The Use of SNOMED CT in Large Language Models: A Scoping Review

Eunsuk Chang, Sumi Sung

Submitted to: Journal of Medical Internet Research
on: June 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

The Use of SNOMED CT in Large Language Models: A Scoping Review

Eunsuk Chang[1] PhD; Sumi Sung[2] PhD

[1]Republic of Korea Air Force Aerospace Medical Center Cheongju, Chungcheongbuk-do KR
[2]Department of Nursing Science, Research Institute of Nursing Science Chungbuk National University Cheongju, Chungcheongbuk-do KR

Corresponding Author:
Sumi Sung PhD
Department of Nursing Science, Research Institute of Nursing Science
Chungbuk National University
1 Chungdae-ro Seowon-gu
Cheongju, Chungcheongbuk-do
KR

Abstract

Background: SNOMED CT serves as a widely adopted standardized terminology in electronic health records and common data models, garnering attention for its secondary applications as a biomedical knowledge source. While large language models commonly face "hallucination" challenges, integrating SNOMED CT as a knowledge base with LLMs has been proposed to improve natural language understanding and generation in the biomedical domain.

Objective: We aimed to review the state-of-the-art methodologies for incorporating SNOMED CT into LLMs to enhance biomedical natural language understanding and generation tasks.

Methods: A comprehensive review of SNOMED CT integration in language models was conducted by querying ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase for publications between 2018 and 2023. Thirty-seven papers were selected for the final review.

Results: BERT and its fine-tuning variants were the mainstream baseline language models in the examined literature. The majority of studies (n=28) incorporated SNOMED CT contents, such as descriptions, relations, and entity types (classes), into the inputs of large language models or training corpora. Other approaches included incorporating SNOMED CT into additional fusion modules of language models or retrieving knowledge from SNOMED CT for inference. SNOMED CT-integrated large language models prevailed in natural language understanding tasks (n=30) such as entity typing, classification, and, most notably, medical concept normalization. The integrated models also encompassed natural language generation tasks (n=9), such as translation, summarization, and question answering. However, only a small number of studies reported performance differences before and after the SNOMED CT integration.

Conclusions: As the utilization of SNOMED CT as a reliable knowledge source becomes more feasible, SNOMED CT-integrated language models hold the potential to warrant model accountability, demonstrating advancements in the tasks of comprehending and generating NL for downstream tasks in the biomedical realm. Future research is anticipated to be more cognizant of the advantage of incorporating SNOMED CT into large language models.

(JMIR Preprints 04/06/2024:62924)
DOI: https://doi.org/10.2196/preprints.62924

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

The Use of SNOMED CT in Large Language Models: A Scoping Review

Abstract

Background: SNOMED CT serves as a widely adopted standardized terminology in electronic health records and common data models, garnering attention for its secondary applications as a biomedical knowledge source. While large language models commonly face "hallucination" challenges, integrating SNOMED CT as a knowledge base with LLMs has been proposed to improve natural language understanding and generation in the biomedical domain.

Objective: We aimed to review the state-of-the-art methodologies for incorporating SNOMED CT into LLMs to enhance biomedical natural language understanding and generation tasks.

Methods: A comprehensive review of SNOMED CT integration in language models was conducted by querying ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase for publications between 2018 and 2023. Thirty-seven papers were selected for the final review.

Results: BERT and its fine-tuning variants were the mainstream baseline language models in the examined literature. The majority of studies (*n*=28) incorporated SNOMED CT contents, such as descriptions, relations, and entity types (classes), into the inputs of large language models or training corpora. Other approaches included incorporating SNOMED CT into additional fusion modules of language models or retrieving knowledge from SNOMED CT for inference. SNOMED CT-integrated large language models prevailed in natural language understanding tasks (*n*=30) such as entity typing, classification, and, most notably, medical concept normalization. The integrated models also encompassed natural language generation tasks (*n*=9), such as translation, summarization, and question answering. However, only a small number of studies reported performance differences before and after the SNOMED CT integration.

Conclusions: As the utilization of SNOMED CT as a reliable knowledge source becomes more feasible, SNOMED CT-integrated language models hold the potential to warrant model accountability, demonstrating advancements in the tasks of comprehending and generating NL for downstream tasks in the biomedical realm. Future research is anticipated to be more cognizant of the advantage of incorporating SNOMED CT into large language models.

Keywords: SNOMED CT; ontology; knowledge graph; large language models; natural language processing; language models

Introduction

The recent emergence of large language models (LLMs), exemplified by Bidirectional Encoder Representations from Transformers (BERT) [1] and Generative Pre-trained Transfomer (GPT) [2], has significantly advanced the capabilities of machine performance in natural language (NL) understanding and generation. Despite achieving state-of-the-art performance on a range of natural language processing (NLP) tasks, LLMs exhibit a deficiency in knowledge when confronted with knowledge-driven tasks [3]. These models acquire factual information from extensive text corpora during training, embedding this knowledge implicitly within their numerous parameters, posing challenges in terms of verification and manipulation [4]. Moreover, numerous studies have demonstrated that LLMs struggle to recall facts and frequently encounter hallucinations, generating statements that are factually inaccurate [5,6]. This poses a significant obstacle to the effective application of LLMs in critical scenarios, such as medical diagnosis and legal judgment [7].

Efforts have been made to address the “black box” nature of LLMs and mitigate potential “hallucination” problems. Approaches include enhancing language model (LM) veracity through strategies like retrieval chain-of-thought prompting [8] and retrieval augmented generation [9]. Another significant avenue involves integrating knowledge graphs (KGs) or ontologies into LMs, utilizing triple relations or KG subgraphs [7,10]. KGs, renowned for their excellence in representing knowledge within a domain, can provide answers when combined with LMs [11], making them valuable for common-sense-based reasoning and fact-checking models [12]. However, LLMs often face challenges when trained and tested predominantly on general domain datasets or KGs, such as Wikipedia and WordNet [13], making it difficult to gauge their performance on datasets containing biomedical texts. The differing word distributions in general and biomedical corpora pose challenges for biomedical text mining models [14].

Biomedicine-specific KGs may be a potential solution to these problems. In the biomedical domain, KGs, also known as ontologies, are relatively abundant, with the Unified Medical Language System (UMLS) [15] being one of the most frequently utilized ontologies [16]. UMLS serves as a thesaurus for biomedical terminology systems such as the Medical Subject Headings (MeSH), International Classification of Diseases (ICD), Gene Ontology, Human Phenotype Ontology (HPO), and SNOMED CT (SCT), curated and managed by the US National Library of Medicine.

Among the UMLS member terminologies, SCT stands out as the most comprehensive biomedical ontology, encompassing a wide range of biomedical and clinical entities, including signs, symptoms, diseases, procedures, and social contexts [17]. These entities are represented by concepts (clinical ideas), descriptions (human-readable terms linked to concepts), and relations (comprising hierarchical *is-a* relations and horizontal attribute relations). As SCT is increasingly integrated into electronic health record (EHR) systems, as required by the Fast Healthcare Interoperability Resource (FHIR) to ensure interoperability among healthcare institutions [18], terminology servers supporting SCT have become ubiquitous. With its ready availability across healthcare institutions, SCT has gained attention with roles as a knowledge source or ontology for representing biomedical and clinical knowledge [17]. In this case, the abstract model of SCT is employed to describe and store biomedical facts in a hierarchical and structured manner, readily available across healthcare institutions.

This scoping review examines the use of SCT as a knowledge source to be incorporated with NLP tasks. The growing interest in both SCT and LLMs prompted us to investigate in detail how these two modalities are integrated to enhance NL understanding (NLU) and generation (NLG).

Methods

This scoping review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-analyses for Scoping Reviews (PRISMA-ScR) framework, which outlines the recommended steps and reporting standards for conducting scoping reviews [19].

Study Identification

To explore scientific literature that describes these models, we conducted our search on ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase on March 12, 2024, using the following query terms: (1) ("language *model" OR "pre-trained *model" OR "language processing" OR "embedding") AND ("SNOMED" OR "Unified Medical Language System" OR "UMLS" OR "*medical") AND ("knowledge graph" OR "ontolog*" OR "knowledge*base" OR "knowledge infusion") and (2) ("SNOMED” AND (“large language model” OR “BERT” OR “GPT”) (query may be modified according to bibliographic databases). Queries were designed to search articles that were published between 2018 and 2023.

Study Selection

The query retrieved a total of 876 articles from the five bibliographic databases, with 634 from the first query and 242 from the second (Figure 1). After deduplication, we then examined the full text of the retrieved articles for the presence of the term “SNOMED.” We prioritized a full-text search first before title/abstract review because many potentially eligible papers do not explicitly mention SNOMED in their titles or abstracts. Using this approach, we identified 325 papers that referenced “SNOMED” in full-text bodies.

To be eligible for this review, SCT had to be incorporated into NLP pipelines, which encompass processes from text cleansing through pre-training and inference to model evaluation, specifically for tasks involving NLU and NLG. We then further excluded studies that met one or more of the following criteria: (1) published in languages other than English; (2) categorized as reviews, surveys, keynotes, or editorial articles; (3) did not incorporate SCT at any stage of NLP pipeline; (4) aimed to create, develop, enrich, or enhance ontologies or graphs; or (5) did not involve the processing of *natural* language text. Studies that solely employed SCT codes for retrieving patients of interest from EHRs or for annotating instances with SCT codes as gold-standard target labels for LM training were omitted.

Synthesis of Results

We defined LLMs as transformer-based language models pre-trained on large-scale corpora [20]. Through discussions and qualitative assessment by the reviewers, we analyzed the included articles according to the following characteristics: chronological and geographic publication trends; baseline LLM and its output; dataset used for training and testing the model; methods in integrating SCT into the LLM; and the model’s end-task and performance.

We elucidated the methodology for incorporating SCT into NLP pipelines in accordance with the categorization methods previously outlined by Pan et al. [7]. These methods categorized methodologies for KG-enhanced LLMs into three distinctive types: (1) KG-enhanced LLM pre-training, (2) KG-enhanced LLM interpretability, and (3) KG-enhanced LLM inference. The end tasks of LLMs after SCT integration include NLU and NLG. NLU involves entity recognition or typing, entity or relation extraction, document classification, question answering (multiple choice), and inference. NLG involves text summarization, question answering (short or essay answers), translation, and dialogue generation. Regarding the performance analysis, we present the nominal percentage gains in performance after SCT integration without analyzing their statistical significance, as the majority of studies did not perform statistical significant testing. We refrained from direct study-to-study comparisons due to concerns about the heterogeneity of testing corpora and evaluation metrics across different studies.

Results

A total of 37 publications were selected for the final scoping review (Figure 1). The detailed descriptions of the characteristics of individual papers and other features, including the language of utilized datasets and SCT descriptions, other ontologies used, and the kinds of entities represented by SCT, are outlined in the supplementary material (Multimedia Appendix 1).

Chronological and Geographic Publication Trends

Table 1 presents the publication trends of the review. Although our literature search covered publications from 2018 onwards, no studies published in 2018 were included in the final review. The largest volume of studies was published in 2022 (*n*=13), followed by those published in 2020 (*n*=10)

When counting countries where the first authors' affiliated institutions are located, the largest number of studies originated from the United States (*n*=10). While the majority of studies were conducted in countries that are members of SNOMED International, some came from non-member countries such as Bulgaria and China, where separate license fees and in-house translation of SCT descriptions to the local language were required.

Table 1. Chronological and geographic publication trends of the included studies.

Study Characteristics	Studies
Publication year	
2019	[21], [22], [23]
2020	[24], [25], [26], [27], [28], [29], [30], [31], [32], [33]
2021	[34], [35], [36]
2022	[37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49]
2023	[50], [51], [52], [53], [54], [55], [56], [57]
Countries	
Australia	[26], [35]
Bulgaria	[34], [52]
Canada	[55]
China (incl. Hong Kong)	[28], [38], [39], [41], [43], [45], [48], [50], [56]
Germany	[47], [51]
India	[22], [31], [32]
Israel	[53]
Spain	[21], [29], [37], [30], [40]
United Kingdom	[54], [57]
United States	[23], [24], [25], [27], [33], [36], [42], [44], [46], [49]
Type of publication	
Journal papers	[23], [24], [25], [26], [36], [42], [43], [44], [45], [46], [50], [55], [56], [57]

Conference papers	[21], [22], [27], [28], [29], [30], [31], [32], [33], [34], [35], [37], [38], [39], [40], [41], [45], [47], [48], [49], [51], [52], [53], [54]

Baseline Large Language Models and Their Outputs

The majority of the included studies employed BERT and its variants as the baseline LLMs for the tasks of NLU and NLG tasks. Variants such as RoBERTa [58] and ALBERT [59] were also utilized to address BERT's relatively small training corpora and long training time [31,37,38,50,53]. To overcome the limited applicability of these general-purpose LLMs to biomedical texts, many studies (*n*=13) employed LLMs trained on large-scale biomedical corpora, such as BioBERT [14] and PubMedBERT [60], which were trained on PubMed articles, and ClinicalBERT [61] and EHRBERT [23], trained on clinical notes. SapBERT [62], initialized by PubMedBERT, was further fined-tuned using contrastive learning with UMLS synonyms to better accommodate SCT synonym descriptions [44,47]. To support biomedical NLP tasks in languages other than English, LLMs trained on corpora in those languages were also adopted, such as medBERT.de [63], designed specifically for the German medical domain [51], and ERNIE-health, pre-trained from Chinese medical records [41]. Departing from the BERT-based models, GPT emerged as a new baseline LLM since 2023. Makhervaks et al. [53] employed BioGPT [64], whose decoder was pre-trained on biomedical corpora, to enhance the generation of artificial sentences. Additionally, Xu et al. [55] utilized GPT-3.5 for ranking suggested annotation terms in their study (Table 2).

Table 2. Large language models used in the included studies.

Base and fine-tuned model	Studies
BERT[a]	
Vanilla BERT	[22], [24], [26], [27], [33], [40], [42], [43], [44], [50], [53], [54], [56], [57]
RoBERTa	[31], [37], [38], [50]
ALBERT	[53]
ELECTRA	[53]
DeBERTa	[53]
mBERT	[37], [45]
BioBERT	[27], [33], [34], [46], [48], [49], [52]
ClinicalBERT	[25], [33], [35], [36]
PubMedBERT	[45], [46]
SAPBERT	[44], [47]
EHRBERT	[23]
SciBERT	[46]

BioELECTRA	[53]
German BERT models	[51]
GPT[b]	
GPT-3.5	[55]
BioGPT	[53]
BART	[57]
Transformer neutral networks	
Transformer NMT[c] Model	[21], [28], [29], [30], [39]
Denoising autoencoder	[32]
ERNIE[d]	
ERNIE-health	[41]

[a] BERT: Bidirectional Encoder Representations from Transformers
[b] GPT: Generative Pre-trained Transformer
[c] NMT: Neural Machine Translation
[d] ERNIE: Enhanced Language Representation with Informative Entities

A primary assertive role of LLMs was representing biomedical entities from text data. While the majority of proposed methods produced embedding vectors to convey contextual information about the biomedical entities that appeared in texts, Kalyan and Sangeetha [31] introduced a Siamese RoBERTa model to generate concept vectors from synonym relationships defined by SCT. These basic outputs of LLMs might undergo additional task-specific layers to perform desired end-tasks, which will be discussed later. Beyond producing embedding representations of entities, some studies required LLMs to perform classification or ranking tasks after fine-tuning, predicting the most likely relevant standard concepts [23,24,26,34,41,55], entity types [35,38,51], sentences [49,53], or matched foreign language words, enabling machine translation [28–30,39]. LLMs with encoder-decoder architectures, such as BART [65], were employed for dedicated NLG tasks [32,57].

Data for Training and Testing Models

When utilizing general-domain LLMs, authors deployed additional fine-tuning or pre-training on biomedical corpora to better adapt their models for biomedical NLP tasks. The pre-training corpora included PubMed or Medline articles [28,30,38,39,46] and other publicly available datasets, such as Wikipedia articles [29] and tweets [37] related to biomedical topics. Synthetic sentences were also utilized to address data scarcity, generated based on SCT descriptions or relations [21,29].

While some studies utilized real-world clinical narrative records [21,30,48,52] or customized (i.e., manually annotated by researchers) data [25,27,41,56] for testing their models, the majority of the studies made use of publicly available datasets, especially when researchers were participating in shared task competitions or dealing with English texts. CADEC [66] and PsySTAR [67], open datasets built from drug review posts in which concept mentions were mapped to SCT concepts, were used for validating and testing concept normalization models [31,45]. The MCN (Medical

Concept Normalization) corpus, drawn from discharge summaries annotated using SCT and RxNorm concepts, was experimented on by concept normalization models [24,26]. The WMT corpora, provided by the annual Conference on Machine Translation shared tasks, were used to test multilingual machine translation tasks by participating researchers [28,29,39]. Makhervaks et al. [53] and Chopra et al. [22] utilized sentence pairs in the MedNLI corpus [68], which are annotated by medical doctors into three categories – contradictory, entailing, and neutral – for the NL inference task. The MedMentions corpus [69] identifies over 350,000 mentions from over 4,000 PubMed abstracts, linking them to the UMLS concepts, and was used in Zotova et al. [40] and Dong et al.'s [54] studies in which SCT was loaded onto the UMLS. The ShARe/CLEF 2013 corpus [70] consists of de-identified clinical notes that annotated disease mentions by the SCT subset of the UMLS and was used for testing concept normalization tasks [44,54].

SNOMED CT Content Integration into Natural Language Processing Pipelines

While Pan et al.'s [7] categorization methods pertain to the integration of LLMs with general-purpose KGs, we treated SCT as a specified form of KG. Their third category, "KG-enhanced LLM inference," was omitted due to the lack of relevant studies in our review. The overarching categorization of all the included methods is shown in Table 3.

Table 3. Summarized categorizations of SNOMED CT-incorporated large language model methods (allowed duplicated counting of studies)

Category	Subcategory
SNOMED CT-enhanced LLM[a] pre-training	Integrating SNOMED CT into LLM inputs (*n*=28)
	Integrating SNOMED CT into additional fusion modules (*n*=5)
SNOMED CT-enhanced LLM inference	Retrieval-augmented knowledge fusion (*n*=5)

[a]LLM: large language model

Integrating SNOMED CT into Large Language Model Inputs

Research in this area concentrated on developing new training objectives for LLMs that incorporate knowledge awareness. More specifically, this line of research aimed to incorporate relevant portions or subsets of SCT as additional input to LLMs during training. Since a disproportionately large number of included studies (*n*=28) fell into this category, we analyzed the methodology by two additional themes: the content of SCT that was integrated into an LLM, and the part of the NLP pipeline into which the aforementioned content was incorporated. After qualitative analysis of the included articles and heuristic discussions among reviewers, we categorized the former into descriptions (which may include descriptions of synonyms), relations, and entity types (classes), and the latter into encoders and training data. SCT content could be incorporated into LLM encoders as embedding vectors, or they could be incorporated as annotations or tags when they were incorporated into the training corpus.

Table 4 shows the distribution of models across SCT contents and NLP pipelines, allowing for duplicated counting of a single study if it adopted two or more methods.

Table 4. Distributions of models across SNOMED CT contents and natural language processing pipelines.

SCT[a] content integrated into NLP pipeline	Part of the NLP[b] pipeline where SCT contents were integrated into	
	Encoder (as vector embedding)	Training corpora (as annotated text)
Description	[31], [35], [41], [43], [44], [54]	[21], [23], [24], [28], [29], [30], [32], [34], [39], [40], [47], [48], [49], [50], [52], [54], [57]
Relation	[31], [45]	[21], [34], [40], [52], [53]
Entity type (class)		[25], [38], [42], [51]

[a]SCT: SNOMED CT

[b]NLP: natural language processing

Integration of SNOMED CT descriptions

Vector representations of SCT concept descriptions were created to facilitate seamless fusion into LLM encoders. The vectors for SCT description embeddings were used to calculate cosine similarity between the original mentions and SCT descriptions for concept normalization tasks [35,41,43,54].

Instead of transforming text descriptions into vector embeddings, NL description texts were directly added to training corpora to expand the size of in-domain vocabulary (Figure 2). The description texts of synonyms were concatenated in the training corpora before being input into an LLM for pre-training [24,47,49,54,57], or they substituted the original entity mentions in the text by standardized terms [32,48]. The descriptions of SCT codes were also prepended to the word sequences as classifier tokens for LLM pre-training [23]. The multilingual feature of SCT descriptions was exploited to address the limited availability of training datasets in foreign languages by adding the translated SCT descriptions to the training corpora [28–30,39,50].

Integration of SNOMED CT relations

This kind of research introduced relevant subgraph information of SCT, representing SCT relations as graph edges, into LLMs (Figure 3). Kalyan and Sangeetha [31] encoded SCT concept descriptions to generate concept embedding vectors and learn representation vectors of concept mentions in the text, further improving the representations by retrofitting the target concept vectors with SCT synonym relations. CODER [45] employed KG embedding methods such as DistMult and ANALOGY [71] to learn relational knowledge from SCT, enabling the quantification of term-relation-term similarity as well as term-term similarity.

A different approach was taken to introduce textual relation triplets defined by SCT to expand the size of training corpora. Soto et al. [21] exploited the relations defined in SCT, such as *is_a* and *occurs_in*, to generate synthetic training corpora. Relations defined in SCT were also used to weakly supervised sentence pairs extracted from PubMed to establish contradiction labels on dataset [53]. Other authors exploited the existing mappings to other ontologies (e.g., ICD-10, UMLS, etc) to enrich the training corpus with the description texts from the linked ontology concepts [34,40,52].

Integration of SNOMED CT entity types

The type information of entities was incorporated into training corpora by distantly labeling the identified entities with SCT semantic tags (e.g., diseases, chemicals, etc.) [25,38] (Figure 4). In other studies, training corpora were annotated with SCT top-level hierarchies [51] or subclasses of the top-level hierarchies [42] to label sentences in accordance with their respective tasks.

Integrating SCT into additional fusion modules

In this approach, no contents of SCT are involved in the pre-training process of LLMs. Instead, concept information was processed separately before being concatenated and fused with LLM embedding output (Figure 5). Authors created knowledge-directed embeddings using the SCT graph, where concepts were represented as nodes and relations as edges, and concatenated them with the LLM contextual embeddings. The merged representations of text and graph embeddings were then passed through a task-specific knowledge fusion module to achieve end-tasks such as semantic similarity measurement [36,46], classification [22,27], and question answering [33,46]. To represent the SCT concepts' graph information, Chang et al. [36] utilized a Graph Convolutional Network [72] for encoding node features and edges. Chopra et al. [22] proposed the Bio-MTDDN model which introduced the shortest path information between corresponding SNOMED CT concepts into knowledge-directed embeddings.

Retrieval-augmented knowledge fusion

In this category of method, SCT is located outside the LLM as a fact-consulting knowledge base, injecting knowledge during the inference phase (Figure 6). In this approach, the module functions as a gazetteer (dictionary), matching mentions in texts against the dictionary of SCT descriptions to filter out irrelevant entities from the models and map textual mentions to the most likely SCT concepts [24,26,37,55,56]. These methods primarily concentrate on entity recognition and question answering, capturing both textual semantic meanings and up-to-date real-world knowledge.

End-Task and Performance Gain After SCT Integration

The majority of the included studies (*n*=30) focused on NLU tasks, such as entity typing and classification. NLG tasks, including translation and summarization, were also attempted by a substantial number of studies (*n*=9), often involving various NLU pipelines before producing the final text output. It should be noted, therefore, that works on NLU may also appear in the NLG category. We also present a comparison of the performance of models with SCT to their counterparts without SCT integration.

Natural Language Understanding

Entity extraction and/or typing

Entity typing or named entity recognition (NER) tasks aim to detect specific types of entities by identifying the spans of their mentions in the text. These can be regarded as multi-classification tasks, where the number of classes is arbitrarily chosen by researchers. To fine-tune LLMs for type classification, entities in texts were annotated by matching domain gazetteer strings (e.g., "BIO" tagging scheme) [37,38,49] or using off-the-shelf automatic concept extractors [27]. The identified entities were then classified into human-annotated entity types [37,38] or topmost nodes in the SCT hierarchies [27,51]. In addition to typing individual entities, the extraction and typing of relations between two entities were also attempted to align the detected entities with FHIR resources [25], protein to chemical and gene to disease [46], and disease to inflicted family members [35].

Many researchers did not provide a comparative performance analysis of their SCT-integrated models against out-of-domain vanilla model. Among the few that reported such comparisons, Jha and Zhang [46] demonstrated a gain in F1 score after the integration of SCT, while Montañés-Salas et al. [37] found a positive impact only on recall. (Table 5)

Table 5. Percentage performance gain in biomedical entity typing tasks after SNOMED CT integration into large language models.

Studies	% F1 gain	% Precision gain	% Recall gain	% AUC[a] gain
Montañés-Salas et al. (2022) [37] (Best 2 model)	-0.11%	-7.97%	+8.60%	N/A[b]
Jha and Zhang (2022) [46] (PubMedBERT on BC2GM)	+4.08%	N/A	-N/A	N/A

[a]AUC: area under the receiver operating characteristic curve
[b]N/A: Not Available

Classification

In this review, we define the classification task as occurring at the sentence or document level, rather than at the word, entity, or phrase level. When implementing classification tasks, semantic similarity [36] or the conditional probability of a positive case given conditions [22,33,53] was calculated, and the case was categorized as positive if the probability exceeded a threshold. The binary classification was performed to determine whether a sentence pair was entailed [33], contradictory [22,53], or similar [36]. The multi-label classification was performed to categorize utterances by clinical encounter components, such as symptoms, complaints, and medications [27], social determinants of health [42], or the narrator's intent [48]

Table 6 shows the percentage performance gain after SCT integration in classification tasks. While Yadav et al. [33] and Zhang et al. [48] estimated the performance of their models in terms of F1 score, precision, and recall, Khosla et al. [27] and Makhervaks et al. [53] measured performance in terms of area under the curve (AUC), which improved by 0.87-14.83% after the integration of SCT. Chang et al. [36] reported the Pearson correlation to assess clinical semantic textual similarity, and the incorporation of SCT into ClinicalBERT improved the model by 1.77% and 2.36% using cui2vec [73] and knowledge graph embeddings, respectively.

Table 6. Percentage performance gain in classification tasks after SNOMED CT integration into large language models.

Studies	% F1 gain	% Precision gain	% Recall gain	% AUC[a] gain	% Accuracy gain
Chopra et al. (2019) [22]	N/A[b]	N/A	N/A	N/A	+0.99%
Yadav et al. (2020) [33]	+26.05%	+36.87%	+16.41%	N/A	+17.28%
Khosla et al. (2020) [27]	N/A	N/A	N/A	+0.87%	N/A
Zhang et al. (2022) [48]					
bioBERT for intent detection	+1.15%	N/A	N/A	N/A	N/A

Semantic matching for content recognition	N/A	-0.90%	+12.15%	N/A	N/A
Makhervaks et al. (2023) [53]					
BERT Base on MedNLI-General	N/A	N/A	N/A	+14.83%	N/A
bio-GPT on MedNLI-General	N/A	N/A	N/A	+10.34%	N/A

[a]AUC: area under the receiver operating characteristic curve
[b]N/A: Not Available

Medical concept normalization (MCN)

The most prominent end-task in NLU was MCN, with a total of 15 studies involved. MCN, the task of linking textual mentions to concepts in an ontology, provides a solution to unify different ways of referring to the same concept. The majority of studies approached concept recognition as a multi-label classification task involving entity extraction and entity typing from words, phrases, or sentences. Models were trained on corpora annotated with SCT concepts and semantic types to identify concept mentions and generate a list of candidate SCT concepts that best match those mentions from testing texts. When training from annotated corpora was not available, MetaMap [74] was utilized to extract biomedical entities mentioned in free texts and map them to ontology concepts [25,26,35,50]. When ranking candidate concepts, representation vectors of mentions and concept descriptions were generated, and their similarity was calculated using cosine similarity [31,35,44,45,54], linear transformation such as support vector classifiers [52], or softmax function [23,41,43]. In a more rule-oriented approach, Borchert and Schapranow [47] calculated weights based on semantic type and preferred term status from a gazetteer to reorder candidate lists. In other studies [24,26,50], sieve-based multi-pass entity linking systems [75] were employed to rank the most likely concepts and achieved superior performance compared to neural classifiers.

Most of the studies observed positive gains in accuracy after SCT integration for the MCN task (Table 7). Two authors reported pre- and post-integration F1, recall, and precision, and observed inconsistent results, with one reporting positive gains in F1 score and precision, while the other demonstrated a loss in F1 score and precision after the integration of SCT.

Table 7. Percentage performance gain in medical concept normalization task after SNOMED CT integration into large language models.

Studies	% F1 gain	% Precision gain	% Recall gain	% Accuracy gain
Peterson et al. (2020) [25]	-1.05%	-1.04%	0.00%	N/A[a]
Wang et al. (2020) [26] (vs training data dictionary with exact match, ignore order "yes")[b]	N/A	N/A	N/A	+27.36%
Hristov et al. (2021) [34]	N/A	N/A	N/A	+73.21%
Dai et al. (2021) [35]	N/A	N/A	N/A	+45.08%

Xu and Miller. (2022) [44] (on ShARe/CLEF 2013)	N/A	N/A	N/A	+0.68%
Dong et al. (2023) [54] (BLINKout on ShARe/CLEF 2013)	+5.87	+15.11%	-3.62%	+10.68%

[a]N/A: Not Available

[b]The training data dictionary was constructed based on the MCN corpus data. The SNOMED CT dictionary includes RxNorm dictionary.

Natural Language Generation

Machine Translation

Several studies that participated in the WMT Biomedical Shared Task [76] described their methods for translating biomedical texts from various foreign languages, such as Spanish, French, German, and Chinese, as well as less-resourced languages, such as Basque, into English or vice versa. Transformer-based multi-lingual neural machine translation systems were the mainstream architecture, trained on dictionaries derived from SCT [28,30,39] or clinical notes artificially generated from SCT terminology contents [21,29].

The translation performance was reported using BLEU scores [77]. While most studies presented improved BLEU scores, by up to 131.66% [21] compared to their out-of-domain models, some studies reported non-superior results [30] (Table 8).

Table 8. Performance comparison of biomedical translation tasks with and without SNOMED CT integration into large language models.

Studies	Translation direction	Performance on test data *without* SCT[a] integration in LLM[b] (BLEU[c] score)	Performance on test data *with* SCT integration in LLM (BLEU score)	% BLEU score gain after SCT integration in LLM
Soto et al. (2019) [21]	EU[d] → ES[e]	10.55	24.44	+131.66%
Soto et al. (2020) [30]	ES → EN[f]	57.25	56.89	-0.63%
	EN → ES	47.19	47.15	-0.08%
Corral and Saralegi (2020) [29]	EN → EU	12.85	13.61	+5.91%
Peng et al. (2020) [28]	EN → FR[g]	38.98	41.66	+6.88%
	FR → EN	38.31	38.44	+0.34%
Wang et al. (2022) [39]	EN → IT[h]	33.53	42.17	+25.77%
	IT → EN	36.43	43.72	+20.01%
	EN → PT[i]	38.73	50.12	+29.41%
	PT → EN	41.84	54.74	+30.83%

	EN → RU[j]	25.25	36.25	+43.56%
	RU → EN	39.76	47.09	+18.44%

[a]SCT: SNOMED CT
[b]LLM: large language model
[c]BLEU: Bilingual Evaluation Understudy Score
[d]EU: Basque
[e]ES: Spanish
[f]EN: English
[g]FR: French
[h]IT: Italian
[i]PT: Portuguese
[j]RU: Russian

Text Summarization

For medical text summarization, encoder-decoder LLMs were utilized to process input embeddings and produce simplified texts. Pattisapu et al. [32] primarily focused on the simplification of verbose sentences. They substituted biomedical mentions with UMLS preferred names and tokenized them at the subword level to produce noisy input sentences for training. In contrast, Searle et al. [57] summarized entire hospital encounters into a few sentences by ranking the most salient ones to constitute the summary. To address the hallucination problem arising from LLMs, SCT semantic tags of the extracted biomedical terms were used to configure guidance signals for clinical problems and interventions.

ROUGE-recall [78] measures how many n-grams in the source text appear in the summarization. Pattisapu et al. [32] reported no gain in ROUGE-recall when incorporating SCT into NLP pipelines. Searle et al. presented ROUGE-F1, a harmonized measure of the recall and precision for ROUGE, and observed improvements by 3.6% (from 11.1 to 11.5) and 48.84% (from 8.6 to 12.8) on the MIMIC-III and King’s College Hospital corpora, respectively, after incorporating the SCT.

Question answering and generation

Generating answers to short-answer or essay questions, other than multiple-choice questions, could be categorized as NLG. The task of question-answering may involve preliminary NLU pipelines, such as intent and content recognition. Zhang et al. [48] developed a clinical communication training dialogue system incorporated with SCT synonyms for the augmentation of textual data and bioBERT for intent recognition. They qualitatively evaluated the performance of the conversation system using scales rated by physician users from 29 training records, which indicated a comparable precision as clinical experts.

Discussion

Principal Findings

In this scoping review, we observed that BERT was the mainstream LLM integrated with SCT. Considering the significant time required to publish state-of-the-art methodologies, especially in peer-reviewed journals [79], it is unsurprising that more recent inventions such as GPT-3.5 and BART were less prevalent in articles published between 2018 and 2023. Researchers in this field exploited biomedically oriented BERT variants, such as bioBERT and PubMedBERT, reflecting the need for biomedical tasks to be trained or fine-tuned on specialized corpora [16]. However, due to privacy and confidentiality concerns, there is a dearth of clinical documents and patient notes, making it difficult to sufficiently train biomedical LLMs to an extent comparable to those in the general domain [80]. SCT can supplement or even substitute biomedical pre-training corpora, addressing the chronic shortage, as noted in this review. A substantial number of included studies utilized SCT to expand pre-training corpora by concatenating synonyms or relations in documents or generating synthetic texts based on SCT descriptions or relations.

We identified three approaches to incorporating SCT into LLMs: LLM input, additional fusion modules, and knowledge retriever, with the former two intervening in the pre-training process of LLMs. While either lexical or graph information from SCT could be incorporated into the pre-training stage, the lexicon of SCT descriptions was the predominant form of integration. This underscores that SCT chiefly introduces synonym information to LLMs, yet relation information remains underutilized in NLP research. The advantage of SCT in defining relations between biomedical entities needs to be highlighted within the biomedical NLP research community, and more sophisticated methodologies to incorporate the directed acyclic graph structure of SCT into LLMs need to be developed.

A significant number of studies included in this review engaged in the concept recognition process from free text, whether as the final task or an intermediate step for subsequent tasks. Recognizing and extracting SCT concepts from the unstructured sections of EHRs is becoming crucial in clinical settings, where substantial patient information such as social history and socioeconomic status remains untapped in free-text clinical notes [81]. Leveraging previously unrepresented SCT concepts from free-text clinical data holds great potential to significantly enhance clinical care and research, especially in the era of smart applications where patient-generated data can be integrated into EHRs through the representation of patient-authored texts with SCT concepts [82].

Only a small fraction of the included models disclosed performance comparisons before and after SCT integration. For example, only 6 out of 15 studies on the MCN task provided information about the gain in F1 scores or accuracy after SCT incorporation. This suggests that many biomedical NLP researchers do not give intense attention to the role of SCT or other ontologies in improving their models. The knowledge-intensive approaches to enhancing LMs, which are often renounced by those favoring deep learning-based approaches, still comprise a small portion of the artificial intelligence research community. However, In the face of immense computational power and the availability of data required by the large language models and deep learning-based systems, an increasing number of researchers now advocate the harmonization of the two approaches [83], and a plethora of KG-enhanced LLMs is developed in the general domain [10,84]. In addition to improving the performance of artificial intelligence models, ontologies and human-curated knowledge bases can address the explainability and controllability of artificial intelligence, probing facts within the human-interpretable form of system architectures [85]. Exploring the trade-offs in combining the two approaches will bring about the next big leap toward trustworthy and reliable artificial intelligence.

Our primary focus in this review was SCT among various biomedical terminology systems and ontologies as a KG integrated with LLMs. Although the UMLS continues to dominate NLP research in the biomedical domain [16], SCT has the potential to expand its influence, given its governance over the healthcare industry. Consequently, the use of SCT as a reliable knowledge source becomes more feasible, considering its presence in various EHR systems or common data models. While this review did not identify real-world SCT-incorporated LLM applications directly tied to EHR systems, SCT is implicitly expected to support these systems as a standardized terminology system bound to syntactic interoperability structures such as FHIR and OpenEHR. Explicit descriptions of SCT in technical specifications or scientific papers by developers of these applications would have been valuable to include in this review.

Limitations

One of the limitations of the current scoping review is that we examined LLMs that accepted SCT only as a working ontology, leaving other biomedical ontologies out of our scope. To the best of our knowledge, however, there is no comprehensive review of the use of other biomedical ontologies within LLMs. Our queries used in this review, especially the first one, retrieved articles that used a variety of biomedical ontologies, such as the UMLS, Medical Subject Headings (MeSH), Gene Ontology (GO), and medical Wikidata. We chose to limit the scope of our review to SCT due to the heterogeneity of components among different ontology systems and the difficulty in delineating the contributions of each ontology in a standardized way. A more consolidated analysis of different ontologies in use within LLMs awaits more comprehensive work.

Another limitation of the current review is that we could not make a conclusive remark on how the integration of SCT improved the performance of LLMs. While the majority of studies observed a positive impact on the performance after the SCT integration, their statistical significance was not provided. Moreover, the heterogeneity of evaluation methods hindered us from conducting a meta-analysis across all the included studies. An evenhanded testing bed, such as a shared task competition under a single testing method requiring all participants to report performance differences before and after KG integration, could provide a controlled evaluation to measure the contributions of KGs in a reliable and objective way.

Conclusion

In conclusion, this scoping review explored the methodologies and performance of integrating SCT into LLMs. The predominant approach involved utilizing SCT concept descriptions or graph embeddings as inputs for LM encoders, many of which were involved in the MCN task. The endeavor to identify and extract SCT concepts from free texts demonstrated instrumental in enhancing the understanding and generation of NL text for downstream tasks in the biomedical realm. Future research is anticipated to be more aware of the advantage of SCT when incorporating it into LLMs.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Republic of Korea government (Ministry of Science and ICT) (RS-2024-00354718).

Conflicts of Interest

None declared.

Abbreviations

AUC: Area Under the Curve
BERT: Bidirectional Encoder Representations from Transformers
EHR: Electronic Health Record
FHIR: Fast Healthcare Interoperability Resource
GPT: Generative Pre-trained Transformer
HPO: Human Phenotype Ontology
ICD: International Classification of Diseases
KG: Knowledge Graph
LLM: Large Language Model
LM: Language Model
MCN: Medical Concept Normalization

MeSH: Medical Subject Headings
NER: Named Entity Recognition
NL: Natural Language
NLG: Natural Language Generation
NLP: Natural Language Processing
NLU: Natural Language Understanding
SCT: SNOMED CT
UMLS: Unified Medical Language System

Multimedia Appendix 1

Summary of included studies.

References

1. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019. 2019; 4171–4186.
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33: 1877–1901.
3. Chen Q, Li FL, Xu G, Yan M, Zhang J, Zhang Y. DictBERT: Dictionary description knowledge enhanced language model pre-training via contrastive learning. International Joint Conference on Artificial Intelligence; 2022. pp. 4086–4092.
4. Hou Y, Jiao W, Liu M, Allen C, Tu Z, Sachan M. Adapters for enhanced modeling of multilingual knowledge and text. Association for Computational Linguistics; 2022. pp. 3931–3946.
5. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2022;55: 1–38. doi:10.1145/3571730
6. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv. 2023. doi:10.48550/arxiv.2311.05232
7. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering. 2024; 1–20.
8. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022; 1–14.
9. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks . Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc.; 2020. pp. 9459–9474.
10. Hu L, Liu Z, Zhao Z, Hou L, Nie L, Li J. A survey of knowledge enhanced pre-trained language models. IEEE Trans Knowl Data Eng. 2023; 1–19. doi:10.1109/TKDE.2023.3310002
11. Lawrence P. Knowledge Graphs + Large Language Models = The ability for users to ask their own questions? . In: Medium [Internet]. 2023 [cited 30 Dec 2023]. Available: https://medium.com/@peter.lawrence_47665/knowledge-graphs-large-language-models-the-ability-for-users-to-ask-their-own-questions-e4afc348fa72

12. Anand V, Ramesh R, Jin B, Wang Z, Lei X, Lin C-Y. Multimodal language modelling on knowledge graphs for deep video understanding. Proceedings of the 29th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2021. pp. 4868–4872. doi:10.1145/3474085.3479220

13. Fellbaum C, editor. WordNet: An Electronic Lexical Database. MIT press; 1998.

14. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–1240. doi:10.1093/bioinformatics/btz682

15. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004;32: D267-70. doi:10.1093/nar/gkh061

16. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained language models in biomedical domain: A systematic survey. ACM Comput Surv. 2024;56: 1–52. doi:10.1145/3611651

17. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: A literature review. J Am Med Inform Assoc. 2021;28: 2017–2026. doi:10.1093/jamia/ocab084

18. Posnack S, Barker W. The Heat is On: US Caught FHIR in 2019. In: Health IT Buzz [Internet]. 2021 [cited 30 Dec 2023]. Available: https://www.healthit.gov/buzz-blog/health-it/the-heat-is-on-us-caught-fhir-in-2019

19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169: 467–473. doi:10.7326/M18-0850

20. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Comput Surv. 2023. doi:10.1145/3605943

21. Soto X, Perez-De-Vinaspre O, Oronoz M, Labaka G. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation. Dublin, Ireland: European Association for Machine Translation; 2019. pp. 8–18.

22. Chopra S, Gupta A, Kaushik A. MSIT_SRIB at MEDIQA 2019: Knowledge directed multi-task framework for natural language inference in clinical domain. Proceedings of the 18th BioNLP Workshop and Shared Task. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. pp. 488–492. doi:10.18653/v1/W19-5052

23. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning Bidirectional Encoder Representations from Transformers (BERT)-based models on large-scale electronic health record notes: An empirical study. JMIR Med Inform. 2019;7: e14830. doi:10.2196/14830

24. Xu D, Gopale M, Zhang J, Brown K, Begoli E, Bethard S. Unified Medical Language System resources improve sieve-based generation and Bidirectional Encoder Representations from Transformers (BERT)-based ranking for concept normalization. J Am Med Inform Assoc. 2020;27: 1510–1519. doi:10.1093/jamia/ocaa080

25. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform. 2020;110: 103541. doi:10.1016/j.jbi.2020.103541

26. Wang Y, Hur B, Verspoor K, Baldwin T. A Multi-pass sieve for clinical concept normalization. Traitement Automatique des Langues. 2020;61: 41–65.

27. Khosla S, Vashishth S, Lehman JF, Rose C. MedFilter: Improving extraction of task-relevant utterances through integration of discourse structure and ontological knowledge. Proceedings

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. pp. 7781–7797. doi:10.18653/v1/2020.emnlp-main.626

28. Peng W, Liu J, Wang M, Li L, Meng X, Yang H, et al. Huawei's submissions to the WMT20 Biomedical Translation Task. Proceedings of the 5th Conference on Machine Translation (WMT). Association for Computational Linguistics; 2020. pp. 857–861.

29. Corral A, Saralegi X. Elhuyar submission to the Biomedical Translation Task 2020 on terminology and abstracts translation. Proceedings of the 5th Conference on Machine Translation (WMT). Association for Computational Linguistics; 2020. pp. 813–819.

30. Soto X, Perez-de-Vinaspre O, Labaka G, Oronoz M. Ixamed's submission description for WMT20 Biomedical shared task: Benefits and limitations of using terminologies for domain adaptation. Proceedings of the 5th Conference on Machine Translation (WMT). Association for Computational Linguistics; 2020. pp. 875–880.

31. Kalyan KS, Sangeetha S. Target concept guided medical concept normalization in noisy user-generated Texts. Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. pp. 64–73. doi:10.18653/v1/2020.deelio-1.8

32. Pattisapu N, Prabhu N, Bhati S, Varma V. Leveraging social media for medical text simplification. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery; 2020. pp. 851–860. doi:10.1145/3397271.3401105

33. Yadav S, Pallagani V, Sheth A. Medical knowledge-enriched textual entailment framework. Proceedings of the 28th International Conference on Computational Linguistics. Association for Computational Linguistics; 2020. pp. 1795–1801.

34. Hristov A, Tahchiev A, Papazov H, Tulechki N, Primov T, Boytcheva S. Application of deep learning methods to SNOMED CT encoding of clinical texts: From data collection to extreme multi-label text-based classification. Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications. 2021. pp. 557–565. doi:10.26615/978-954-452-072-4_063

35. Dai X, Rybinski M, Karimi S. SearchEHR: A family history search system for clinical decision support. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2021. pp. 4701–4705. doi:10.1145/3459637.3481986

36. Chang D, Lin E, Brandt C, Taylor RA. Incorporating domain knowledge into language models by using graph convolutional networks for assessing semantic textual similarity: Model development and performance comparison. JMIR Med Inform. 2021;9: e23101. doi:10.2196/23101

37. Montañés-Salas R, López-Bosque I, García-Garcés L, del-Hoyo-Alonso R. ITAINNOVA at SocialDisNER: A Transformers cocktail for disease identification in social media in Spanish. Proceedings of the 29th International Conference on Computational Linguistics. Association for Computational Linguistics; 2022. pp. 71–74.

38. Ying H, Luo S, Dang T, Yu S. Label refinement via contrastive learning for distantly-supervised named entity recognition. Findings of the Association for Computational Linguistics: NAACL 2022. Stroudsburg, PA, USA: Association for Computational Linguistics;

2022. pp. 2656–2666. doi:10.18653/v1/2022.findings-naacl.203

39. Wang W, Meng X, Yan S, Tian Y, Peng W. Huawei BabelTar NMT at WMT22 Biomedical Translation Task: How we further improve domain-specific NMT. Proceedings of the Seventh Conference on Machine Translation (WMT). Association for Computational Linguistics; 2022. pp. 930–935.

40. Zotova E, Cuadros M, Rigau G. ClinIDMap: Towards a clinical IDs mapping for data interoperability. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA); 2022. pp. 3661–3669.

41. Tang G, Liu T, Cai X, Gao S, Fu L. Standardization of clinical terminology based on hybrid recall and Ernie. Proceedings of 2022 3rd International Symposium on Artificial Intelligence for Medicine Sciences. New York, NY, USA: Association for Computing Machinery; 2022. pp. 19–23. doi:10.1145/3570773.3570782

42. Han S, Zhang RF, Shi L, Richie R, Liu H, Tseng A, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform. 2022;127: 103984. doi:10.1016/j.jbi.2021.103984

43. Chen Y, Hu D, Li M, Duan H, Lu X. Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. Int J Med Inform. 2022;159: 104676. doi:10.1016/j.ijmedinf.2021.104676

44. Xu D, Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. J Biomed Inform. 2022;130: 104080. doi:10.1016/j.jbi.2022.104080

45. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform. 2022;126: 103983. doi:10.1016/j.jbi.2021.103983

46. Jha K, Zhang A. Continual knowledge infusion into pre-trained biomedical language models. Bioinformatics. 2022;38: 494–502. doi:10.1093/bioinformatics/btab671

47. Borchert F, Schapranow M-P. HPI-DHC @ BioASQ DisTEMIST: Spanish biomedical entity linking with pre-trained transformers and cross-lingual candidate retrieval. CEUR Workshop Proceedings. Conference and Labs of the Evaluation Forum; 2022. pp. 244–258.

48. Zhang X, Yu BXB, Liu Y, Chen G, Ng GW-Y, Chia N-H, et al. Conversational system for clinical communication training supporting user-defined tasks. Proceedings of the 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE). IEEE; 2022. pp. 396–403. doi:10.1109/TALE54877.2022.00071

49. Morine MJ, Priami C, Coronado E, Haber J, Kaput J. A comprehensive and holistic health database. Proceedings of 2022 IEEE International Conference on Digital Health (ICDH). IEEE; 2022. pp. 202–207. doi:10.1109/ICDH55609.2022.00039

50. Li L, Zhai Y, Gao J, Wang L, Hou L, Zhao J. Stacking-BERT model for Chinese medical procedure entity normalization. Math Biosci Eng. 2023;20: 1018–1036. doi:10.3934/mbe.2023047

51. Llorca I, Borchert F, Schapranow M-P. A meta-dataset of German medical corpora: Harmonization of annotations and cross-corpus NER evaluation. Proceedings of the 5th Clinical Natural Language Processing Workshop. Stroudsburg, PA, USA: Association for Computational Linguistics; 2023. pp. 171–181. doi:10.18653/v1/2023.clinicalnlp-1.23

52. Hristov A, Ivanov P, Aksenova A, Asamov T, Gyurov P, Primov T, et al. Clinical text classification to SNOMED CT codes using transformers trained on linked open medical

ontologies. Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings. 2023. pp. 519–526. doi:10.26615/978-954-452-092-2_057

53. Makhervaks D, Gillis P, Radinsky K. Clinical contradiction detection. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2023. pp. 1248–1263. doi:10.18653/v1/2023.emnlp-main.80

54. Dong H, Chen J, He Y, Liu Y, Horrocks I. Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery; 2023. pp. 452–462. doi:10.1145/3583780.3615036

55. Xu J, Mazwi M, Johnson AEW. AnnoDash, a clinical terminology annotation dashboard. JAMIA Open. 2023;6: ooad046. doi:10.1093/jamiaopen/ooad046

56. Liu F, Liu M, Li M, Xin Y, Gao D, Wu J, et al. Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quant Imaging Med Surg. 2023;13: 3873–3890. doi:10.21037/qims-22-1158

57. Searle T, Ibrahim Z, Teo J, Dobson RJB. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained Transformer models. J Biomed Inform. 2023;141: 104358. doi:10.1016/j.jbi.2023.104358

58. Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. Proceedings of the 20th Chinese National Conference on Computational Linguistics. Huhhot, China: Chinese Information Processing Society of China; 2021. pp. 1218–1227.

59. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for self-supervised learning of language representations. arXiv. 2019; 1909.11942.

60. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2022;3: 1–23. doi:10.1145/3458754

61. Alsentzer E, Murph J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, MN, USA: Association for Computational Linguistics; 2019. pp. 72–78.

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

63. Bressem KK, Papaioannou J-M, Grundmann P, Borchert F, Adams LC, Liu L, et al. medBERT.de: A comprehensive German BERT model for the medical domain. Expert Syst Appl. 2024;237: 121598. doi:10.1016/j.eswa.2023.121598

64. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. arXiv. 2022; 2210.10341.

65. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. pp. 7871–7880.

66. Karimi S, Metke-Jimenez A, Kemp M, Wang C. CADEC: A corpus of adverse drug event annotations. J Biomed Inform. 2015;55: 73–81. doi:10.1016/j.jbi.2015.03.010

67. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data Brief. 2019;24: 103838. doi:10.1016/j.dib.2019.103838

68. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. pp. 1586–1596. doi:10.18653/v1/D18-1187

69. Mohan S, Li D. MedMentions: A large biomedical corpus annotated with UMLS concepts. arXiv. 2019; 1902.09476. doi:10.48550/arxiv.1902.09476

70. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B, editors. Information access evaluation multilinguality, multimodality, and visualization. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. pp. 212–231. doi:10.1007/978-3-642-40802-1_24

71. Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. Proceedings of the 34th International Conference on Machine Learning. 2017;70: 2168–2178.

72. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations 2017. ICLR; 2017.

73. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. Pac Symp Biocomput. 2020;25: 295–306. doi:10.1142/9789811215636_0027

74. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001; 17–21.

75. D'Souza J, Ng. Sieve-based entity linking for the biomedical domain. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Association for Computational Linguistics; 2015. pp. 297–302.

76. Barrault L, Biesialska M, Bojar O, Costa-jussà MR, Federmann C, Graham Y, et al. Findings of the 2020 Conference on Machine Translation (WMT20). Proceedings of the Fifth Conference on Machine Translation. 2020.

77. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Morristown, NJ, USA: Association for Computational Linguistics; 2001. pp. 311–318. doi:10.3115/1073083.1073135

78. Lin C-Y, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04. Morristown, NJ, USA: Association for Computational Linguistics; 2004. pp. 605–612. doi:10.3115/1218955.1219032

79. Björk B-C, Solomon D. The publishing delay in scholarly peer-reviewed journals. Journal of Informetrics. 2013;7: 914–923. doi:10.1016/j.joi.2013.09.001

80. Spasic I, Nenadic G. Clinical text data in machine learning: Systematic review. JMIR Med Inform. 2020;8: e17984. doi:10.2196/17984

81. Jonnagaddala J, Liaw S-T, Ray P, Kumar M, Chang N-W, Dai H-J. Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform. 2015;58 Suppl: S203–S210. doi:10.1016/j.jbi.2015.08.003

82. Sezgin E, Hussain S-A, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: Feasibility study with real-world ata. JMIR Formativ Res. 2023;7: e43014. doi:10.2196/43014

83. Humm BG, Archer P, Bense H, Bernier C, Goetz C, Hoppe T, et al. New directions for applied knowledge-based AI and machine learning. Informatik Spektrum. 2023;46: 65–78. doi:10.1007/s00287-022-01513-9

84. Yang L, Chen H, Li Z, Ding X, Wu X. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Trans Knowl Data Eng. 2024; 1–20. doi:10.1109/TKDE.2024.3360454

85. Confalonieri R, del Prado FM, Agramunt S, Malagarriga D, Faggion D, Weyde T, et al. An ontology-based approach to explaining artificial neural networks. arXiv preprint. 2019; 1906.08362.

Figures

Figure 1. Flow diagram of article selection process. *SCT*, SNOMED CT.

Figure 2. Integrating SNOMED CT description information into large language models.

Figure 3. Integrating SNOMED CT relation information into large language models.

Figure 4. Integrating SNOMED CT entity type information into large language models.

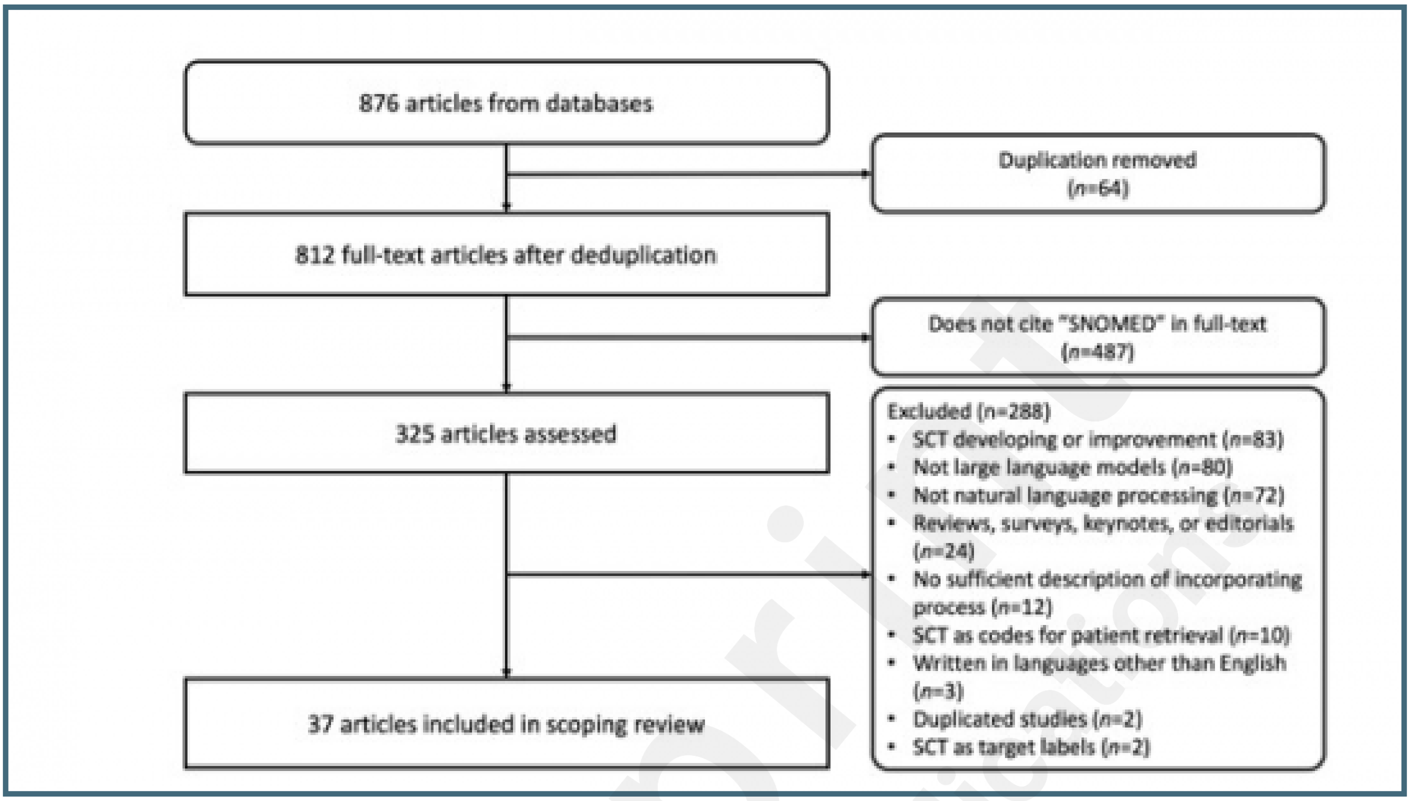
Figure 5. Integrating SCT into additional fusion modules. *LM*. language model

Figure 6. Retrieval-augmented knowledge fusion. *LLM*, large language model; *LM*, language model

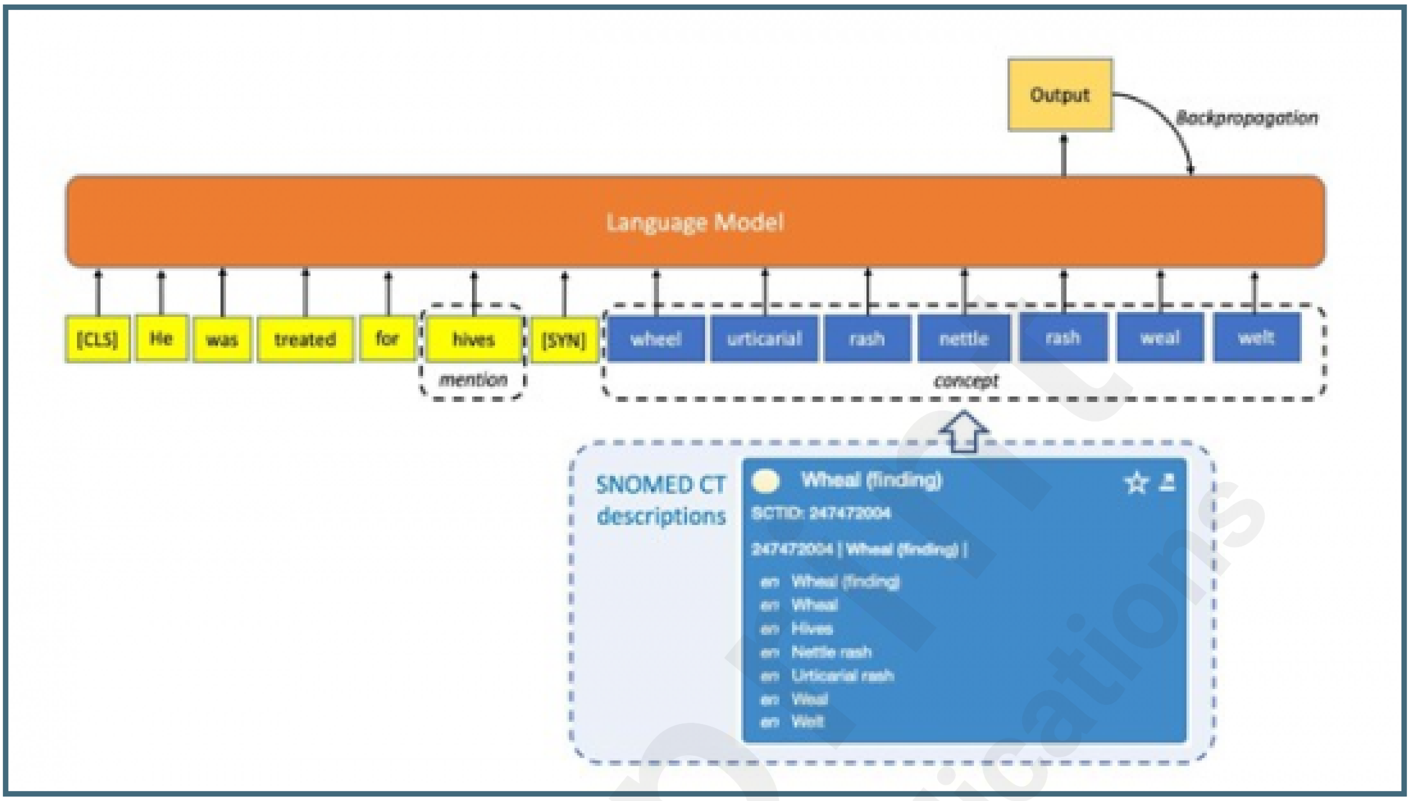
Supplementary Files

Figures

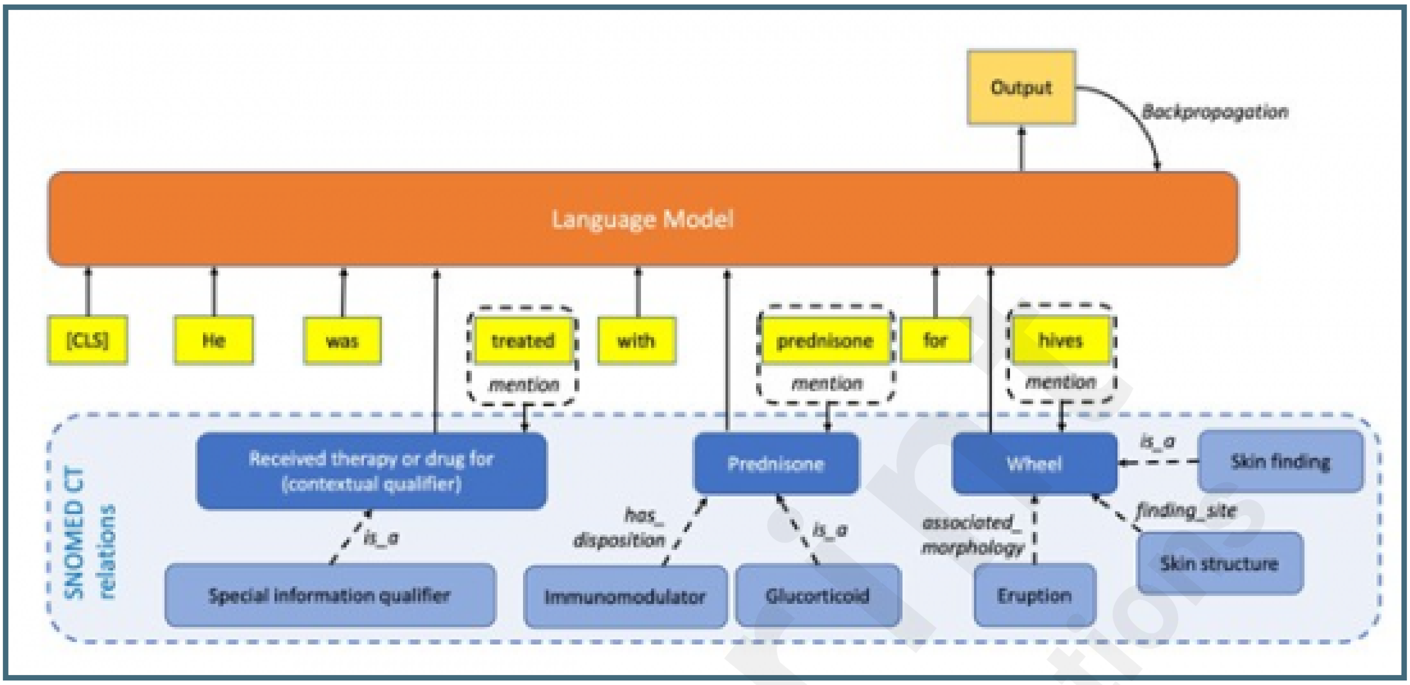
Flow diagram of article selection process. SCT, SNOMED CT.


876 articles from databases
Duplication removed (*n*=64)
812 full-text articles after deduplication
Does not cite "SNOMED" in full-text (*n*=487)
325 articles assessed
Excluded (n=288)
• SCT developing or improvement (*n*=83)
• Not large language models (*n*=80)
• Not natural language processing (*n*=72)
• Reviews, surveys, keynotes, or editorials (*n*=24)
• No sufficient description of incorporating process (*n*=12)
• SCT as codes for patient retrieval (*n*=10)
• Written in languages other than English (*n*=3)
• Duplicated studies (*n*=2)
• SCT as target labels (*n*=2)
37 articles included in scoping review

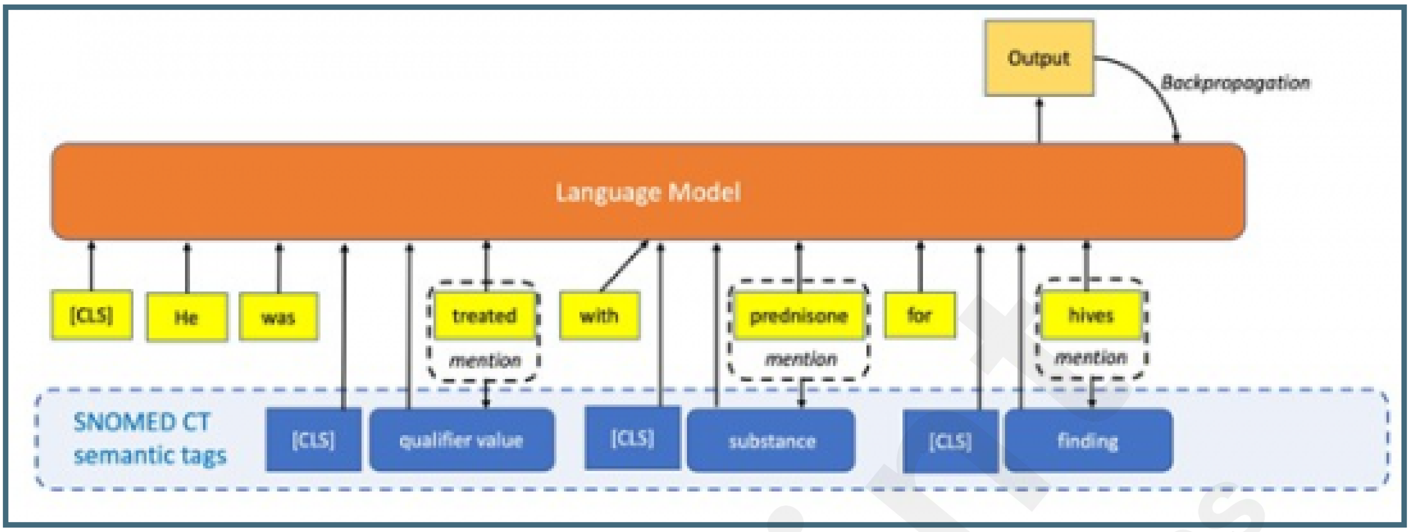
Integrating SNOMED CT description information into large language models.



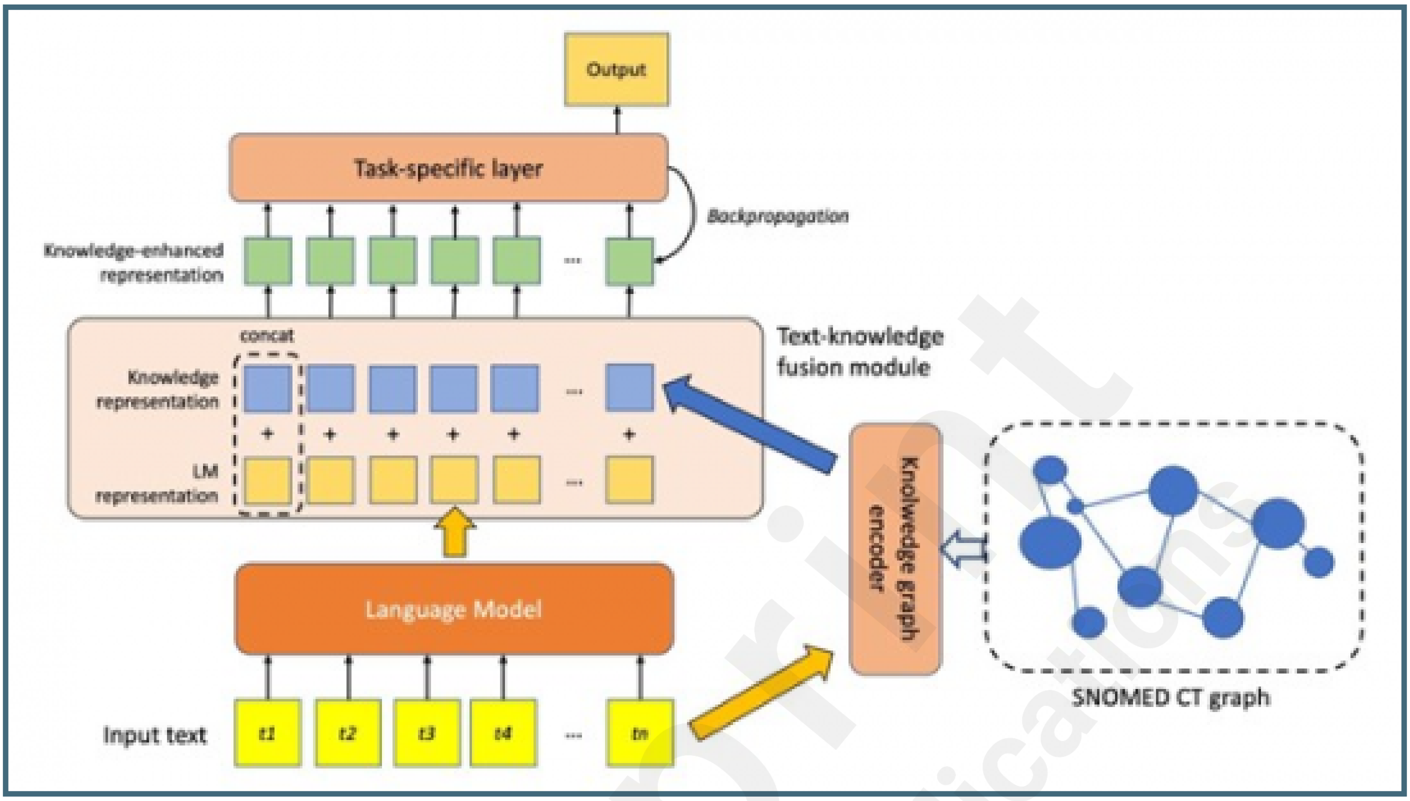
Integrating SNOMED CT relation information into large language models.



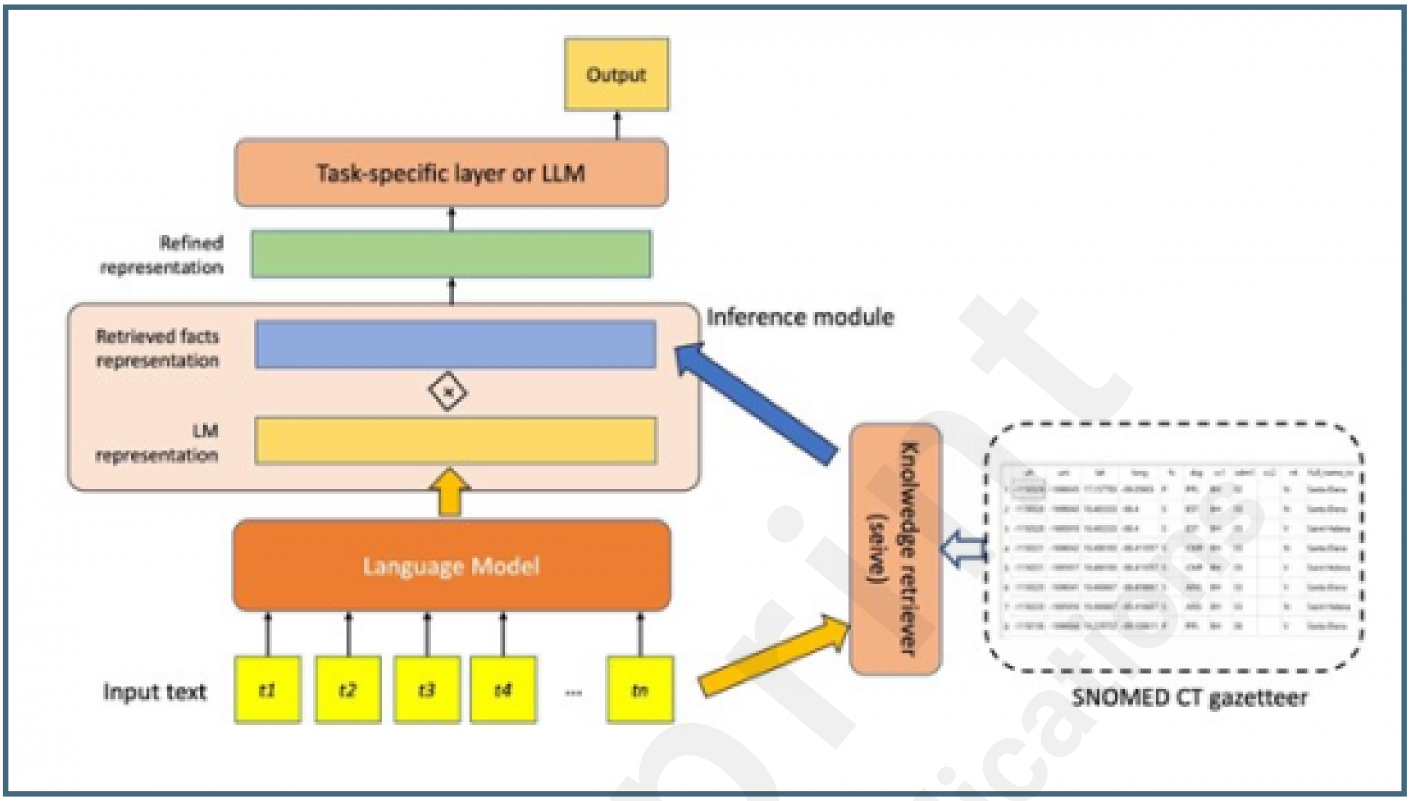
Integrating SNOMED CT entity type information into large language models.



Integrating SCT into additional fusion modules. LM. language model.



Retrieval augmented knowledge fusion. LLM, large language model; LM, language model.



Multimedia Appendixes

Summary of included studies.
URL: http://asset.jmir.pub/assets/88e9f19ec65687b6405b05e187ce9864.xlsx