. Given this observation, we select SapBERT as the fundamental model for this study and focus on analyzing the impact of data quality on MCN performance. Furthermore, we adopt PubMedBERT as the backbone for SapBERT in this study, based on its strong performance and balanced stability across the datasets. Specifically, PubMedBERT fine-tuned with SapBERT achieves a top-1 accuracy of 87.64% on *AskAPatient*, outperforming all other configurations, including BioBERT, ClinicalBERT and UMLSBERT, etc. On the *TwADR-L* dataset, while BioBERT (46.20%) and UMLSBERT (46.13%) slightly surpass PubMedBERT (45.13%), the performance differences are marginal (within 1.07%). Considering the overall consistency and stability observed across both datasets, PubMedBERT strikes a balance between performance and domain-specific adaptability, making it a suitable choice as the backbone model for SapBERT in this

Table 4
Experiment framework. The same experiments were conducted on the TwADR-L and AskAPatient datasets, respectively.

Training	DA method	Duplication	Concept–concept	DA size	Testing	Purpose
Org-Tr	–	w	w	–	Org-Te	Test the impact of DA, duplication, concept–concept mapping, and different test settings
Org-Tr	–	w/o	w	–	Org-Te	
Org-Tr	–	w	w/o	–	Org-Te	
Org-Tr	–	w/o	w	o	Org-Te + DA	
Org-Tr	FSL	w	w/o	20	Org-Te + DA	
Org-Tr	FSL	w/o	w/o	20	Org-Te + DA	
DA	ZSL	w	w	100	Org-Te	Test models trained with DA, tested with original data with different settings
DA	ZSL	w/o	w	100	Org-Te	
DA	ZSL	w/o	w	m	Org-Te	
Org-Tr + DA	ZSL	w	w	100	Org-Te	Test different DA methods and sizes, different testing data for DA under different settings
Org-Tr + DA	ZSL	w	w	m	Org-Te + DA	
Org-Tr + DA	ZSL	w/o	w	100	Org-Te	
Org-Tr + DA	ZSL + context	w/o	w	n	Org-Te	
Org-Tr + DA	ZSL + synonyms	w/o	w	n	Org-Te	
Org-Tr + DA	FSL	w	w/o	o	Org-Te	
Org-Tr + DA	FSL	w/o	w/o	o	Org-Te	
Org-Tr + DA	FSL	w	w/o	n	Org-Te + DA	
Org-Tr + DA	FSL	w/o	w/o	n	Org-Te + DA	

Notes: original training (Org-Tr), original testing (Org-Te), data augmentation (DA) with GPT 3.5. Parameters: m = 1,5,10,20,40,80, n = 5, 10, 20, o = 5,10,20,40,80. The difference between o and m lies in the addition of a single data point. Adding just one data point shows no significant performance difference compared to the original model.

Table 5
The performance of SapBERT and other baselines models on data collections TwADR-L and AskAPatient regarding top-1 accuracy.

Model	TwADR-L	AskAPatient
CNN (2017)	19.46	55.46
RNN (2017)	25.30	65.04
MTA-CharCNN (2019)	**46.46**	84.65
BERT + fine-tuning (2021)	41.71	84.91
PubMedBERT (2020)	45.13	**87.64**
+ SapBERT (Fine-tuned)		
BioBERT (2020)	46.20	85.94
+ SapBERT (Fine-tuned)		
ClinicalBERT (2019)	45.06	87.23
+ SapBERT (Fine-tuned)		
UMLSBERT (2020)	46.13	86.38
+ SapBERT (Fine-tuned)		
BioBART (2022)	–	87.13
CODER (2022)	31.46	70.11

study. Building on these results, SapBERT demonstrates remarkable adaptability in handling complex medical terminology through its advanced pretraining strategies and optimization techniques. Its precise alignment of biomedical terms with their corresponding concepts has established new performance benchmarks across multiple MEL datasets, including *AskAPatient* and *TwADR-L*. Given its consistent performance and stability, SapBERT provides a solid foundation for exploring the impact of data quality on medical concept normalization tasks in this study. For simplicity, in the following sections, we refer to the PubMedBERT-based SapBERT model simply as SapBERT, unless stated otherwise.

4.2. Quality evaluation for data augmentation

In this step, we follow the method in Section 3.2.1 to generate the data. As discussed previously, quality evaluation and control is the key to assure that the augmented data can enhance the performance of NLP models instead of decreasing them. Fitting for the application in this study, correctness and comprehensiveness are selected as the most critical data quality dimensions.

4.2.1. Correctness

The total number of records being evaluated and the evaluation results are presented in Table 6.

From Table 6, we observe that the agreement scores for the *AskAPatient* and *TwADR-L* datasets demonstrate consistently high values, reflecting the quality of the generated data. Specifically, for the *AskAPatient* dataset, the agreement score is 0.9342 without few-shot examples and 0.9465 with few-shot examples. Similarly, for the *TwADR-L* dataset, the agreement score is 0.9526 without few-shot examples and 0.9375 with few-shot examples. Details regarding the computation of agreement scores are provided in Section 3.2.3.

4.2.2. Comprehensiveness

We visualize the embedding similarity distributions, as a measurement of comprehensiveness of the dataset in Fig. 4. In Fig. 4, the three sub-figures on the top illustrate the embedding similarity distributions for informal phrases of the AskAPatient, comparing original, GPT-generated, and combined dataset (original dataset + GPT-generated dataset) with or without duplication present. the three sub-figures on the bottom show the same comparison of TwADR-L. In the figure, blue represents the embedding similarity (cosine) distribution after bootstrap iteration for the original dataset, red represents the same distribution of GPT-generated dataset, and green represents the combined dataset. Under each sub-figure, we also encapsulate the mean and standard deviation for the embedding similarity distributions. M_Org, M_GPT, and M_Com donates the average embedding similarity for the original, GPT-generated, and combined datasets, respectively. SD_Org, SD_GPT, and SD_Com donates the variability (standard deviation) within these embedding similarity scores, highlighting the dispersion of data points around the mean value.

From Fig. 4, we make the following observations:

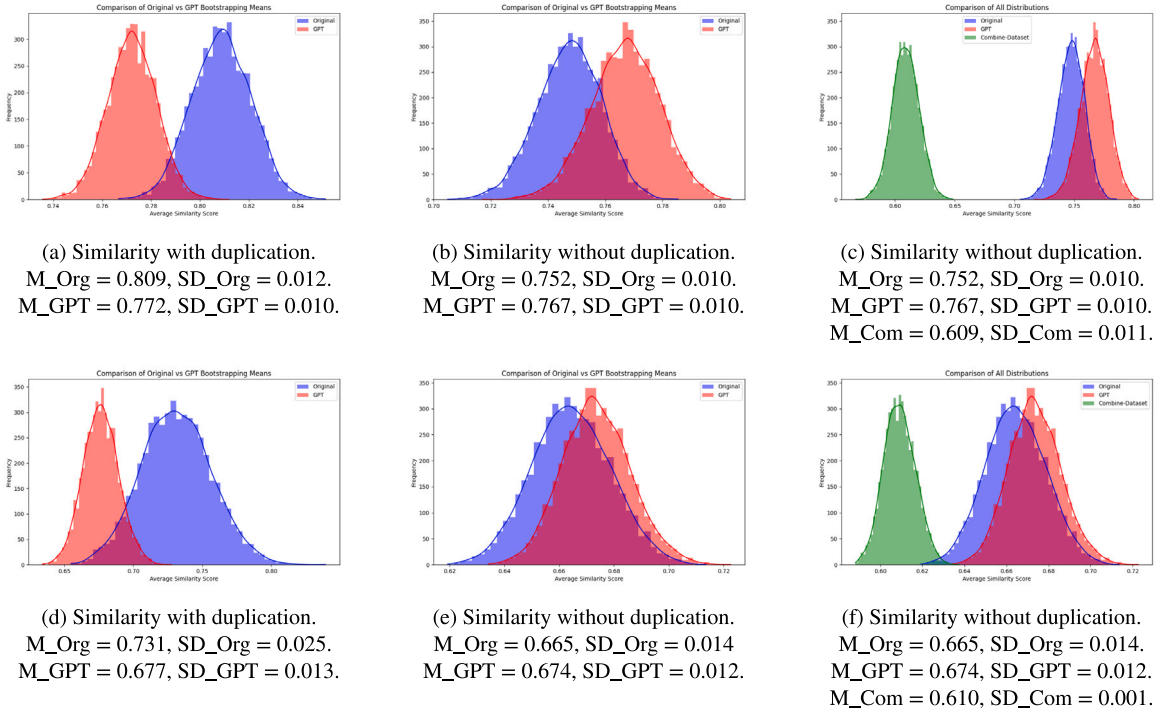
- Duplication has a direct impact on the data quality evaluation regarding comprehensiveness dimension (comparison between a and b for AskAPatient, c and d for TwADR-L).
- The data (informal phrases in this study) generated by ChatGPT using zero-short prompting lacks of comprehensiveness as can be seen from the red distributions: the generated data has a high semantic similarity.
- The distribution of data generated by ChatGPT using zero-short prompting is significantly different than the distribution of the original datasets, indicating that data augmentation with zero-short prompting might distort the original datasets, as can be seen from sub-figures c and f.

4.3. Impact of data augmentation on MCN performance

The quality evaluation results in Section 4.2 provide a better understanding of the data augmentation quality with ChatGPT-based

Table 6
Human evaluation results of the augmented data. 10% randomly sampled records for each concept were used for human evaluation. #C indicates the number of correct labels, #I indicates the number of incorrect labels, and #I indicates the number of conflict labels by the two annotators. A-score indicates the agreement score between the two annotators and DA method represents data augmentation method.

Datasets	#C	#I	#O	A-score	DA method
AskAPatient	4669	170	341	0.9342	Zero-Shot
TwADR-L	10,203	271	526	0.9526	Zero-Shot
AskAPatient	4723	180	277	0.9465	Few-Shot
TwADR-L	10,092	214	694	0.9375	Few-Shot



(a) Similarity with duplication. M_Org = 0.809, SD_Org = 0.012. M_GPT = 0.772, SD_GPT = 0.010.

(b) Similarity without duplication. M_Org = 0.752, SD_Org = 0.010. M_GPT = 0.767, SD_GPT = 0.010.

(c) Similarity without duplication. M_Org = 0.752, SD_Org = 0.010. M_GPT = 0.767, SD_GPT = 0.010. M_Com = 0.609, SD_Com = 0.011.

(d) Similarity with duplication. M_Org = 0.731, SD_Org = 0.025. M_GPT = 0.677, SD_GPT = 0.013.

(e) Similarity without duplication. M_Org = 0.665, SD_Org = 0.014 M_GPT = 0.674, SD_GPT = 0.012.

(f) Similarity without duplication. M_Org = 0.665, SD_Org = 0.014. M_GPT = 0.674, SD_GPT = 0.012. M_Com = 0.610, SD_Com = 0.001.

Fig. 4. Comprehensiveness evaluation results by calculating the BERT similarity within original dataset and ChatGPT-generated dataset (ZSL) separately with duplicated records (a and d), without duplicated records (b and e), combining original and ChatGPT-generated dataset (ZSL) without duplicated records (c and f). (a–c) represent AskAPatient and (d–f) represent TwADR-L. M donates mean value, SD donates standard deviation value.

zero-short prompting. In this section, we will quantify the impact of zero-shot data augmentation on the performance of medical concept normalization models from different perspectives. The experiment results are presented in Tables 7 and 8.

4.3.1. *Duplication of data items*

The impact of duplication on MCN performance, as shown in Tables 7 and 8, is three-fold:

1. Comparing the model performance on the original training and testing dataset with/without duplication (rows 1 and 6 for AskAPatient on Table 7, rows 1 and 6 for TwADR-L on Table 8), we see that the accuracy is 16.36% (top-1) and 6.8% (top-5) higher with duplication of AskAPatient, and 17.17% (top-1) and 19.16% higher with duplication of TwADR-L, indicating the model performance might be over-claimed if trained on the original data directly.
2. Duplication issue in the dataset has a more significant impact on the model trained with GPT-generated data, which is demonstrated in rows 3 & 5 for AskAPatient on Table 7 and rows 3 & 5 for TwADR-L on Table 8. The difference of the model performance is higher on top-1 and top-5 accuracy.
3. In terms of the dataset combined with the original data and GPT-generated data, rows 7 and 8 on both tables, duplication issue also causes misleading (around 20% & 10% for top-1 & top-5 accuracy regarding AskAPatient, and 20% & 25% for top-1 & top-5 accuracy regarding TwADR-L) performance improvement.

The above experiment results is align with the findings in [26], when incorporating data augmentation, the negative impact on MCN models is more significant. Therefore, we try to mitigate this impact by removing duplication from the training and testing data in the remaining experiments.

4.3.2. *Volume of data*

Our previous studies [39,84] have shown that the size of augmentation data also impacts the model performance. More is not always better; more data without meeting the quality requirement may introduce more noise, which will cause a defect in the model. In many scenarios, we need to select the appropriate amount of high-quality data. Therefore, in this study, we incrementally add more GPT-generated data for the training and check the changing of the performance. The purpose is to compare the model performance with different amount of augmentation data to optimize the size being employed in the rest of the experiments. The results are presented in Table 7 for AskAPatient and in Table 8 for TwADR-L.

To better visualize the influence of augmentation data size, we conducted 12 model experiments, which were divided into two strategies: chatGPT training and combined training. The first strategy used only chatGPT-generated data as the training dataset (Gpt-Tr) and tested on the original testing dataset. The second strategy combined original and chatGPT-generated data as the training dataset (Com-Tr) and tested on the same original testing dataset. For each strategy, we experimented with 6 models, each with a different size (1, 5, 10, 20, 40, and 80)

Table 7
Model performance for AskAPatient dataset based on SapBERT under different data quality settings, including with/without duplication, size of the augmentation data, and testing data with/without augmentation data. All the augmentation data (DA) used in this table is based on zero-shot prompting (ZSL) with ChatGPT.

Dataset	Training	DA method	Duplication	Concept–concept	DA size	Testing	Accuracy (%)	
							Top-1	Top-5
	Org-Tr	–	w	w	–	Org-Te	87.64	95.01
	Org-Tr	ZSL	w	w	100	Gpt-Te	32.95	52.69
	Gpt-Tr	ZSL	w	w	100	Org-Te	59.37	73.96
	Gpt-Tr	ZSL	w	w	100	Gpt-Te	83.92	95.17
	Gpt-Tr	ZSL	w/o	w	100	Org-Te	35.70	56.08
	Org-Tr	–	w/o	w	–	Org-Te	71.28	88.21
	Com-Tr	ZSL	w	w	100	Org-Te	84.69	93.35
	Com-Tr	ZSL	w/o	w	100	Org-Te	64.59	83.36
	Org-Tr	–	w	w/o	–	Org-Te	86.66	94.65
	Gpt-Tr	ZSL	w	w	1	Org-Te	60.24	77.54
	Gpt-Tr	ZSL	w	w	5	Org-Te	65.45	85.40
	Gpt-Tr	ZSL	w	w	10	Org-Te	66.47	85.13
	Gpt-Tr	ZSL	w	w	20	Org-Te	64.23	82.00
	Gpt-Tr	ZSL	w	w	40	Org-Te	61.80	77.69
	Gpt-Tr	ZSL	w	w	80	Org-Te	58.78	73.81
	Com-Tr	ZSL	w/o	w	1	Org-Te	71.14	88.30
	Com-Tr	ZSL	w/o	w	5	Org-Te	72.72	89.01
	Com-Tr	ZSL	w/o	w	10	Org-Te	71.53	88.33
AskAPatient	Com-Tr	ZSL	w/o	w	20	Org-Te	68.84	85.97
	Com-Tr	ZSL	w/o	w	40	Org-Te	66.47	85.35
	Com-Tr	ZSL	w/o	w	80	Org-Te	64.92	83.28
	Com-Tr	ZSL	w/o	w	5	Com-Te	66.33	81.54
	Com-Tr	ZSL	w/o	w	10	Com-Te	67.12	83.28
	Com-Tr	ZSL	w/o	w	20	Com-Te	72.29	88.36
	Com-Tr	ZSL	w/o	w	40	Com-Te	80.03	93.22
	Com-Tr	ZSL	w/o	w	80	Com-Te	85.94	95.37
	Org-Tr	ZSL	w/o	w	5	Com-Te	56.32	69.33
	Org-Tr	ZSL	w/o	w	10	Com-Te	46.28	64.48
	Org-Tr	ZSL	w/o	w	20	Com-Te	43.49	60.01
	Org-Tr	ZSL	w/o	w	40	Com-Te	37.32	55.32
	Org-Tr	ZSL	w/o	w	80	Com-Te	35.28	54.19
	Com-Context	ZSL	w/o	w	5	Org-Te	–	88.47
	Com-Synonym	ZSL	w/o	w	5	Org-Te	–	88.63
	Com-Context	ZSL	w/o	w	10	Org-Te	–	87.89
	Com-Synonym	ZSL	w/o	w	10	Org-Te	–	88.27
	Com-Context	ZSL	w/o	w	20	Org-Te	–	88.01
	Com-Synonym	ZSL	w/o	w	20	Org-Te	–	88.08

of the generated data in the training dataset. The results demonstrate that when implementing data augmentation, due to the comprehensiveness issue of the data, the negative impact on the training is increasing with more augmented data. From Table 7, we observe that the model performance initially improves with the incremental addition of GPT-generated data, peaking when the dataset includes 10 samples. Specifically, the Top-1 accuracy increased from 60.24% to 66.47% as the augmentation data size grew from 1 to 10 samples. However, beyond this point, a decline in performance is evident. With 20, 40, and 80 samples, the accuracy drops progressively to 64.23%, 61.80%, and 58.78%, respectively. The Top-5 accuracy follows the same trend, with an initial increase followed by a decrease as the data size grows beyond 10 samples.

A similar trend is displayed in Table 8 for the TwADR-L dataset. The Top-1 accuracy increased from 28.73% to 31.79% as the augmentation data size grew from 1 to 10 samples. However, as the number of samples increased to 20, 40, and 80, the accuracy fell to 29.67%, 26.56%, and 23.72%, respectively. The Top-5 accuracy for the TwADR-L dataset also follows this trend, reinforcing the observation that more data can introduce noise and diminish model performance beyond a certain point.

4.3.3. Context and synonym in prompts

In the context of zero-shot learning, we further investigated the effects of context-specific and synonym-based augmentation strategies on the performance of MCN models. The prompts used for generating these data types were:

- Context-Specific Augmentation: "Please generate 20 informal phrases from social text, each in a specific context or scenario, that can be mapped to the medical concept *[current_formal_concept]*".
- Synonym-Based Augmentation: "Please generate 20 informal phrases from social text that include synonyms or similar meaning words for the medical concept *[current_formal_concept]*".

The results, shown in Tables 7 and 8, reveal interesting trends when comparing the performance of these augmentation methods to that of the original training data.

For the AskAPatient dataset, the original data (Org-Tr) without any augmentation achieved a Top-5 accuracy of 88.21%. When context-specific augmentation (Com-Context) was applied with 5 samples, the Top-5 accuracy slightly increased to 88.47%, and for synonym-based augmentation (Com-Synonym), it was 88.63%. However, as the number of augmented samples increased to 10 and 20, the performance showed fluctuations rather than consistent improvement. The Top-5 accuracy for context-specific augmentation (Com-Context) decreased to 87.89% at 10 samples and slightly recovered to 88.01% at 20 samples. Similarly, for synonym-based augmentation (Com-Synonym), the Top-5 accuracy decreased to 88.27% at 10 samples and showed a minimal increase to 88.08% at 20 samples.

For the TwADR-L dataset, the original data (Org-Tr) without any augmentation achieved a Top-5 accuracy of 47.98%. When context-specific augmentation (Com-Context) was applied with 5 samples, the Top-5 accuracy was 46.54%, and for synonym-based augmentation (Com-Synonym), it was 47.03%, which is lower than the performance of the original data. As the number of augmented samples increased to 10 and 20, both augmentation methods showed a slight decrease in performance. The Top-5 accuracy for context-specific augmentation (Com-Context) decreased from 46.54% with 5 samples to 45.72%

Table 8
Model performance for TwADR-L dataset based on SapBERT under different data quality settings, including with/without duplication, size of the augmentation data, and testing data with/without augmentation data. All the augmentation data (DA) used in this table is based on zero-shot prompting (ZSL) with ChatGPT.

Dataset	Training	DA method	Duplication	Concept–concept	DA size	Testing	Accuracy (%)	
							Top-1	Top-5
	Org-Tr	–	w	w	–	Org-Te	45.13	67.14
	Org-Tr	ZSL	w	w	100	Gpt-Te	32.95	52.69
	Gpt-Tr	ZSL	w	w	100	Org-Te	24.53	40.22
	Gpt-Tr	ZSL	w	w	100	Gpt-Te	83.30	94.93
	Gpt-Tr	ZSL	w/o	w	100	Org-Te	14.79	27.31
	Org-Tr	–	w/o	w	–	Org-Te	27.96	47.98
	Com-Tr	ZSL	w	w	100	Org-Te	40.72	62.30
	Com-Tr	ZSL	w/o	w	100	Org-Te	19.41	38.86
	Org-Tr	–	w	w/o	–	Org-Te	37.70	57.53
	Gpt-Tr	ZSL	w	w	1	Org-Te	28.73	46.20
	Gpt-Tr	ZSL	w	w	5	Org-Te	31.32	47.65
	Gpt-Tr	ZSL	w	w	10	Org-Te	31.79	48.63
	Gpt-Tr	ZSL	w	w	20	Org-Te	29.67	45.89
	Gpt-Tr	ZSL	w	w	40	Org-Te	26.56	41.72
	Gpt-Tr	ZSL	w	w	80	Org-Te	23.72	39.38
	Com-Tr	ZSL	w/o	w	1	Org-Te	27.33	47.04
	Com-Tr	ZSL	w/o	w	5	Org-Te	27.97	48.50
	Com-Tr	ZSL	w/o	w	10	Org-Te	27.48	49.02
TwADR-L	Com-Tr	ZSL	w/o	w	20	Org-Te	26.49	47.26
	Com-Tr	ZSL	w/o	w	40	Org-Te	22.57	42.12
	Com-Tr	ZSL	w/o	w	80	Org-Te	20.38	40.25
	Com-Tr	ZSL	w/o	w	5	Com-Te	41.57	63.38
	Com-Tr	ZSL	w/o	w	10	Com-Te	51.25	71.39
	Com-Tr	ZSL	w/o	w	20	Com-Te	66.47	84.34
	Com-Tr	ZSL	w/o	w	40	Com-Te	74.55	90.25
	Com-Tr	ZSL	w/o	w	80	Com-Te	82.24	94.58
	Org-Tr	ZSL	w/o	w	5	Com-Te	22.94	37.67
	Org-Tr	ZSL	w/o	w	10	Com-Te	20.13	35.28
	Org-Tr	ZSL	w/o	w	20	Com-Te	19.87	35.10
	Org-Tr	ZSL	w/o	w	40	Com-Te	19.56	34.94
	Org-Tr	ZSL	w/o	w	80	Com-Te	19.80	35.06
	Com-Context	ZSL	w/o	w	5	Org-Te	–	46.54
	Com-Synonym	ZSL	w/o	w	5	Org-Te	–	47.03
	Com-Context	ZSL	w/o	w	10	Org-Te	–	45.72
	Com-Synonym	ZSL	w/o	w	10	Org-Te	–	46.38
	Com-Context	ZSL	w/o	w	20	Org-Te	–	44.32
	Com-Synonym	ZSL	w/o	w	20	Org-Te	–	45.41

with 10 samples and further to 44.32% with 20 samples. Similarly, the synonym-based augmentation (Com-Synonym) showed a Top-5 accuracy of 46.38% at 10 samples, decreasing to 45.41% at 20 samples.

When comparing the performance across the two datasets, AskAPatient consistently showed higher baseline accuracy with original data than TwADR-L. However, both datasets exhibited a similar pattern where synonym-based augmentation marginally outperformed context-specific augmentation. Despite these slight improvements, neither augmentation strategy provided a significant enhancement over the accuracy levels achieved by the models trained on the original data alone.

4.3.4. Testing data

In machine learning tasks, it is crucial that the testing data closely follows the distribution of the training data. If augmentation data significantly alters the distribution of the training dataset, the testing dataset should also be updated to reflect these changes. To explore the optimal design of a testing dataset after the inclusion of augmentation data, we conducted a series of experiments using different testing datasets. Specifically, we tested the models on the original testing data, ChatGPT-generated data, and a combination of both original and ChatGPT-generated data.

The results of these experiments for the AskAPatient dataset are detailed in Table 7. This table illustrates the model performance under various data quality settings, including the presence or absence of duplication, different sizes of augmentation data, and different types of testing data.

We observe that using only the original testing data (Org-Te) or only ChatGPT-generated testing data (Gpt-Te) produces distinct outcomes. When the training data consists solely of GPT-generated samples (Gpt-Tr), the model achieves a Top-1 accuracy of 59.37% and a Top-5 accuracy of 73.96% on the original testing dataset. Conversely, when tested on GPT-generated testing data, the model performance significantly improves, achieving a Top-1 accuracy of 83.92% and a Top-5 accuracy of 95.17%. This indicates that the model performs better on data that is similar in distribution to its training set, and this trend was consistent even when duplication was removed.

When combining the original training data with GPT-generated data (Com-Tr) and testing on the original dataset, the model achieves a Top-1 accuracy of 84.69% and a Top-5 accuracy of 93.35%, with scores of 64.59% for Top-1 accuracy and 83.36% for Top-5 accuracy when duplication was removed. Testing on the combined dataset (Com-Te) further improves the performance, with the Top-1 and Top-5 accuracies reaching 85.94% and 95.37%, respectively. Without duplication, the performance was 64.59% and 83.36% for Top-1 and Top-5 accuracy.

Furthermore, we evaluated the impact of different sizes of generated samples within the combined training data (Com-Tr) on the combined testing data (Com-Te) without duplication. For example, with Com-Tr using 5, 10, 20, 40, and 80 samples, the model achieves Top-1 accuracies of 66.33%, 67.12%, 72.29%, 80.03%, and 85.94%, respectively. The corresponding Top-5 accuracies are 81.54%, 83.28%, 88.36%, 93.22%, and 95.37%. This trend shows that increasing the size of the generated samples in the combined training set generally improves model performance on the combined testing set.

In contrast, when the original training data (Org-Tr) was augmented with different sizes of generated samples and tested on the combined testing data (Com-Te), the performance showed a decreasing trend with increased augmentation data size. For instance, with 5, 10, 20, 40,

and 80 samples, the model's Top-1 accuracies were 56.32%, 46.28%, 43.49%, 37.32%, and 35.28%, respectively, while the Top-5 accuracies were 69.33%, 64.48%, 60.01%, 55.32%, and 54.19%.

A similar analysis was conducted for the TwADR-L dataset, as shown in Table 8. Here, we see that the model's performance trends align with those observed for the AskAPatient dataset. Using GPT-generated samples (Gpt-Tr) for training and testing on the GPT-generated testing data yields the highest performance, with Top-1 and Top-5 accuracies of 83.30% and 94.93%, respectively. Conversely, using only the original testing data shows lower performance, highlighting the importance of testing data distribution alignment, with or without duplication.

For the combined training data (Com-Tr) tested on the combined testing data (Com-Te), the model's Top-1 accuracies are 41.57%, 51.25%, 66.47%, 74.55%, and 82.24% for 5, 10, 20, 40, and 80 samples, respectively. The Top-5 accuracies follow a similar trend, reaching up to 94.58%.

4.4. Discussion

The results of our study underscore the critical importance of data quality in medical concept normalization (MCN) when utilizing zero-shot data augmentation with ChatGPT. We found that duplication within datasets can significantly inflate model performance, potentially leading to misleading conclusions if not properly addressed, as noted in prior research [85,86]. This issue was evident across both the AskAPatient and TwADR-L datasets, where models trained on duplicated data demonstrated substantially higher accuracy compared to those trained on de-duplicated data. Moreover, while data augmentation can enhance diversity, our findings align with existing literature [39,84] in showing that simply increasing the quantity of augmented data does not always improve performance; in fact, beyond a certain point, it often introduces noise that diminishes model effectiveness. This was observed in both datasets, where performance peaked at an optimal size of augmented data but declined as more samples were added. Additionally, our exploration of the impact of different testing data on model performance reveals that alignment between the distribution of training and testing data is crucial for accurate evaluation, as supported by previous studies [87,88]. Models trained on GPT-generated data performed significantly better on similar testing data, while those trained on a combination of original and augmented data showed improved performance when tested on a mix of both. These insights highlight the need for careful management of data quality and quantity in MCN tasks.

5. Experimental results: Data quality enhancement for MCN performance improvement

This section presents the outcomes of our efforts to enhance data quality and its subsequent impact on the performance of Medical Concept Normalization (MCN) models. By utilizing few-shot learning (FSL) to generate augmented data, we aimed to improve the comprehensiveness and coherence of the datasets. The data generation and filtering processes, which were crucial in refining the dataset for this purpose, are detailed in Section 3.2.2. The specific enhancement techniques employed and their effects on model performance will be discussed in the following sections (see Fig. 5).

5.1. Comprehensiveness improvement with FSL-based data augmentation

Figs. 4 and 6 illustrate the impact of data augmentation techniques on the comprehensiveness of medical concept normalization. Specifically, we compare the BERT similarity scores between the original dataset, ChatGPT-generated datasets using Zero-Shot Learning (ZSL), and Few-Shot Learning (FSL), as well as their combinations. In Fig. 4, the combined dataset (original + ZSL-generated) is represented by the green plots, showing moderate similarity scores. Conversely, in Fig. 6, the purple plots represent this same combined dataset, while the green plots represent the combined dataset (original + FSL-generated). The green plots achieve higher similarity scores compared to the purple plots, indicating that FSL-generated data aligns more closely with the original dataset.

The increase in similarity scores from ZSL to FSL demonstrates the effectiveness of FSL in generating data that is more representative of the original dataset. This suggests that FSL can produce higher-quality synthetic data [89], improving the overall comprehensiveness and reliability of the augmented dataset. The consistent improvement in similarity scores with the use of FSL-based data augmentation highlights its potential for enhancing data quality. By generating data that more accurately reflects the original dataset, FSL contributes to better performance in medical concept normalization tasks. This comparison underscores the superiority of FSL over ZSL in generating high-quality synthetic data, thus advancing data quality enhancement methodologies in the field of medical informatics.

5.2. Dataset coherence on performance improvement

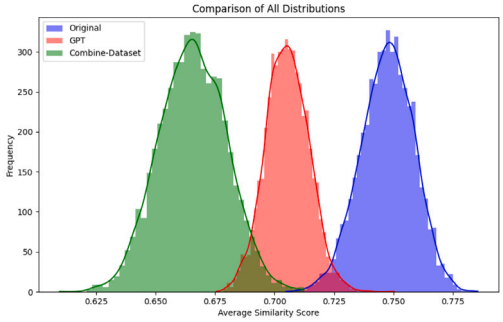
Table 9 illustrates the performance of the SapBERT model under various data quality settings, emphasizing the critical role of dataset coherence in determining model accuracy. The table compares results using the original training dataset (Org-Tr) and a combined dataset (Com-Tr) augmented with GPT-generated data, evaluated with and without data augmentation (FSL) and duplication. By comparing accuracies across identical duplication and concept–concept settings, we can discern the impact of few-shot learning (FSL) with varying sizes of augmentation data on the model's performance.

For the **AskAPatient** dataset, incorporating duplication consistently boosts performance, with a notable peak at 10 augmentation data points, where the model achieves a top-1 accuracy of 88.01%—an improvement over the original dataset's 86.66%. However, as the augmentation size increases to 40 and 80, top-1 accuracy slightly tapers off to 86.37% and 85.82%, respectively, although top-5 accuracy remains relatively stable around 95%. A similar trend emerges for the **TwADR-L** dataset: with duplication, top-1 accuracy climbs from 37.70% (original data) to 47.93% (5 data points), then gradually declines as augmentation expands to 40 (46.93%) and 80 (46.21%).

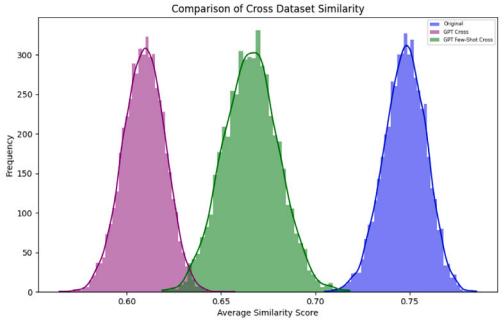
When duplication is removed, moderate amounts of augmentation also yield improvements over the original dataset. For AskAPatient, top-1 accuracy rises from 69.98% to 72.16% with 10 data points but decreases again with 40 (70.03%) and 80 (69.24%). In TwADR-L, similarly, performance gains observed at 5 or 10 data points wane at higher augmentation levels of 40 and 80. Overall, these findings indicate that while introducing a certain quantity of GPT-generated data can significantly enhance classification accuracy, excessively large augmentation sets may introduce noise and diminish gains, underscoring the importance of balancing augmentation size.

Beyond measuring classification accuracy, we further investigated the coherence of GPT-generated data by analyzing the relationship between in-context sample size and data consistency. Specifically, we focused on comparing the characteristics of generated data when prompted with concepts that have *fewer than ten examples* versus those guided by *ten or more examples* in the few-shot learning setup. This analysis aimed to evaluate whether larger in-context examples lead to more coherent and consistent outputs.

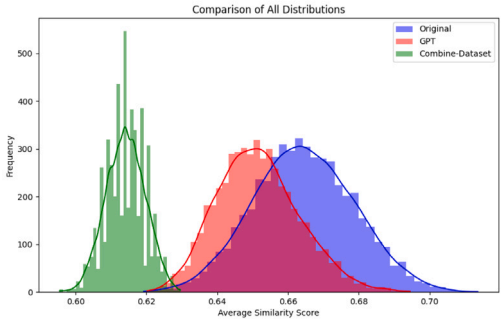
To assess coherence, we utilized BERT-based embeddings to represent each generated informal phrase in both the *TwADR-L* and *AskAPatient* datasets. For each concept, we computed pairwise cosine similarity scores among its generated phrases, measuring the internal consistency of the outputs. To ensure statistical robustness, we conducted a 5,000-iteration bootstrapping procedure. In each iteration, concepts were randomly sampled, and their average within-concept similarity scores were computed to capture distributional trends across different sample sizes.



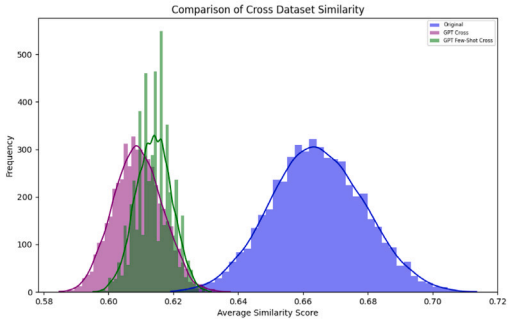
(a) Similarity within different datasets.
M_Org = 0.752, SD_Org = 0.010.
M_GPT = 0.708, SD_GPT = 0.008.
M_Com = 0.666, SD_Com = 0.014.



(b) Similarity cross different datasets.
M_Org = 0.752, SD_Org = 0.010.
M_GPT = 0.609, SD_GPT = 0.011.
M_Com = 0.666, SD_Com = 0.014.



(c) Similarity within different datasets.
M_Org = 0.665, SD_Org = 0.014.
M_GPT = 0.651, SD_GPT = 0.012.
M_Com = 0.614, SD_Com = 0.005.



(d) Similarity cross different datasets.
M_Org = 0.665, SD_Org = 0.014.
M_GPT = 0.610, SD_GPT = 0.007.
M_Com = 0.614, SD_Com = 0.005.

Fig. 5. Comprehensiveness evaluation results by calculating the BERT similarity within the original dataset, ChatGPT-generated dataset (FSL), and combined datasets separately (a and c). In (b) and (d), cross-dataset BERT similarity comparisons are shown between the combined datasets (ZSL + original in purple, and FSL + original in green) and the original dataset. (a and b) represent the AskAPatient dataset, while (c and d) represent the TwADR-L dataset. All similarity scores were calculated after removing duplicated records.

Figs. 6(c) and 6(d) present the results for the *TwADR-L* dataset, while Figs. 6(a) and 6(b) illustrate similar patterns for *AskAPatient*. The line plots in Figs. 6(a) and 6(c) show the mean similarity scores as a function of the few-shot sample size, ranging from 2 to 10. The shaded regions represent the range (minimum–maximum) of similarity scores observed for each sample size.

A key observation is that smaller sample sizes (e.g., 2–4) exhibit higher similarity scores and narrower shaded regions, indicating that the generated data is more internally consistent but potentially lacks diversity. This suggests that when the LLM is guided by fewer examples, it tends to overfit to the patterns present in the prompts, resulting in outputs that are closely aligned but semantically less varied. Conversely, as the sample size increases, particularly beyond 6–10 examples, the mean similarity scores show slightly more fluctuation, and the shaded regions expand. This trend reflects greater variety and semantic diversity in the generated outputs, albeit with slightly reduced internal coherence.

In the *TwADR-L* dataset (Fig. 6(c)), the mean similarity scores exhibit a slight dip around sample size 6, accompanied by a noticeably wider shaded range. However, this pattern is primarily attributed to the smaller number of concepts (only 9) available at this sample size, compared to over 20 concepts for other sample sizes. The limited data introduces higher variance, resulting in a broader range, rather than indicating intrinsic diversity in the generated outputs. A similar trend is observed in the *AskAPatient* dataset (Fig. 6(a)), where the mean similarity peaks at size 6 before slightly declining as the sample size approaches 10. Unlike *TwADR-L*, the shaded ranges remain relatively narrow across sample sizes, reflecting a more stable distribution due to consistently larger concept counts.

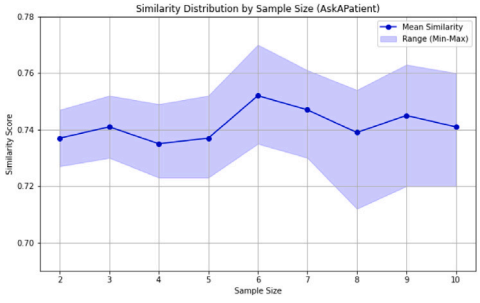
The bootstrapping distributions in Figs. 6(d) and 6(b) further highlight the influence of sample size on the diversity of GPT-generated outputs. For *TwADR-L*, expansions generated using concepts with fewer than ten examples exhibit a higher mean similarity distribution (blue) with a narrower spread, indicating that smaller sample sizes tend to produce outputs that are more internally consistent but less diverse. In contrast, concepts guided by ten or more examples yield a slightly lower mean similarity (red) but with a broader distribution, suggesting increased variability and semantic diversity in the generated data.

A slightly different trend is observed in the *AskAPatient* dataset. Concepts with more than ten examples produce outputs with a slightly higher mean similarity compared to those with fewer than ten examples. However, the spread of similarity scores is noticeably wider, reflecting greater diversity in the generated outputs. In contrast, concepts with fewer than ten examples yield a narrower distribution, indicating higher internal consistency but reduced diversity. The results indicate that larger in-context sample sizes encourage the model to explore broader semantic variations while maintaining reasonable coherence, whereas smaller sample sizes constrain the outputs to be more homogeneous. Consequently, tasks requiring high fidelity may benefit from smaller sample sizes, while tasks prioritizing coverage and generalization may benefit from larger in-context examples.

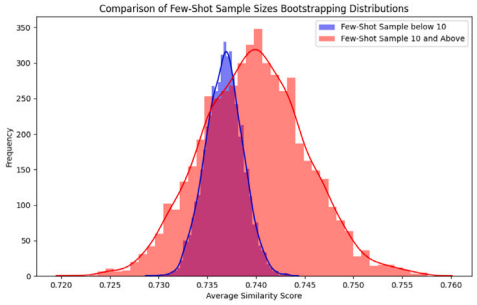
Our analysis highlights the role of dataset coherence and diversity in augmenting training data effectively. Larger in-context examples enable the generation of more diverse outputs, capturing broader semantic

Table 9
Model performance based on SapBERT under different data quality settings, including with/without duplication, size of the augmentation data, and testing data with/without augmentation data. All the augmentation data (DA) used in this table is based on few-shot prompting (FSL) with ChatGPT.

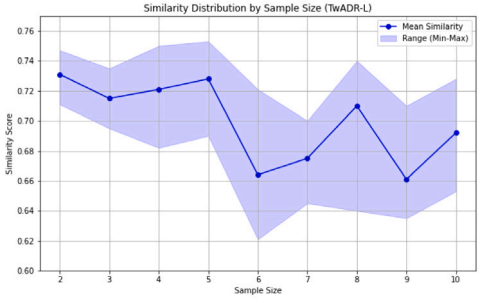
Dataset	Training	DA method	Duplication	Concept–concept	DA size	Testing	Accuracy (%)	
							Top-1	Top-5
AskAPatient	Org-Tr	–	w/o	w/o	–	Org-Te	69.98	87.33
	Com-Tr	FSL	w/o	w/o	5	Org-Te	71.73	88.42
	Com-Tr	FSL	w/o	w/o	10	Org-Te	72.16	88.20
	Com-Tr	FSL	w/o	w/o	20	Org-Te	70.51	88.85
	Com-Tr	FSL	w/o	w/o	40	Org-Te	70.03	87.95
	Com-Tr	FSL	w/o	w/o	80	Org-Te	69.24	86.74
	Org-Tr	–	w	w/o	–	Org-Te	86.66	94.65
	Com-Tr	FSL	w	w/o	5	Org-Te	87.75	95.16
	Com-Tr	FSL	w	w/o	10	Org-Te	88.01	95.06
	Com-Tr	FSL	w	w/o	20	Org-Te	87.02	95.38
	Com-Tr	FSL	w	w/o	40	Org-Te	86.37	95.09
	Com-Tr	FSL	w	w/o	80	Org-Te	85.82	94.94
	Org-Tr	FSL	w	w/o	20	Nt	38.41	57.06
	Org-Tr	FSL	w/o	w/o	20	Nt	30.66	51.08
	Com-Tr (80%)	FSL	w	w/o	20	Nt	59.62	81.34
	Com-Tr (80%)	FSL	w/o	w/o	20	Nt	54.61	78.96
TwADR-L	Org-Tr	–	w/o	w/o	–	Org-Te	28.97	50.15
	Com-Tr	FSL	w/o	w/o	5	Org-Te	32.54	52.31
	Com-Tr	FSL	w/o	w/o	10	Org-Te	32.17	50.72
	Com-Tr	FSL	w/o	w/o	20	Org-Te	32.08	48.57
	Com-Tr	FSL	w/o	w/o	40	Org-Te	30.92	47.39
	Com-Tr	FSL	w/o	w/o	80	Org-Te	29.34	46.71
	Org-Tr	–	w	w/o	–	Org-Te	37.70	57.53
	Com-Tr	FSL	w	w/o	5	Org-Te	47.93	69.16
	Com-Tr	FSL	w	w/o	10	Org-Te	47.72	69.03
	Com-Tr	FSL	w	w/o	20	Org-Te	47.79	67.07
	Com-Tr	FSL	w	w/o	40	Org-Te	46.93	66.67
	Com-Tr	FSL	w	w/o	80	Org-Te	46.21	65.83
	Org-Tr	FSL	w	w/o	20	Nt	17.07	32.17
	Org-Tr	FSL	w/o	w/o	20	Nt	14.02	28.48
	Com-Tr (80%)	FSL	w	w/o	20	Nt	40.58	63.68
	Com-Tr (80%)	FSL	w/o	w/o	20	Nt	39.30	62.44



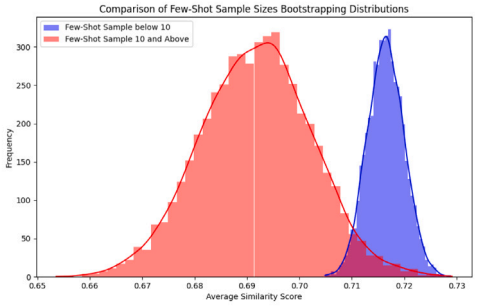
(a) Line plot for different sample size similarity.



(b) Comparison for smaller and above 10 sample size similarity



(c) Line plot for different sample size similarity.



(d) Similarity cross different datasets. Comparison for smaller and above 10 sample size similarity

Fig. 6. Coherence evaluation results by calculating the BERT similarity within the generated dataset. (a and b) represent the AskAPatient dataset, while (c and d) represent the TwADR-L dataset. All similarity scores were calculated after removing duplicated records.

variations, while smaller in-context examples tend to produce more consistent and homogeneous outputs. This trade-off between diversity and coherence underscores the importance of tailoring few-shot prompting strategies to specific task requirements, balancing precision and variability based on downstream applications.

5.3. Data augmentation settings on performance improvement

5.3.1. Duplication of data items

In alignment with our previous zero-shot duplication analysis, we investigated the impact of duplication on few-shot learning. During these experiments, we excluded all concept-to-concept data points where the informal phrase and medical concept were identical, as these do not aid in improving model performance. The original training data contains duplication, which has a significant effect on model accuracy. For the AskAPatient dataset presented in Table 9, the removal of duplication leads to a noticeable drop in performance. Specifically, without duplication, the original training data achieves a top-1 accuracy of 69.98% and a top-5 accuracy of 87.33%. In contrast, with duplication, these accuracies rise to 86.66% and 94.65%, respectively. A similar pattern is observed for the TwADR-L dataset, where the absence of duplicated data points significantly reduces the model's accuracy. Without duplication, the model achieves a top-1 accuracy of 28.97% and a top-5 accuracy of 50.15%. However, the presence of duplication improves the accuracies to 37.70% for top-1 and 57.53% for top-5. These findings underscore the complexities introduced by duplication in the original data [90], which can hinder the model's ability to generalize effectively across diverse concepts.

5.3.2. Influence of data volume

Table 9 provides a comprehensive evaluation of the influence of data augmentation size on model performance across two datasets, *AskAPatient* and *TwADR-L*. The experiments systematically varied the augmentation size, evaluating the model under both small-scale (5, 10, and 20 data points) and larger-scale (40 and 80 data points) few-shot learning scenarios. The augmented datasets were generated using few-shot prompting techniques, allowing the model to incorporate varying degrees of additional context.

The results indicate that moderate levels of augmentation improve accuracy, but diminishing returns and slight degradations appear at larger augmentation sizes. For the *AskAPatient* dataset, in the absence of duplication, top-1 accuracy improved from 69.98% (original data) to 71.73%, 72.16%, and 70.51% with 5, 10, and 20 additional data points, respectively. However, as augmentation increased further to 40 and 80 data points, performance slightly dropped to 70.03% and 69.24%. A similar trend was observed in the top-5 accuracy, which peaked at 88.85% with 20 additional points before declining to 87.95% and 86.74%.

This trend is mirrored in the *TwADR-L* dataset. Top-1 accuracy increased from 28.97% to 32.54%, 32.17%, and 32.08% with 5, 10, and 20 points, but larger augmentation sizes (40 and 80) led to slight decreases, with scores dropping to 30.92% and 29.34%. Similarly, top-5 accuracy peaked at 52.31% before declining to 47.39% and 46.71%.

The findings highlight the role of few-shot learning in leveraging limited data to improve model performance. Even with minimal to moderate augmentation (e.g., 5–20 examples), noticeable performance gains were observed, demonstrating the effectiveness of few-shot prompting in capturing task-relevant patterns. However, the diminishing returns observed at larger augmentation sizes emphasize the need to optimize data volume, balancing between coherence and diversity to prevent performance degradation. These insights are particularly critical for tasks such as medical text processing, where both precision and generalization are essential.

5.3.3. Evaluation using new testing set (Nt)

To address the challenge of model generalization in real-world scenarios, a novel testing set, referred to as N_t, was introduced. N_t was created by combining 20% of the generated data, not used in training, with the original test data (t0). This approach allowed us to evaluate the model's performance on a more diverse and realistic dataset. For the AskAPatient dataset, the model trained on the combined dataset and tested on N_t shows top-1 accuracies ranging from 38.41% to 59.62% and top-5 accuracies from 57.06% to 81.34%. In contrast, for the TwADR-L dataset tested on N_t, top-1 accuracies range from 17.07% to 40.58%, and top-5 accuracies from 32.17% to 63.68%. These variations demonstrate the challenges and the potential of using such a mixed test set to truly evaluate the model's adaptability and generalization to new, unseen data, providing a more comprehensive assessment of its performance in real-world settings.

5.4. Discussions

Our study demonstrates that few-shot learning (FSL) significantly outperforms zero-shot learning (ZSL) in data augmentation for medical concept normalization (MCN), primarily by enhancing the comprehensiveness and alignment of generated data with the original dataset. FSL-generated data not only exhibits higher semantic similarity to the original data but also improves model generalization as highlighted by research [89], leading to more robust and accurate performance across various testing scenarios. Importantly, our findings underscore the nuanced impact of data quality dimensions — volume, accuracy, and comprehensiveness — on AI performance. While increasing the volume of augmented data can enhance model accuracy, it must be done judiciously to avoid introducing noise that could undermine performance, as cautioned by previous studies [91,92]. Additionally, ensuring the accuracy of the data, particularly in avoiding issues like duplication, is crucial for reliable model outcomes [90]. These insights highlight the superiority of FSL in generating high-quality synthetic data, making it a more effective approach for advancing MCN tasks in practice.

6. An approach for quality evaluation and improvement for deep learning

In this section, we introduce a framework aimed at automating the quality evaluation and enhancement of datasets, with a specific focus on Medical Concept Normalization (MCN) tasks. The core of this approach leverages BERT-based similarity measurements to assess the semantic quality of data. By analyzing the similarity between data points through BERT embeddings, we can evaluate the degree of redundancy and diversity within the dataset. This method ensures that the data remains comprehensive and accurately reflects the domain, which is crucial for the effective normalization of medical concepts.

To further improve the dataset, we employ prompt engineering techniques using ChatGPT, with a particular emphasis on leveraging few-shot learning (FSL). In this approach, prompts are carefully designed not only to generate new data that aligns closely with the original dataset but also to introduce meaningful variation. By using FSL, we provide the model with a few examples of original data, which serve as a reference or sample to guide the generation process. These examples ensure that ChatGPT produces data that is consistent with the nuances and characteristics of the existing dataset.

The prompts typically combine action keywords, such as "Generate" or "Paraphrase", with specific inputs, which in the case of FSL, include a small, representative set of original data points. This selection of inputs plays a critical role in the quality of the output, as the model uses these examples to understand the desired structure, tone, and content. By strategically choosing these examples, we can exert precise control over the relevance, diversity, and novelty of the generated data,

ensuring that it remains contextually appropriate and adds value to the dataset.

This FSL-driven approach allows for the efficient creation of high-quality, novel data while minimizing the need for extensive manual intervention. By iteratively refining prompts and evaluating the output against BERT similarity scores, we can optimize the balance between similarity to the original data and the introduction of useful variations. This process ultimately enhances the performance and robustness of deep learning models in Medical Concept Normalization tasks, ensuring that the generated data not only enriches the dataset but also aligns closely with the specific requirements of the domain.

7. Conclusion

In this study, we explored the application of large language models (LLMs), particularly ChatGPT, to improve data quality in medical concept normalization (MCN) task. We focused on two widely-used datasets, TwADR-L and AskAPatient, to evaluate the impact of various data augmentation strategies on MCN task performance. Our approach involved utilizing zero-shot and few-shot prompting techniques to generate additional training data, which were then rigorously evaluated for correctness and comprehensiveness. To quantify the semantic coherence between the original and ChatGPT-generated data, we employed BERT-based embedding similarity analysis, calculating the cosine similarity between embeddings of phrases associated with the same medical concept. This allowed us to assess the extent to which the augmented data retained the semantic richness of the original dataset. Additionally, we conducted extensive experiments to examine the effects of data augmentation size, duplication, and specific augmentation methods (context-specific and synonym-based) on model performance.

Our research demonstrated that while LLMs like ChatGPT are capable of generating high-quality data for MCN tasks, the effectiveness of data augmentation varies significantly depending on the strategy employed. Zero-shot data augmentation, though effective in increasing data quantity, sometimes introduced semantic drift, particularly when overused, which led to diminished model performance. These findings highlight the importance of balancing augmentation size and maintaining data comprehensiveness to avoid the pitfalls of data noise. In contrast, few-shot learning techniques proved to be more effective than zero-shot approaches. The data generated through few-shot learning exhibited higher semantic similarity to the original dataset, indicating that providing the model with a few examples enables it to produce augmented data that better preserves the contextual integrity of the original data. This suggests that few-shot learning is a more reliable strategy for generating high-quality augmented data in MCN tasks.

Despite these promising results, our study has several limitations. First, the reliance on ChatGPT for data generation may introduce biases inherent to the LLM itself, which were not fully explored or mitigated in this study. Additionally, the datasets used — TwADR-L and AskAPatient — have intrinsic data quality issues, such as duplication and lack of comprehensiveness, which may have influenced the outcomes. Lastly, our experiments were conducted on a limited number of datasets, which may affect the generalizability of our findings to other MCN datasets or domains.

Building on the insights gained from this study, future research should focus on several key areas. First, a deeper investigation into the biases introduced by LLMs during data generation is necessary to enhance the reliability of augmented data. These biases might influence the fairness and accuracy of the resulting datasets, impacting the overall model performance [93,94]. Additionally, collaborating with linguistic experts to design more effective prompts will lead to more contextually appropriate and semantically rich outputs [95], which might improve the quality and relevance of the generated data. Finally, future work could explore the integration of more advanced data augmentation techniques, such as domain-specific fine-tuning of LLMs or hybrid approaches that combine LLM-generated data with traditional augmentation methods. These strategies are essential for tailoring LLMs to specific tasks, thereby enhancing the performance of MCN models.

Table 10
Abbreviations and acronyms used in this article.

Abbreviations	Full name of concepts
MCN	Medical Concept Normalization
LLMs	Large Language Models
NLP	Natural Language Processing
ZSL	Zero-Shot Learning
FSL	Few-Shot Learning
ORG	Original Dataset
COM	Combined Dataset
DA	Data Augmentation
ML	Machine Learning
DL	Deep Learning
TL	Transfer Learning
RL	Reinforcement Learning
CNE	Deep Neural Embedding
CNN	Convolutional Neural Network
DNN	Deep neural network
AL	Active Learning
NER	Name Entity Recognition
UMLS	Unified Medical Language System
MEL	Medical Entity Linking

CRediT authorship contribution statement

Haihua Chen: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ruochi Li:** Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Ana Cleveland:** Writing – review & editing, Validation, Investigation, Formal analysis. **Junhua Ding:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

After completing the paper, we employ ChatGPT-4o to identify writing typos. Subsequently, manual review and revision are performed to address these typos.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is partially support by NSF grants #2231519, #2244259 and #2225229.

Appendix. Abbreviations

See Table 10.

References

[1] K. Lee, S.A. Hasan, O. Farri, A. Choudhary, A. Agrawal, Medical concept normalization for online user-generated texts, in: 2017 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2017, pp. 462–469.
[2] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, V. Varma, Distant supervision for medical concept normalization, J. Biomed. Inform. 109 (2020) 103522.
[3] X. Jing, The unified medical language system at 30 years and how it is used and published: systematic review and content analysis, JMIR Med. Inform. 9 (8) (2021) e20675.

[4] E. Chang, J. Mostafa, The use of SNOMED CT, 2013–2020: a literature review, J. Am. Med. Inform. Assoc. 28 (9) (2021) 2017–2026.

[5] N. Baumann, How to use the medical subject headings (MeSH)., Int. J. Clin. Pract. 70 (2) (2016) 171–174.

[6] J.E. Harrison, S. Weber, R. Jakob, C.G. Chute, ICD-11: an international classification of diseases for the twenty-first century, BMC Med. Inform. Decis. Mak. 21 (2021) 1–10.

[7] H. Le, R. Chen, S. Harris, H. Fang, B. Lyn-Cook, H. Hong, W. Ge, P. Rogers, W. Tong, W. Zou, RxNorm for drug name normalization: a case study of prescription opioids in the FDA adverse events reporting system, Front. Bioinform. 3 (2024) 1328613.

[8] H.J. Lowe, G.O. Barnett, Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches, Jama 271 (14) (1994) 1103–1108.

[9] R.I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.

[10] S. Karimi, A. Metke-Jimenez, M. Kemp, C. Wang, Cadec: A corpus of adverse drug event annotations, J. Biomed. Inform. 55 (2015) 73–81.

[11] N. Limsopatham, N. Collier, Normalising medical concepts in social media texts by learning semantic representation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1014–1023.

[12] A. Sarker, M. Belousov, J. Friedrichs, K. Hakala, S. Kiritchenko, F. Mehryary, S. Han, T. Tran, A. Rios, R. Kavuluru, et al., Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1274–1283.

[13] M. Zolnoori, K.W. Fung, T.B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola, N.D. Shah, Y.S.S. Wu, C.E. Eldredge, J. Luo, et al., The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications, Data Brief 24 (2019) 103838.

[14] Y.-F. Luo, W. Sun, A. Rumshisky, MCN: a comprehensive corpus for medical concept normalization, J. Biomed. Inform. 92 (2019) 103132.

[15] S. Scepanovic, E. Martin-Lopez, D. Quercia, K. Baykaner, Extracting medical entities from social media, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 170–181.

[16] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 3122–3137.

[17] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, C.P. Rosé, Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets, J. Biomed. Inform. 121 (2021) 103880.

[18] R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, J. Biomed. Inform. 57 (2015) 28–37.

[19] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, V. Malykh, Medical concept normalization in social media posts with recurrent neural networks, J. Biomed. Inform. 84 (2018) 93–102.

[20] Z. Miftahutdinov, E. Tutubalina, Deep neural models for medical concept normalization in user-generated texts, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 393–399.

[21] E. Tutubalina, A. Kadurin, Z. Miftahutdinov, Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6710–6716.

[22] H. Cho, D. Choi, H. Lee, Re-ranking system with BERT for biomedical concept normalization, IEEE Access 9 (2021) 121253–121262.

[23] F. Remy, K. Demuynck, T. Demeester, BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights, J. Am. Med. Inform. Assoc. (2024) ocae029.

[24] U. Naseem, J. Kim, M. Khush, A.G. Dunn, A linguistic grounding-infused contrastive learning approach for health mention classification on social media, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 529–537.

[25] N. Pattisapu, S. Patil, G. Palshikar, V. Varma, Medical concept normalization by encoding target knowledge, in: Machine Learning for Health Workshop, PMLR, 2020, pp. 246–259.

[26] H. Chen, J. Chen, J. Ding, Data evaluation and enhancement for quality improvement of machine learning, IEEE Trans. Reliab. 70 (2) (2021) 831–847.

[27] S. Ji, Y. Gao, P. Marttinen, Knowledge-augmented graph neural networks with concept-aware attention for adverse drug event detection, 2024, arXiv preprint arXiv:2301.10451.

[28] A. Wang, C. Liu, J. Yang, C. Weng, Fine-tuning large language models for rare disease concept normalization, J. Am. Med. Inform. Assoc. (2024) ocae133.

[29] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, V. Munigala, Overview and importance of data quality for machine learning tasks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3561–3562.

[30] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance, 2022, arXiv preprint arXiv:2207.14529.

[31] Q. Tian, M. Liu, L. Min, J. An, X. Lu, H. Duan, An automated data verification approach for improving data quality in a clinical registry, Comput. Methods Programs Biomed. 181 (2019) 104840.

[32] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, L.M. Aroyo, "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–15.

[33] A. Ferré, P. Langlais, An analysis of entity normalization evaluation biases in specialized domains, BMC Bioinformatics 24 (1) (2023) 227.

[34] A. Ng, A chat with andrew on mlops: From model-centric to data-centric AI, 2021, URL https://www.youtube.com/watch?v=06-AZXmwHjo. [Online; (Accessed 14 April 2024)].

[35] D. Zha, Z.P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, X. Hu, Data-centric artificial intelligence: A survey, 2023, arXiv preprint arXiv:2303.10158.

[36] J. Cui, R. Zhang, R. Li, F. Zhou, Y. Song, E. Gehringer, A comparative analysis of GitHub contributions before and after an oss based software engineering class, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, 2024, pp. 576–582.

[37] S.E. Whang, Y. Roh, H. Song, J.-G. Lee, Data collection and quality challenges in deep learning: A data-centric AI perspective, VLDB J. (4) (2023) 791–813.

[38] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. 41 (3) (2009) 1–52.

[39] H. Chen, L.F. Pieptea, J. Ding, Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches, IEEE Trans. Reliab. 71 (2) (2022) 657–673.

[40] W. Elouataoui, S. El Mendili, Y. Gahi, An automated big data quality anomaly correction framework using predictive analysis, Data (12) (2023) 182.

[41] A. Dirkson, S. Verberne, A. Sarker, W. Kraaij, Data-driven lexical normalization for medical social media, Multimodal Technol. Interact. 3 (3) (2019) 60.

[42] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238.

[43] J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, F. Li, Unified named entity recognition as word-word relation classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 10965–10973.

[44] D. Kartchner, J. Deng, S. Lohiya, T. Kopparthi, P. Bathala, D. Domingo-Fernández, C. Mitchell, A comprehensive evaluation of biomedical entity linking models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14462–14478.

[45] D. Agarwal, R. Angell, N. Monath, A. McCallum, Entity linking via explicit mention-mention coreference modeling, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022.

[46] J. Ding, X. Li, V.N. Gudivada, Augmentation and evaluation of training data for deep learning, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 2603–2611.

[47] N. Gupta, S. Mujumdar, H. Patel, S. Masuda, N. Panwar, S. Bandyopadhyay, S. Mehta, S. Guttula, S. Afzal, R. Sharma Mittal, et al., Data quality for machine learning tasks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 4040–4041.

[48] S.E. Whang, Y. Roh, H. Song, J.-G. Lee, Data collection and quality challenges in deep learning: A data-centric ai perspective, VLDB J. 32 (4) (2023) 791–813.

[49] J. Cui, R. Li, K. Lou, C. Liu, Y. Xiao, Q. Jia, E. Gehringer, R. Zhang, Can pre-class github contributions predict success by student teams? in: Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Software Engineering Education and Training, 2022, pp. 40–49.

[50] Y. Liu, Y. Wang, K. Zhou, Y. Yang, Y. Liu, Semantic-aware data quality assessment for image big data, Future Gener. Comput. Syst. 102 (2020) 53–65.

[51] N. Tran, H. Chen, J. Bhuyan, J. Ding, Data curation and quality evaluation for machine learning-based cyber intrusion detection, IEEE Access 10 (2022) 121900–121923.

[52] Y. Gong, G. Liu, Y. Xue, R. Li, L. Meng, A survey on dataset quality in machine learning, Inf. Softw. Technol. (2023) 107268.

[53] J. Yoon, S. Arik, T. Pfister, Data valuation using reinforcement learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 10842–10851.

[54] S. Mishra, A. Arunkumar, B. Sachdeva, C. Bryan, C. Baral, Dqi: Measuring data quality in nlp, 2020, arXiv preprint arXiv:2005.00816.

[55] J. Ding, X. Li, X. Kang, V.N. Gudivada, A case study of the augmentation and evaluation of training data for deep learning, J. Data Inf. Qual. (JDIQ) 11 (4) (2019) 1–22.

[56] F.H.K.D.S. Tanaka, C. Aranha, Data augmentation using GANs, 2019, arXiv preprint arXiv:1904.09135.

[57] Z. Han, W. Li, A method for improving data quality of large-scale simulation, in: 2021 14th International Symposium on Computational Intelligence and Design, ISCID, IEEE, 2021, pp. 164–168.

[58] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An empirical survey of data augmentation for limited data learning in nlp, Trans. Assoc. Comput. Linguist. 11 (2023) 191–211.

[59] J. Cui, F. Zhou, R. Zhang, R. Li, C. Liu, E. Gehringer, Predicting students' software engineering class performance with machine learning and pre-class GitHub metrics, in: 2023 IEEE Frontiers in Education Conference, FIE, IEEE, 2023, pp. 1–9.

[60] X. Han, K. Zhu, S. Liang, Z. Zheng, G. Zeng, Z. Liu, M. Sun, QASnowball: An iterative bootstrapping framework for high-quality question-answering data generation, 2023, arXiv preprint arXiv:2309.10326.

[61] L. Wu, F. Zhang, C. Cheng, S. Song, Supervised contrast learning text classification model based on data quality augmentation, ACM Trans. Asian Low- Resour. Lang. Inf. Process. (2024).

[62] Y. Zhou, C. Guo, X. Wang, Y. Chang, Y. Wu, A survey on data augmentation in large model era, 2024, arXiv preprint arXiv:2401.15422.

[63] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 462–477.

[64] A. Abaskohi, S. Rothe, Y. Yaghoobzadeh, LM-CPPF: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 670–681.

[65] A. Latif, J. Kim, Evaluation and analysis of large language models for clinical text augmentation and generation, IEEE Access (2024).

[66] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D.C. Comeau, et al., Opportunities and challenges for ChatGPT and large language models in biomedicine and health, Brief. Bioinform. 25 (1) (2024) bbad493.

[67] C. Zheng, G. Wu, C. Li, Toward understanding generative data augmentation, Adv. Neural Inf. Process. Syst. 36 (2024).

[68] Y. Shen, L. Heacock, J. Elias, K.D. Hentel, B. Reig, G. Shih, L. Moy, ChatGPT and other large language models are double-edged swords, 307, (2) 2023, e230163,

[69] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–45.

[70] N. Lee, T. Wattanawong, S. Kim, K. Mangalam, S. Shen, G. Anumanchipali, M.W. Mahoney, K. Keutzer, A. Gholami, LLM2LLM: Boosting LLMs with novel iterative data enhancement, 2024, arXiv preprint arXiv:2403.15042.

[71] J. Cui, R. Zhang, R. Li, Y. Song, F. Zhou, E. Gehringer, Correlating students' class performance based on github metrics: A statistical study, in: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, 2023, pp. 526–532.

[72] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Medica 22 (3) (2012) 276–282.

[73] Y.-C. Lin, P. Hoffmann, E. Rahm, Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models, SN Comput. Sci. 3 (5) (2022) 387.

[74] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[75] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, 2019, arXiv preprint arXiv:1904.03323.

[76] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, A. Wong, Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus, 2020, arXiv preprint arXiv:2010.10391.

[77] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Heal. (HEALTH) 3 (1) (2021) 1–23.

[78] J. Niu, Y. Yang, S. Zhang, Z. Sun, W. Zhang, Multi-task character-level attentional networks for medical concept normalization, Neural Process. Lett. 49 (2019) 1239–1256.

[79] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, 2022, arXiv preprint arXiv:2204.03905.

[80] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv preprint arXiv:1910.13461.

[81] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, S. Yu, CODER: Knowledge-infused cross-lingual medical term embedding for term normalization, J. Biomed. Inform. 126 (2022) 103983.

[82] T.M. Lai, C. Zhai, H. Ji, KEBLM: knowledge-enhanced biomedical language models, J. Biomed. Inform. 143 (2023) 104392.

[83] M. Sänger, S. Garda, X.D. Wang, L. Weber-Genzel, P. Droop, B. Fuchs, A. Akbik, U. Leser, HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools, 2024, arXiv preprint arXiv:2402.12372.

[84] H. Zhao, H. Chen, H.-J. Yoon, Enhancing text classification models with generative AI-aided data augmentation, in: 2023 IEEE International Conference on Artificial Intelligence Testing, AITest, IEEE, 2023, pp. 138–145.

[85] Q. Chen, J. Zobel, K. Verspoor, Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study, Database 2017 (2017) baw163.

[86] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, N. Carlini, Deduplicating training data makes language models better, 2021, arXiv preprint arXiv:2107.06499.

[87] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, P.J. Kennedy, Training deep neural networks on imbalanced data sets, in: 2016 International Joint Conference on Neural Networks, IJCNN, IEEE, 2016, pp. 4368–4374.

[88] Q. Wei, R.L. Dunbrack Jr., The role of balanced training and testing data sets for binary classifiers in bioinformatics, PLoS One 8 (7) (2013) e67863.

[89] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, J. Han, Tuning language models as training data generators for augmentation-enhanced few-shot learning, in: International Conference on Machine Learning, PMLR, 2023, pp. 24457–24477.

[90] A. Chowdhury, J. Alspector, Data duplication: an imbalance problem, in: ICML'2003 Workshop on Learning from Imbalanced Data Sets (II), Washington, DC, 2003.

[91] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, The curse of recursion: Training on generated data makes models forget, 2024, arXiv Preprint arXiv:2305.17493v3.

[92] A.J. Peterson, AI and the problem of knowledge collapse, 2024, arXiv preprint arXiv:2404.03502.

[93] I.O. Gallegos, R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed, Bias and fairness in large language models: A survey, Comput. Linguist. (2024) 1–79.

[94] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A.J. Ratner, R. Krishna, J. Shen, C. Zhang, Large language model as attributed training data generator: A tale of diversity and bias, Adv. Neural Inf. Process. Syst. 36 (2024).

[95] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, et al., Prompt engineering for healthcare: Methodologies and applications, 2023, arXiv preprint arXiv:2304.14670.