MedCT: A Clinical Terminology Graph for Generative AI Applications in Healthcare

Ye Chen[1,3][0009−0007−6024−6562], Dongdong Huang[1,2], Yuqiang Shen[1], Haoyun Xu[3], Lin Sheng[3], Qingli Zhou[1], and Kai Wang[1,2,4][0000−0003−4328−8799]

[1] Department of Respiratory and Critical Care Medicine,
The Fourth Affiliated Hospital of School of Medicine, Zhejiang University, China
[2] Zhejiang Key Laboratory of Precision Diagnosis and Treatment for Lung Cancer, China
{8016009,yuqiangs,zhouql,kaiw}@zju.edu.cn
[3] Tiger Research, Shanghai, China
{yechen,haoyun.xu,lin.sheng}@tigerbot.com
[4] Corresponding author
kaiw@zju.edu.cn

Abstract. While recent advances in large language models (LLMs) offer great promise in healthcare, the safety-critical nature of the domain requires a thoughtful strategy to mitigate risks of hallucinations and potential harms. We propose a graph of domain knowledge and empirically validated its effectiveness in graph-augmented LLM generation. This presents a promising approach to safe, factual, and in turn more precise LLM applications. We first introduce the world's first clinical terminology for the Chinese healthcare community, namely MedCT, accompanied by a clinical foundation model MedBERT and an entity linking model MedLink. The MedCT system enables standardized representation of clinical data, successively stimulating the development of new medicines, treatment pathways, and better patient outcomes. Moreover, the MedCT knowledge graph provides a principled mechanism to minimize the hallucination problem of LLMs, therefore achieving significant levels of accuracy and safety in LLM-based clinical applications. Our experiments show that the MedCT system achieves state-of-the-art (SOTA) performance in semantic matching and entity linking tasks, not only for Chinese but also for English. We also conducted a longitudinal field experiment by applying MedCT and LLMs in a representative spectrum of clinical tasks, including electronic health record (EHR) auto-generation and medical document search. Our study shows a multitude of values of MedCT for clinical workflows and patient outcomes, especially in the new genre of clinical LLM applications. We present our approach in sufficient engineering detail, such that implementing a clinical terminology for other non-English societies should be readily reproducible. To encourage further research on LLM-based healthcare digitalization and promote the wellbeing of humankind, we are releasing our terminology, models, algorithms, and real-world clinical datasets for development.

Keywords: Large Language Model · LLM · Application · Healthcare · Clinical Terminology · Knowledge Graph.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

1 Introduction

Standard clinical terminologies, e.g., SNOMED CT, LOINC, ICD, can enable a multitude of values for global healthcare systems. For individual patients and clinicians, terminology or ontology coded electronic health records (EHR) greatly boost the consistency and interoperability of clinical data, and in turn increase the opportunities for real-time decision support for care delivering, retrospective reporting and analytics for research, precision medicine, and management [25].

While clinical terminology, ontology, or knowledge graph has been widely perceived as pivotal for healthcare practice and research, the daunting cost of building and optimization has hindered their wide adoption or effective use. SNOMED CT is considered to be the most comprehensive and widely adopted clinical terminology in the world [4]. It was established in 1965, has undergone more than twenty years development with over one hundred million dollar investment [26]. The remarkable advances of large language models (LLMs) [5] have inspired us to explore rapidly building high-quality and reliable terminologies for the healthcare domain. In particular, we address the problem of developing a clinical terminology for the Chinese clinical domain, namely MedCT. Previous work on coding Chinese clinical terms [9,13] mainly focused on better semantic matching to existing terminology predominantly in English, but none has developed a clinical terminology truly grounded on underserved languages.

As LLMs have been increasingly applied and deployed to real-world healthcare and clinical settings [29], hallucinations or fabricated information, remains one of the prominent challenges. However, the safety-critical nature of the healthcare domain requires a more deterministic approach to restraining hallucination, therefore motivating us to explore the other side of the AI world. In particular, we augment LLM generation with a model of truth, the MedCT knowledge graph. We further deployed the MedCT terminology to a representative spectrum of real world clinical applications. To summarize, we believe that we have made the following contributions to the global healthcare system in the AI era, and the rest of the paper is logically structured likewise.

1. MedCT: the world's first open Chinese clinical terminology at the scale comparable to SNOMED CT.
2. A suite of models and algorithms for readily adoption of the above terminology, namely, MedBERT, a pretrained foundation model, and MedLink, a fine-tuned entity linking model.
3. A holistic approach with implementation details for rapid and cost-efficient development of clinical terminology for other underserved languages.
4. A wide and representative spectrum of real-world clinical applications utilizing the MedCT system, to demonstrate its value propositions and provide a reference framework of truth-augmented LLM applications in healthcare.
5. Finding and observations from the field with regards to the status quo of applying LLMs in real-world clinical setting, e.g., large or small models, LLM or classical NLP techniques, general or domain-specialized models.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

2 Methodology

We bootstrap our development from SNOMED CT, that is considered to be the most comprehensive and widely adopted clinical terminology, therefore inheriting decades of its achievements. We first pretrained a LLM, namely Tigerbot-3 [10], continually from Llama-3.1 [1] to strengthen biomedical base knowledge (with medical training data in Table 2) and multilingual coverage (especially Chinese for our applications). We then applied LLM to contextualize and translate the SNOMED concepts into Chinese, thus forming our initial MedCT terminology. Further, we collaborated with a tertiary care hospital for truth-grounding the terminology, through annotating real-world EHRs with MedCT while revising the terminology for correction and localization.

At the core of the models and algorithms to utilize MedCT is a clinical foundation model, called MedBERT. We pretrained MedBERT from scratch using a thoughtfully curated clinical dataset, and yielded SOTA performance in semantic understanding. Next, with the MedCT annotated clinical data, we trained MedLink, fine-tuned models for clinical terminology named entity recognition (NER) and linking (NEL). After we deployed MedCT in the field, the learning process is iteratively reinforced, for both the MedCT terminology and entity linking models. Our work is inspired by the SNOMED CT entity linking challenge [16], and our method largely follows the model-based winning solution SNOBERT [22]. We managed to push the boundary further in model performance and multilingual coverage, by leveraging LLMs and well-curated real-world clinical data. Table 1 illustrates some examples of clinical notes with annotated MedCT concepts.

Table 1. Clinical Note Snippets with Concept Annotations

Notes	Concept ID (hier.), name & syn.
. . . 右肾上腺[1]结节，建议进一步检查。	29392005 (body) Right adrenal gland 右肾上腺
. . . 支气管扩张试验：吸入沙丁胺醇400ug. . .	415299008 (procedure) Reversibility trial by bronchodilator 支气管扩张剂可逆性试验
. . . 附见：两侧胸腔少量积液伴邻近肺不张。	425802001 (finding) Bilateral pleural effusion 双侧胸腔积液
. . . 神志清，精神一般，胃纳差，睡眠差. . .	64379006 (finding) Decrease in appetite 没胃口，食欲下降

[1] We color-coded concept hierarchy as: green-body, blue-procedure, and red-finding.

2.1 MedBERT: a clinical foundation model

Of the central importance for most clinical NLP tasks is a foundation model to encode broad and basic semantics in the domain. Previous work shows that for domains with a copious amount of unlabeled texts, pretraining language models from scratch yielded substantial gains over continual pretraining from general-domain models [17]. Biomedicine is one of such high-resourced domains. Specifically, we pretrain a BERT model, namely MedBERT, from scratch using a biomedical dataset curated with the following design considerations, with statistics and sources outlined in Table 2.

1. A large corpora of biomedical literature and publications with comprehensive and timely coverage of the domain, e.g., the PubMed Central (PMC) repository [2].
2. Data from the field and directly relevant to downstream tasks, e.g., clinical guidelines [7,8,12,27,34] and real-world clinical notes MIMIC-IV [21].
3. Clinical terminologies and their contexts, e.g., SNOMED CT and MedCT terms and descriptions.
4. Multilingual coverage, i.e., English and Chinese.

Table 2. MedBERT Training Data

Source	Dataset (lang)	Examples	Disk size
Publications	PMC abstracts (en) [2]	24,732,786	26G
	PMC full-texts (en) [2]	3,775,772	109G
	PMC patients (en) [35]	167,034	444M
	PubMedQA contexts (en) [20]	211,269	280M
	Open medical books (en) [33]	13,000	11G
	Chinese literature (zh) [23]	27,704	14G
	Trad. Chinese medicine books (zh) [18]	17	13M
Guidelines	Clinical guidelines (en) [7,8]	11,184	527M
	Clinical guidelines (zh) [6,14]	4,364	643M
Clinical notes	MIMIC-IV v2.2 clinical notes (en) [21]	2,653,148	5.8G
	Chinese EHR and clinical notes (zh)	3,109,181	904M
Terminology	SNOMED and MedCT (en) [25]	723,552	23M
Total	—	**35,429,011**	**168G**

We compared the prediction accuracy of the fill-mask task between our MedBERT and other SOTA biomedical and general-domain models, as the results exhibited in Table 3. First, we verified that domain-specific training has advantages, as biomedical models outperform general-domain BERT models by about twenty percentage points. Second, multilingual expansion is critical. Although the evaluation dataset only has less than 10% Chinese data, the

DOI: 10.1007/978-3-031-97557-8_3

multilingual and Chinese BERT surpass the English-only models by a large margin. Furthermore, the scale and quality of the training data tends to yield better model performance, as seen that BiomedBERT trained with PMC full text wins those with PubMed abstracts only.

Table 3. MedBERT Evaluation

Domain	Model	Accuracy
Biomedical	BiomedBERT-base-fulltext [17]	0.5633
	BiomedBERT-large-abstract	0.5100
	BiomedBERT-base-abstract	0.4209
	SciBERT [3]	0.5819
	MedBERT	**0.8344**
General	BERT-base-multilingual [15]	0.5333
	BERT-base-Chinese	0.5582
	BERT-large	0.3199
	BERT-base	0.3440

2.2 MedLink: clinical entity recognition and linking

We implemented a two-stage approach to recognizing clinical entities from free-text notes and linking the entities to the built MedCT concepts, as follows.

1. First stage: A NER segmentation task to detect spans of texts as clinical entity mentions.
2. Second stage: A NEL ranking task to predict the MedCT concepts for the recognized entities from the first stage.

For the first stage NER task, we fine-tuned a token classification model from the MedBERT foundation model, as described in Section 2.1. We classify each token into four classes: `{finding, procedure, body, none}`, using the BIO format [28], therefore a token tagging task with seven labels: `{O, B-find, I-find, B-proc, I-proc, B-body, I-body}`. At the second stage NEL task, we need to link segmented entity mentions to concepts in the MedCT ontology. This is a semantic matching task, which we therefore simply formulate it as a ranking problem in the embedding space. Specifically, we chose `SapBERT` [24] for English embedding, and its cross-lingual extension `SapBERT-all-lang` for multilingual and Chinese tasks. We measure the performance of trained models with character-level concept-averaged intersection-over-union (IoU). Table 4 exhibits the experimental results. Our MedLink model achieves SOTA performance in both English and Chinese clinical NER and NEL tasks. We conjecture that a stronger multilingual foundation model MedBERT and copious annotated real-world clinical training data largely contribute into the gain.

DOI: 10.1007/978-3-031-97557-8_3

Table 4. MedLink Evaluation

Type	Base model	English NEL (IoU on MIMIC)	Chinese NEL (IoU on MedCT)
Biomed	BiomedBERT-base-fulltext [17]	0.4797	0.0091
	BiomedBERT-large-abstract	0.4952	0.0005
	BiomedBERT-base-abstract	0.4976	0.0003
	SciBERT [3]	0.4993	0.0026
	MedBERT	**0.5065**	**0.3012**
General	BERT-base-multilingual [15]	0.4717	0.1006
	BERT-base-Chinese	0.4508	0.1516
	BERT-large	0.4868	0.0007
	BERT-base	0.4774	0.0002

3 Experiments and Applications

3.1 Large or small models

In this experiment, we compared the two methodologies in the medical NER and NEL applications: small specialized models versus large general models (LLMs). The experimental results are shown in Table 5. Let us first consider the most generalized LLM approach that does not rely on any specialized data. Under this setting, GPT-4o only yields 0.11 IoU on English data and 0.17 IoU on Chinese data, substantially inferior to our MedLink small model approach, that is 0.51 IoU for English and 0.30 IoU for Chinese. Although the LLM results are visually plausible, its numerical measurement of performance is suboptimal. LLM approach with GPT-4o incurs considerably more inference time than small model approach with MedCT, more than 10 times for English and 30 times for Chinese tests. Moreover, for this medical NER task, open-source model (Llama-3) performs comparably as close model (GPT-4o).

Table 5. MedLink vs. LLM approach

Model	English NEL (51 MIMIC notes)		Chinese NEL (1860 MedCT notes)	
	IoU	Time	IoU	Time
MedLink	**0.5065**	1m40s	**0.30117**	4m15s
GPT-4o[1]	0.1146	13m46s	0.1739	116m46s
Llama-3.1-70B[2]	0.1116	102m58s	0.1689	661m26s

[1,2] Both GPT-4o and Llama-3.1-70B model downloading were executed as of this writing in January, 2025.

DOI: 10.1007/978-3-031-97557-8_3

3.2 Retrospective retrieval of health records

In this experiment, we wish to validate and measure the value of the clinical terminology MedCT in the application of health record retrieval. A majority of retrospective retrieval of EHRs involves finding cases with similar or related diseases, in reference past evidences in testing, diagnosis, treatment and outcome. Therefore, from the `MedCT-clinical-notes` dataset we curated in-house, we took a corpus of discharge summaries for this retrieval experiment. The corpus contains 13,863 examples or discharge notes, entered from all departments during the first quarter of year 2024 in a tertiary care hospital. The data was organized into relevant textual fields including demographics, admission, treatment pathway, discharge summary and instruction. We interviewed a panel of 12 senior physicians to collected a set of 20 queries representative of real-world clinical practice and research. The clinical query set was chosen with non-trivial complexity such that a straightforward keyword match conceivably cannot yield satisfactory results. One example is “post-stroke with pneumonia”.

We implemented two retrieval strategies, the classic sparse or dense retrieval and the MedCT-augmented retrieval. For the MedCT-augmented retrieval approach, we first offline tagged each document with MedCT concepts, and then indexed the list of concept ids along with texts per document. These annotated concepts should capture almost all relevant clinical information in the health records. At online retrieval time, we annotated full text queries with MedCT concepts, and then ranked documents with a hybrid strategy based on both text-based sparse or dense retrieval and strict concept id matching. For evaluation, we asked the same panel clinicians to annotate relevant examples from a random sample of 2K discharge notes, for each of the 20 queries. This ground truth allows us to measure precision, recall and the balanced F_1 score as the performance metrics for our retrieval task, as reported in Table 6. Our experiments show that retrieval augmented with MedCT graph substantially outperforms modern text-based search. In particular, MedCT boosts the search recall by a 15% lift over sparse retrieval.

Table 6. EHR retrieval augmented with MedCT

Retrieval method	Precision[1]	Recall[2]	F_1-score[3]
Sparse	0.5294	0.5015	0.5151
Dense	0.0706	0.0995	0.0826
Hybird	0.3882	0.2527	0.3061
MedCT-augmented	**0.6235**	**0.5745**	**0.5980**

[1,2,3] All metrics are measured at top 10 retrieved results, representative of a typical search scenario.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

3.3 Health records auto-generation by LLMs

Next we consider the task of health records generation. Nearly half of physician's time is devoted to digital paperwork, rather than direct patient care [30]. The statistics is even worse in regions and countries with shortfall of health workforce. In our field study at a tertiary care hospital in China, reportedly near 90% of residency doctors' time is absorbed in writing clinical notes and medical records. Among various health records, discharge summary is arguably the most important document a hospitalist writes. The discharge summary is a semi-structured narrative document for communicating clinical information about patients. However, nowadays hospitalists have little time to write good quality discharge summaries, along with often delay to deliver to downstream physicians, causing disruption in the continuity of care and risking poor patient outcomes.

Therefore it is appealing to apply LLMs to generate health record drafts for physicians, to review, mildly edit and submit. This is a text summarization task with moderate complexity yet high practical significance [11]. A good model, likely with domain or task-specific training, would both speed up clinical documentation and improve the quality of health records. In our deployment to a tertiary care hospital in Zhejiang, China, hospitalists reported about 40% reduction in time spent in writing discharge summaries with the help of LLM generation, while observing improvement in both quality and information density.

However, general-purpose LLMs, if used as-is in a vanilla fashion, typically cannot meet the safety requirements of the medical domain [31]. In our controlled experiments, for instance, we observed that vanilla LLMs hallucinated medical misinformation such as made-up procedures (e.g., Laparoscopy for radical pulmonary surgery) and medications (e.g., Cefuroxime) in discharge summary generation. This represents a significant risk of applying LLMs in an ignorant way in mission-critical domains like healthcare. Many of these hallucinations were trivial to identified by qualified physicians, which also symbolizes the large gap between human intelligence and LLMs, especially in domain knowledge.

In order to address the hallucination problem intrinsic to LLMs, we guide the LLM generation with a knowledge graph as source of truth. We believe that our approach brings together the strengths of both worlds, LLMs and specialized small models; and moreover presents a systematic and measurable way to minimize hallucination. As illustrated in the detailed prompts in Figure 1, the MedCT-guided generation instructs the LLM to attend to major clinical concepts such as "chief complaint" and "physical examination" and therefore should capture key clinical information more comprehensively and accurately.

To evaluate the generation results, we recruited a panel of nine hospitalists to review summary generations from the above two methods, along with the human summary by doctors, in a blind fashion, and then cast Likert-scale to each testing example. Evaluating text summarization models is nontrivial, especially with automatic metrics. In general domains, such as TL;DR Reddits and CNN/DM news article summarization, previous works have used ROUGE or reward models to predict human preference [32]. But these metrics are only rough proxies to real human perceived summary quality, and should not apply indiscriminately

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

to different domains or even different tasks. For example, in book and news article summarization, coherence is often used to measure how easy the summary is to read on its own. But for the task of health record summarization in the clinical domain, with the time pressure and norm use of medical abbreviation and terminology, conciseness and clarity weigh more than coherence. After three sessions of panel discussions, we developed a set of metrics for our health record summarization task. The metrics cover both general language quality and clinical significance, from perspectives of accuracy, completeness, clarity, relevance, conciseness, and clinical depth. Moreover, our evaluation metrics, along with the annotated preference dataset, shall be instrumental to develop automatic metrics for text summarization in the clinical domain.

# Zero-shot prompt for vanilla LLM approach:
`{input context: hospital course, discharge diagnosis}`

The above is a detailed hospital course and discharge diagnosis from a medical record. Please summarize it into an accurate and concise medical summary. The summary should include the reason for admission, basis for diagnosis, main treatment measures and their effects, changes in condition, and status at discharge.

Medical summary:

# Zero-shot prompt for MedCT-guided LLM approach:
`{input context: hospital course, discharge diagnosis}`

The above is a detailed hospital course and discharge diagnosis from a medical record. Please summarize it into an accurate and concise medical summary. The summary should include the reason for admission, basis for diagnosis, main treatment measures and their effects, changes in condition, and status at discharge.

Also, the above medical records include the following entities, please include these medical entities in the medical course summary.
`{entity mentions from the input}`

Medical summary:

Fig. 1. Prompts for discharge summary auto-generation

Our evaluation dataset contains 91 examples of discharge notes, with detailed hospital courses and discharge diagnosis as raw input (denoted as raw), along with discharge summaries written by human hospitalists (denoted as human). We then infer the underlying LLM with two prompting approaches as in Figure 1 (denoted as LLM and MedCT, respectively). The average length in character is 3,545 for the input clinical notes, 542 for the human summary, 274 for the vanilla LLM, and 317 for the MedCT method. LLMs tend to condense more than human, while the MedCT-augmented method conveys richer information than simple LLM prompting. This is as expected, since we instruct the LLM to preserve information regarding clinical entities. For a rapid computerizable evaluation, we compute cosine similarities, in the embedding space projected by MedBERT, between raw input, human and machine generations. With similar compression rates, cosine similarity is a reasonable proxy to how well a summary captures the original text's main points. As shown in Table 7, our MedCT-augmented generation achieved best cosine scores, notably even higher than human summaries. Furthermore once again, for tasks that require deep domain knowledge such as clinical notes summarization, proprietary and open-source models empirically yielded comparable performance.

Table 7. EHR summarization cosine similarity

cosine	**Raw**			**Human**	
	Human	LLM	MedCT	LLM	MedCT
GPT-4o	0.8940	0.8984	**0.9288**	0.9242	**0.9257**
Llama-3.1-70B	0.8940	0.8897	**0.9066**	**0.9163**	0.9150

While the programmable cosine similarity gives a rapid proxy to summarization quality, especially for model iteration and comparison, the gold standard is still human review. We distributed the 91 testing examples to nine hospitalists with tenure from ten years or above. For each example, we gave original input clinical notes, and three summary generations, from human doctors, simple LLM, and LLM augmented with MedCT graph (denoted as human, LLM and MedCT, respectively). More importantly, the review was conducted in a blind manner. The three generations were randomly shuffled per example (only organizers held the true orders), and hence reviewers did not know which model or human generated the summary to be scored. The physician reviewers were instructed to rate summaries using 5-point Likert-scale by real-world clinical standards in reference to the above six-dimension metrics. One of the entries is factually human summary by real doctor anyway. The results are shown in Table 8.

Overall, our graph-guided LLM approach achieves highest human ratings, winning five out of six review dimensions. Notably, the gains from the perspectives of "clinical depth" and "relevance" are particularly substantial, over both LLM and human generations. By additionally prompting LLM with MedCT-recognized clinical entities, the underlying models exhibited precise attentions to clinical concepts in the free-text notes. As illustrated in the snippets in Figure 2, the MedCT-augmented approach captured the finding of "pulmonary infection" in hospital course summary, while simple LLM and human missed this concept. From the input notes of this patient, "pulmonary infection" is one of the complications through most of her or his hospital course. Both machine generations yielded comparable or better summarization with respect to factual human summaries.

Table 8. EHR summarization human review scores

	Human	LLM	MedCT
Accuracy	4.42	4.41	**4.42**
Completeness	3.99	4.19	**4.20**
Clarity	4.44	**4.45**	4.40
Relevance	4.56	4.77	**4.84**
Conciseness	4.23	4.27	**4.33**
Cinical depth	3.95	4.07	**4.19**
Overall	25.58	26.15	**26.36**

The win-or-tie rate of LLM approach is 0.59 over human, and even higher augmented by MedCT with a good odd of 0.68 being at least as good as human. This finding confirms with previous computerized evaluation of cosine similarity, suggesting that LLMs with domain specialization indeed can help clinical workflow tasks such as health record auto-generation.

Input clinical notes:
Chief complaint: The patient's temperature was the highest yesterday at 38°C,... Diagnosis: 1. Open craniocerebral injury,...3. Contusion of both lungs, bilateral pleural effusion with incomplete expansion of both lower lungs, and pulmonary infection ...

Human summary:
After admission, he was given mannitol to lower intracranial pressure,...insulin combined with metformin to lower blood sugar, and parenteral nutrition support.

LLM summary:
In terms of changes in condition, although the patient suffered from various complications, after careful treatment by the doctor, the patient gradually regained consciousness and .... Finally, the patient was discharged from the hospital on January 28, 2024, and planned to continue rehabilitation training after discharge.

LLM+MedCT summary:
After treatment, the patient's condition gradually stabilized, but there were still some problems, such as repeated fever, low level of consciousness, and lung infection . Finally, the patient was discharged from the hospital on January 28, 2024, and planned to undergo rehabilitation training after discharge.

Fig. 2. A snippet from clinical notes summarization

DOI: 10.1007/978-3-031-97557-8_3

4 Conclusions

Healthcare arguably remains one of the most prosperous domains beneficial from the rapid development of artificial intelligence (AI) in general and large language models (LLMs) in particular.

> “We’re going to have a family doctor who’s seen a hundred million patients and they’re going to be a much better family doctor.” [19]
>
> *Geoffrey Hinton*

However, even the state-of-the-art general foundation models, e.g., GPT-4o and Llama-3.1, merely scratch the surface encoding deep domain knowledge from high-resourced yet largely private domains such as healthcare. Moreover, the probabilistic root of LLMs, hence the tendency of hallucination, hinders their wide practical adoption in privacy and safety critical tasks. In the context of leveraging LLMs’ extraordinary capabilities of semantic understanding, generativeness and interactiveness, while ensuring their safety, unbiasedness, and honesty in real-world applications, we developed and released MedCT. To the best of our knowledge. MedCT is the world’s first clinical terminology built for and grounded from non-English community, specifically Chinese. We presented our comprehensive approach to building the clinical terminology knowledge graph, truth grounding and optimizing from real clinical data, training models for named entity recognition and linking to the graph. Our approach leverages LLMs as an integral part of development tools, along with abundant real-world clinical data and annotations by experienced physicians. Consequently, our MedCT models achieve new state-of-the-art in medical NER and NEL tasks (w.r.t. BiomedBERT and SciBERT etc.), especially in a rapid and cost-efficient manner (w.r.t. SNOMED CT etc.).

The values of our MedCT clinical terminology are even more pronounced in complementing LLM applications in clinical setting. A knowledge graph such as MedCT not only injects clinical domain commonsense into foundation models, but also as a source of truth gouges model generations to be more safe, truthful and reliable. We deployed MedCT in a wide variety of clinical applications, including clinical information retrieval and document summarization. Our findings from human blind reviews are inspiring, in that MedCT-augmented LLMs can achieve human-like or even better results in various tasks in clinical workflow and research tasks. Meanwhile, we also found that general-purpose LLM is no silver bullet, especially for domains with knowledge depth. As our approach stands, practical yet mission-critical applications of LLM still require domain specialization, for example in conjunction with classical yet surgical machine learning techniques.

We believe that we are at the dawn of unleashing the values of AI and LLMs for great humanity. In the hope of facilitating further development in

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

the healthcare domain, we open-source release the MedCT suite of models and datasets [5], which include:

1. The MedCT bilingual (English and Chinese) clinical terminology dictionary, with 223K medical concepts.
2. The MedCT named entity recognition (NER) model: MedLink.
3. A biomedical foundation model: MedBERT, that achieves state-of-the-art performance in a variety of downstream tasks, e.g., clinical NER/NEL, search, and summarization.
4. Our MedCT-clinical-notes dataset, including:
   - For the NER and NEL tasks, 7.4K real-world clinical notes in Chinese, and 61K entity mention annotations per MedCT graph.
   - For the search task, 20 clinical queries, and 2K discharge notes with relevance annotations.
   - For the clinical note summarization task, 91 raw discharge notes and summaries by human, LLM and MedCT-augmented generations, along with preference Likert-scale annotated by human physicians.

Acknowledgements

This work was supported by the Natural Science Foundation of China Key Program (U23A20467) and the Zhejiang Provincial Science and Technology Program (Grant No. 2024C03270) under the Pioneer and Leading Goose Research and Development Program.

References

1. AI@Meta: Llama 3.1 model card. https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct (10 2024)
2. Beck, J.: Report from the field: Pubmed central, an xml-based archive of life sciences journal articles. International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML (08 2010)
3. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. EMNLP 2019 (03 2019), https://arxiv.org/pdf/1903.10676.pdf
4. Benson, T.: Principles of Health Interoperability HL7 and SNOMED. Springer London (2010). https://doi.org/10.1007/978-1-84882-803-2, http://dx.doi.org/10.1007/978-1-84882-803-2
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. arXiv:2005.14165v4 [cs.CL] (05 2020), https://arxiv.org/pdf/2005.14165.pdf

[5] For a detailed description and ongoing releases, please see MedCT repository:
Github: https://github.com/TigerResearch/MedCT;
Huggingface: https://huggingface.co/collections/TigerResearch/medct-6744641d6f19b9d70a56f848

6. Cai, Q., Chai, R., Chen, G.: Clinical practice guidelines for intratympanic drug delivery. Chinese Journal of Otorhinolaryngology-skull Base Surgery **30**(1) (2024)
7. Cancer Care Ontario: Cancer care ontario (cco) guidelines & advice. https://www.cancercareontario.ca/en/guidelines-advice (2024)
8. Centers for Disease Control and Prevention: Centers for disease control and prevention: Guidelines and recommendations. https://www.guidelinecentral.com/guidelines/CDC/ (2024)
9. Chen, Y., Hu, D., Li, M., Duan, H., Lu, X.: Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. International Journal of Medical Informatics **159** (Mar 2022)
10. Chen, Y.: Tigerbot 3 model card. https://huggingface.co/TigerResearch/tigerbot-70b-chat-v6 (10 2024)
11. Chen, Y., Couto, I., Cai, W., Fu, C., Dorneles, B.: Softtiger: A clinical foundation model for healthcare workflows. AAAI 2024 Spring Symposium on Clinical Foundation Models (03 2024), `https://arxiv.org/pdf/2403.00868.pdf`
12. Cochrane Library: Cochrane database of systematic reviews. https://www.cochranelibrary.com/about/about-cochrane-reviews (2024)
13. Dai, R., Zhang, X., Li, F., Li, C.: Research on normalization of chinese clinical terms based on keyword extraction and data augmentation technology. In: Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science. pp. 1291–1298. ISAIMS 2023, ACM (Oct 2023)
14. Dai, Y., Li, G.: Oncology pharmacy clinic standards (trial). Chinese Pharmaceutical Journal **56**(9) (2021)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (10 2018), `https://arxiv.org/pdf/1810.04805.pdf`
16. Driven Data Inc.: SNOMED CT Entity Linking Challenge. https://www.drivendata.org/competitions/258/competition-snomed-ct/ (03 2024)
17. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing (07 2020). `https://doi.org/https://doi.org/10.1145/3458754`, `https://arxiv.org/pdf/2007.15779.pdf`
18. Guo, Z.: Explanation of huangdi neijing (1988)
19. Hinton, G.: Large language models in medicine. they understand and have empathy. https://erictopol.substack.com/p/geoffrey-hinton-large-language-models (12 2023)
20. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (11 2019)
21. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., wei H. Lehman, L., Celi, L.A., Mark, R.G.: Mimic-iv, a freely accessible electronic health record dataset. Scientific Data **10**(1) (2023)
22. Kulyabin, M., Sokolov, G., Galaida, A., Maier, A., Arias-Vergara, T.: Snobert: A benchmark for clinical notes entity linking in the snomed ct clinical terminology (05 2024), `https://arxiv.org/pdf/2405.16115.pdf`
23. Library, Z.U.: Chinese medical journal full text database. http://www.yiigle.com/ (2024)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3

24. 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 pp. 4228–4238 (06 2021), https://arxiv.org/pdf/2010.11784.pdf
25. Liu, K., Hogan, W.R., Crowley, R.S.: Natural language processing methods and systems for biomedical ontology learning. Journal of Biomedical Informatics **44**(1), 163–179 (2011). https://doi.org/https://doi.org/10.1016/j.jbi.2010.07.006, https://www.sciencedirect.com/science/article/pii/S153204641000105X, ontologies for Clinical and Translational Research
26. National Library of Medicine: United States National Library of Medicine. SNOMED Clinical Terms® To Be Added To UMLS® Metathesaurus®. http://www.nlm.nih.gov/research/umls/Snomed/snomed_announcement.html (2003)
27. NICE: National Institute for Health and Care Excellence: Nice guidelines. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-guidelines (2024)
28. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995), https://aclanthology.org/W95-0107
29. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2, https://doi.org/10.1038/s41586-023-06291-2
30. Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G.: Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. Annals of Internal Medicine (2016)
31. Stanceski, K., Zhong, S., Zhang, X., Khadra, S., Tracy, M., Koria, L., Lo, S., Naganathan, V., Kim, J., Dunn, A.G., Ayre, J.: The quality and safety of using generative ai to produce patient-centred discharge instructions. npj Digital Medicine **7**(1) (Nov 2024). https://doi.org/10.1038/s41746-024-01336-w, http://dx.doi.org/10.1038/s41746-024-01336-w
32. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (09 2020), https://arxiv.org/pdf/2009.01325.pdf
33. Walther, D.S.: Applied kinesiology synopsis 2nd edition (01 1988)
34. World Health Organization: Who guidelines. https://www.who.int/publications/who-guidelines (2024)
35. Zhao, Z., Jin, Q., Chen, F., Peng, T., Yu, S.: Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. Sci Data **10**, 909 (2023). https://doi.org/https://doi.org/10.1038/s41597-023-02814-8, https://arxiv.org/pdf/2202.13876.pdf

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_3