An LLM-based Framework for Biomedical Terminology Normalization in Social Media via Multi-Agent Collaboration

Yongqi Fan[♢], **Kui Xue**[♠], **Zelin Li**[♣], **Xiaofan Zhang**[♡♠], **Tong Ruan**[♢*]
[♢]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
[♠]Intelligent Healthcare, Shanghai Artificial Intelligence Laboratory, Shanghai, China
[♣]Northwestern University, USA [♡]Shanghai Jiao Tong University, Shanghai, China
johnnyfans@mail.ecust.edu.cn, zelinli2025@u.northwestern.edu
xuekui@pjlab.org.cn, xiaofan.zhang@sjtu.edu.cn, ruantong@ecust.edu.cn

Abstract

Biomedical Terminology Normalization aims to identify the standard term in a specified termbase for non-standardized mentions from social media or clinical texts, employing the mainstream "Recall and Re-rank" framework. Instead of the traditional pretraining-finetuning paradigm, we would like to explore the possibility of accomplishing this task through a tuning-free paradigm using powerful Large Language Models (LLMs), hoping to address the costs of re-training due to discrepancies of both standard termbases and annotation protocols. Another major obstacle in this task is that both mentions and terms are short texts. Short texts contain an insufficient amount of information that can introduce ambiguity, especially in a biomedical context. Therefore, besides using the advanced embedding model, we implement a Retrieval-Augmented Generation (RAG) based knowledge card generation module. This module introduces an LLM agent that expands the short texts into accurate, harmonized, and more informative descriptions using a search engine and a domain knowledge base. Furthermore, we present an innovative tuning-free agent collaboration framework for the biomedical terminology normalization task in social media. By leveraging the internal knowledge and the reasoning capabilities of LLM, our framework conducts more sophisticated recall, ranking and re-ranking processes with the collaboration of different LLM agents. Experimental results across multiple datasets indicate that our approach exhibits competitive performance. We release our code and data on the github repository JOHNNY-fans/RankNorm.

1 Introduction

Biomedical Terminology Normalization is a basic research task in clinical natural language processing, linking non-standard mentions extracted from social media or clinical texts to normalized terms in a standard termbase, e.g., UMLS, MedDRA, ICD, SNOMED CT, to find the standard terms that have the same semantics as them. (Ruch et al., 2008; Leaman et al., 2013; Leal et al., 2015; Luo et al., 2019; Lee and Uzuner, 2020). It plays a cornerstone role in clinical data analysis (Xu et al., 2024), Clinical decision support system (CDSS) (Papadopoulos et al., 2022), Diagnosis-Related Group system (DRGs) (Wang et al., 2020) and medical dialogue system (Xu et al., 2019).

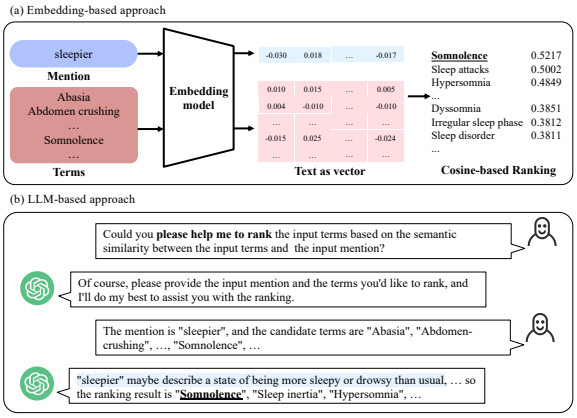

Figure 1: Comparison of Embedding-based Approach and LLM-based approach for Terminology Normalization Tasks.

Mainstream approaches typically employ the "recall and rerank" framework to accomplish this task. This involves initially recalling some candidates from the standard database and re-ranking them more precisely. Due to the success of the pre-trained language model BERT (Kenton and Toutanova, 2019), most of the recent work adopts the pretraining-finetuning paradigm, i.e., using a BERT-level pre-trained model as the backbone, subsequently fine-tune it on specific datasets (Miftahutdinov and Tutubalina, 2019; Xu et al., 2020; Liang et al., 2021). This means we need to completely retrain the model when the standard termbase

*Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10712–10726
January 19–24, 2025.

changes, which is not generalizable. Another bottleneck is that both mentions and terms in this task are short texts. Short text often contains insufficient information and introduces ambiguities, especially in the biomedical context, posing a considerable challenge.

However, new trends and solutions have emerged in the Large Language Models (LLMs) era. Advanced embedding models, considered foundational for computing semantic similarity and retrieval, such as instructor-xl (Su et al., 2022), BGE (Xiao et al., 2023), and OpenAI's Text Embeddings (OpenAI, 2022, 2024). These models are trained using effective methods and substantial supervised data, exhibiting superior performance. Meanwhile, very large language models appear to learn from the vast amount of data they process. They can perform tasks without gradient steps or fine-tuning, relying solely on task definitions and few-shot demonstrations provided in their contexts (Brown et al., 2020). This method, known as Language Prompting or simply "Prompting", has now become a new paradigm for accomplishing downstream tasks.

Therefore, we intend to leverage the LLM and explore new paradigm-based solutions based on the mainstream "Recall and Rank" framework for the terminology normalization task. In Figure 1, we provide a simple comparison chart of the traditional and LLM-based approaches.

To address the short-text challenge, we introduce "Knowledge Cards". They expand on the names of mentions or terms and provide descriptive information through knowledge distillation from LLMs. We introduce an LLM agent that uses search engines and knowledge bases to generate these expanded knowledge cards. Additionally, we propose a Knowledge-Enhanced Retrieval approach that employs an advanced embedding model, which considers both the name and the knowledge card during retrieval.

Meanwhile, we have discovered that ranking can also be achieved by reasoning using the LLM. For instance, RankGPT Sun et al. (2023) utilizes an LLM to rank documents effectively based on user queries. We propose a training-free LLM-based multi-agent collaboration framework to improve the performance, building on the "recall and rerank" framework. This framework is designed for the terminology normalization task and harnesses the capabilities of advanced embedding models and LLMs to enhance the entire process.

Specifically, we introduce a terminology expert agent that manages both the Knowledge-Enhanced Retrieval module as the rough recall module and the "Top-k Ranking" module to further refine the selection of candidate terms. Additionally, we aim to obtain conclusions from different professional perspectives and achieve more reasonable answers through ensemble learning. Therefore, we expand our system to include three additional agents: a clinical doctor agent, an outpatient doctor agent, and an internet doctor agent to conduct further detailed ranking. These agents collaborate in a multi-agent framework to perform detailed rankings.

As shown in Figure 2, the overall framework and our contributions can be summarized as follows:

- We design a training-free multi-agent collaboration framework for term normalization task that utilizes advanced embedding models and LLMs to acquire the candidate terms via Knowledge-Enhanced Retrieval and obtain the final standard terms through LLM-based ranking with demonstrations.
- We propose a knowledge enhancement approach that introduces an LLM agent to use search engines and knowledge bases to extend short medical texts into knowledge cards containing enhanced descriptive information and medical knowledge, which benefits the performance of recall phase.
- We developed a multi-agent collaborative recall and ranking workflow using prompt engineering techniques such as chain-of-thought and demonstration selection. After the terminology expert agents enhance the recall phase, the "Top-K Ranking" module inspired by the divide-and-conquer algorithm is used to further refine the list of candidate terms. In addition, by aggregating the ranking conclusions from different agents in the "Multi-Persona Re-ranking" module, we further improve the performance of the re-ranking phase.

2 Related Work

2.1 Biomedical Terminology Normalization

Biomedical term normalization (Leaman et al., 2013; Ji et al., 2020; Li et al., 2017) is one of the fundamental tasks within biomedical natural language processing and medical domain (Xu et al., 2024; Papadopoulos et al., 2022; Wang et al., 2020;

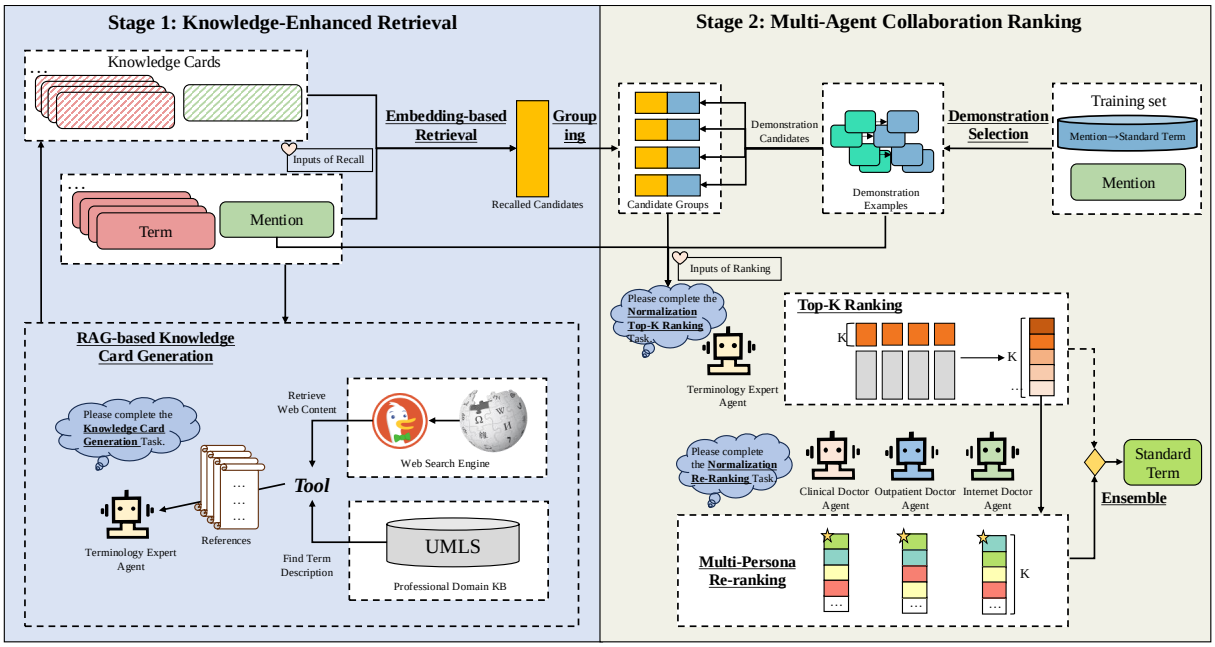

Figure 2: The proposed framework. The left side is the Knowledge-Enhanced Retrieval stage, and the right side shows the LLM-based Multi-Agent Collaboration Ranking flow.

Xu et al., 2019), aiming at finding standard terms for various clinical statements.

Early approaches for clinical term normalization relied on dictionary lookup (Lee et al., 2016) and heuristic string-matching techniques (Leal et al., 2015), both requiring significant manual effort. With advancements in Artificial Intelligence, Machine Learning, and Deep Learning methods have emerged (Savova et al., 2008; Sui et al., 2022; Zhou et al., 2021b; Ji et al., 2021; Zhou et al., 2021a).

Given the massive scale of knowledge bases, directly ranking the entire standard terminology set is challenging. A two-stage process "recall and rank" has thus become essential. For instance, Liang et al. (2021) proposed a framework based on "recall, rank, and fusion," including a model-based online negative sampling strategy in the recall stage. Similarly, Xu et al. (2020) developed an architecture featuring a BERT-based candidate generator and list-wise ranker. However, there have also been explorations for another paradigm, where Yuan et al. (2022) has proposed a new generative approach without candidate recall based on pre-training of the knowledge base and fine-tuning by fusing synonym information.

Recall modules can leverage traditional models like BM25, and TF-IDF, but vector-based semantic similarity is now mainstream. Ji et al. (2020) pioneered using BM25 scores for recall evaluation. Additionally, Liu et al. (2020) introduced the ABTSBM method for ICD-9-CM3 normalization, employing an N-gram algorithm to generate candidate terminologies. Niu et al. (2019) presented a multi-task character-level attentional network to learn character structure features, while Yan et al. (2020) suggested a generative sequence framework with prefix tree decoding to produce realistic medical procedure entities. Moreover, Lai et al. (2022) has attempted to use a sequence-to-sequence (seq2seq) model to generate descriptive information for entities to improve recall performance.

Ranking modules typically use scoring or classification models to identify the correct standard term from candidates. For example, Leaman et al. (2013) introduced a linear pair-wise model to rank standard terminologies based on vector similarity and negative sampling strategies. Additionally, several studies treat normalization as a classification task, such as Liu et al. (2020)'s BERT-based classifier and Ji et al. (2020)'s fine-tuning of existing BERT models. The latest work (Xu et al., 2023) is based on prompt-based learning to improve the accuracy of ranking by paying more attention to fine-grained information.

2.2 Leveraging Large Language Models

Pretrained language models (Radford et al., 2018; Kenton and Toutanova, 2019) have recently demonstrated significant improvements in various NLP tasks. Motivated by the scaling law, which suggests

Dataset	NAME	KC	RAG	HR@1	HR@5	HR@10	HR@20	HR@50	HR@100	HR@200
AskPatient	✓	✗	✗	66.35	87.22	92.33	95.42	97.69	99.11	99.46
	✓	✓	✗	66.38	85.03	90.08	94.34	97.15	98.59	99.12
	✓	✓	✓	**70.80**	**91.30**	**95.47**	**97.67**	**99.06**	**99.41**	**99.57**
TwADR-L	✓	✗	✗	35.39	61.67	68.26	76.17	84.37	89.00	93.55
	✓	✓	✗	38.26	62.23	71.13	77.86	85.63	89.98	94.74
	✓	✓	✓	**39.38**	**63.70**	**72.67**	**79.89**	**86.83**	**90.89**	**94.81**
SMM4H-17	✓	✗	✗	47.36	64.56	78.16	85.08	90.52	93.04	95.28
	✓	✓	✗	57.64	73.12	80.04	84.84	90.84	93.48	94.80
	✓	✓	✓	**57.68**	**78.20**	**83.60**	**87.92**	**93.52**	**94.80**	**95.72**

Table 1: The Knowledge-Enhanced Retrieval experiment result, where "NAME" denotes the names of mentions and terms be used in retrieval, "KC" denotes the knowledge cards be used in retrieval, "RAG" denotes the Retrieval Augmented Generation technique be used when generating knowledge cards, "HR@num" denotes the hit rate of candidate terms containing the correct answer, and "num" denotes the number of candidate terms recalled.

that increasing model size enhances capacity (Kaplan et al., 2020), researchers have scaled up model parameters (Ouyang et al., 2022), resulting in Large language models (LLMs) with unique capabilities for a variety of downstream tasks.

The concept of In-Context Learning (ICL) was formalized by GPT-3 (Brown et al., 2020), which showed that LLMs could generate expected outputs by completing prompts based on natural language instructions and task demonstrations, without further training (Zhao et al., 2023). For instance, Nori et al. (2023) studied how different prompting techniques, such as chain-of-thought and kNN examples, enhance LLM performance in medicine. RankGPT(Sun et al., 2023) explored using large models for document ranking and introduced new paradigms for this task.

Retrieval-Augmented Generation (RAG) is another crucial LLM technique (Lewis et al., 2020; Gao et al., 2023; Asai et al., 2023) that improves response accuracy by retrieving relevant reference information and reducing hallucinations (Tonmoy et al., 2024). Additionally, LLM agents, powered by advanced language models, are autonomous systems (Wang et al., 2024; Guo et al., 2024; Zhao et al., 2024) designed to interact, make decisions, and perform tasks across various domains. For example, multi-agent debate systems (Chan et al., 2023) have been used for detailed, automated performance evaluations.

3 Method

We present a comprehensive overview of our solution. It is a training-free multi-agent collaboration framework based on LLM and comprises two primary stages. The "Knowledge-Enhanced Retrieval" stage generates knowledge cards (KCs) using an agent and recalls high-quality candidate terms. The "Multi-Agent Collaboration Ranking" stage includes the "Top-K Ranking" and the "Multi-Persona Re-ranking" modules, which minimize the range of candidate terms and find the optimal standard term through multi-agent collaboration. Specific framework details are displayed in Figure 2.

3.1 Knowledge-Enhanced Retrieval

3.1.1 RAG-based KC Generation

This step focuses on generating knowledge cards using advanced LLMs. The knowledge is then explicitly employed to enhance the semantics of mentions and terms.

We begin by introducing a terminology expert agent and designing a task for generating Knowledge Cards, guided by a carefully crafted prompt. This agent is specifically designed to extract terminology-related knowledge, with the objective of producing types, descriptions, explanations, and meanings for input terms, and generating knowledge cards in specified formats. To enhance the quality of these cards, we also integrate a search engine and a specialized terminology database to provide reliable references. The prompt includes chain-of-thought instructions that guide the LLM to analyze the input terms and produce the corresponding descriptive content for the knowledge cards. The detailed prompt content is shown in Figure A1.

3.1.2 Embedding-based Retrieval

We employ "Embedding + Knowledge Card" as our final retrieval strategy, whereby both the term name and its expanded information via knowledge cards are encoded as vectors by a text embedding

model. These vectors are then concatenated to form a knowledge-enhanced representation for the term, followed by the similarity score computation. The algorithm flow for this approach is presented in Algorithm 1. The vector retrieval engine embeds every standard term t in the standard terminology base T and its corresponding knowledge card K_t, and concatenates the term name embedding and knowledge card embedding into a vector $\hat{\mathbf{t}} \in \hat{\mathbf{T}}$. Meanwhile, the mention m, and its associated knowledge card K_m are encoded as $\hat{\mathbf{m}}$ through the same operation. The cosine similarity between the mention m and each standard term t in the entire terminology base is measured, and some standard terms with high similarity to the mention term m are selected and added to the candidate term set C.

Algorithm 1: Algorithm of Knowledge-Enhanced Retrieval

Input: mention m
standard terminology base T
knowledge cards $K_m, K_t \in K_T$
Output: standard term s of mention m
candidate terms C of mention m

foreach t *in* T **do**
  embedToVecWithKC$(t, K_t) \rightarrow \hat{\mathbf{t}} \in \hat{\mathbf{T}}$;
end
embedToVecWithKC$(m, K_m) \rightarrow \hat{\mathbf{m}}$;
searchSimTerm$(m, T, \hat{\mathbf{m}}, \hat{\mathbf{T}}) \rightarrow C$;
searchMaxSimTerm$(m, T, \hat{\mathbf{m}}, \hat{\mathbf{T}}) \rightarrow s$;

3.2 Multi-Agent Collaboration Ranking

3.2.1 Agent Initialization

In addition to the terminology expert agent, we introduced three additional agents: a clinical doctor agent, an internet doctor agent, and an outpatient doctor agent. During the ranking phase, these agents are assigned specific roles through system prompts, guiding the LLM to focus on different biomedical perspectives. By incorporating insights from various medical stages where the mention might occur, we aim to achieve more comprehensive reasoning. The content prompts they use to complete tasks are provided in Appendix A.

3.2.2 Prompts and Data Preparation

The task definition for the LLM is to rank a given candidate terms list and then output the most relevant terms with the input mentions.

Chain-of-thought instructions are introduced for the agent to perform step-by-step reasoning to improve the task accuracy, including learning the pattern from the given demonstrations, analyzing the meaning of the input mention, giving the basis, and then outputting the ranking result.

Output format is a necessary component for achieving a more automated and controllable algorithm. We format the agent's output in JSON to facilitate the extraction of the conclusions and content we need.

Input of Ranking consists of a mention and some candidate terms from memory. Heuristically, we group the candidates so that the number of elements in each group remains at a suitable level. Moreover, discarding sequential grouping, Here's a polished version of your sentence:

We use a **balanced grouping** strategy, where candidates C are randomly assigned to groups G based on their cosine scores. After the initial ranking, candidates are sequentially and randomly placed into the group with the fewest members. This approach guarantees consistency in the number and distribution of each group. Since the agent can access k-NN demonstration examples from memory, we add the standard terms from these examples as expanded candidates to each group and obtain supplemented $\tilde{G}$.

Algorithm 2: Algorithm of Demonstration Selection

Input: given mention m
training dataset $(d, t) \in D$
knowledge cards $K_m, K_d \in K_D$
Output: k-NN demonstration examples E of input mention m

foreach d,_ *in* D **do**
  embedToVecWithKC$(d, K_d) \rightarrow \hat{\mathbf{d}} \in \hat{\mathbf{D}}$
end
embedToVecWithKC$(m, K_m) \rightarrow \hat{\mathbf{m}}$;
searchSimTrain$(m, D, \hat{\mathbf{m}}, \hat{\mathbf{D}}) \rightarrow E$;

Demonstration Selection. Demonstrations are highly effective for enabling LLMs to perform in-context learning and complete tasks successfully (Ye et al., 2023). To leverage this, we developed a demonstration selection module that identifies high-quality examples from the training data using the k-nearest neighbors algorithm. By applying knowledge-enhanced retrieval between the input mention and those in the training data, we

identify the most appropriate demonstration examples E for a given input mention m from the training set D. The detailed algorithm flow is presented in Algorithm 2.

3.2.3 Ranking and Re-ranking

The ranking process begins with the term expert agent performing a "Top-K Ranking" task, aimed at refining the candidate term list down to a manageable number K. Following this, the "Multi-Persona Re-ranking" module reorders these terms, allowing the three medical persona agents to select the most appropriate standard term corresponding to the mention. The detailed algorithm flow is illustrated in Algorithm 3.

Algorithm 3: LLM-based Ranking Algorithm

Input: mention m, candidates C, Term Expert A_t, Clinical Doctor A_c, Outpatient Doctor A_o, Internet Doctor A_i
Output: normalized result s

candidateGrouping$(C) \rightarrow G$;
addDemocandidate$(G) \rightarrow \tilde{G}$;
foreach $\tilde{g} \in \tilde{G}$ **do**
  A_t: topkRanking$(m, \tilde{g}) \rightarrow V$;
end
A_t: topkRanking$(m, V) \rightarrow \tilde{C}$;
foreach $A \in \{A_c, A_o, A_i\}$ **do**
  A: re-ranking$(m, \tilde{C}) \rightarrow R$;
end
ensemble$(R, \tilde{C}) \rightarrow s$;

Top-K Ranking. Using a divide-and-conquer approach, the term expert agent A_t identifies the top K terms v from each group, merges them, and then selects the top K terms from the newly combined candidate set V. The final output is a streamlined set $\tilde{C}$ containing only a few candidate terms.

Multi-Persona Re-ranking. To identify the most suitable term from a refined set of candidate terms $\tilde{C}$ as the standard term corresponding to the mention, we adjusted the ranking prompt by removing the constraint of selecting K terms. Instead, we focused on filtering relevant terms and re-ranking them. Three medical persona agents, A_c, A_o, and A_i, each provide their own assessment, and the final term s is determined through ensemble learning. This process uses a voting mechanism that prioritizes average rankings and term frequency. In case of a tie, the terminology expert's ranking is used as the deciding factor.

4 Experiment

4.1 Datasets

Following the complete setting of (Xu et al., 2020), we conduct our experiment on three datasets, AskPatient (Limsopatham and Collier, 2016), TwADR-L (Limsopatham and Collier, 2016), and SMM4H-17 (Sarker et al., 2018).

AskAPatient: The AskAPatient dataset[1] comprises 17,324 annotations of adverse drug reactions (ADRs) sourced from blog entries. These annotations are linked to 1,036 medical concepts, encompassing 22 semantic categories derived from a segment of the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and the Australian Medicines Terminology (AMT). Our methodology aligns with the 10-fold cross-validation framework utilized in the study by (Limsopatham and Collier, 2016), which presents 10 separate training, validation, and testing divisions.

TwADR-L: Encompassing 5,074 expressions of ADRs extracted from social media platforms, the TwADR-L dataset[1] aligns these expressions with 2,220 concepts from the Medical Dictionary for Regulatory Activities (MedDRA), spanning 18 semantic categories. Our approach also adheres to the 10-fold cross-validation model established by (Limsopatham and Collier, 2016).

SMM4H-17: SMM4H-17[2] includes 9,149 hand-picked ADR expressions from Twitter posts. These expressions are linked to 22,500 concepts, incorporating 61 semantic types from MedDRA Preferred Terms (PTs). The training dataset includes 5,319 expressions from the publicly released set while reserving the 2,500 expressions from the original test set for evaluation purposes.

4.2 Implementation Details

For the Knowledge-Enhanced Retrieval, we use text-embedding-3-large (OpenAI, 2024) as Embedding model, and set the number of candidates as 200. The search engine tool for the term expert agent is DuckDuckGo (DuckDuckGo, 2008), and the additional terminology knowledge comes from the UMLS2023ab version (Bodenreider, 2004).

[1] https://zenodo.org/records/55013
[2] https://data.mendeley.com/datasets/rxwfb3tysd/1

Method	AskPatient	TwADR-L	SMM4H-17
Unsupervised methods			
TF-IDF	55.47	22.93	22.16
BM25	55.46	23.00	24.20
text-embedding-ada-002 (OpenAI, 2022)	69.31	38.68	55.96
text-embedding-3-large (OpenAI, 2024)	66.35	35.39	47.36
∗ text-embedding-ada-002 + KnowledgeCard	**71.74**	39.23	56.36
∗ text-embedding-3-large + KnowledgeCard	70.80	**39.38**	**57.68**
Supervised methods			
WordCNN (Limsopatham and Collier, 2016)	81.41	44.78	-
WordGRU+Attend+TF-IDF (Tutubalina et al., 2018)	85.71	-	-
BERT+TF-IDF (Miftahutdinov and Tutubalina, 2019)	-	-	89.64
CharCNN + Attend+MT (Niu et al., 2019)	84.65	46.46	-
CharLSTM + WordLSTM (Han et al., 2017)	-	-	87.20
LR + MeanEmbedding (Belousov et al., 2017)	-	-	87.70
BERT Multiclass	86.35	42.56	86.52
† BERT + BERT-rank + ST-reg (Xu et al., 2020)	87.46	47.02	88.24
† BERT Multiclass (BioBERT v1.2)	**89.74**	45.89	89.40
Generative + PT + FT (Yuan et al., 2022)	89.30	-	-
∗ Ours (GPT3.5)	88.54	**52.28**	**90.84**

Table 2: Comparison of different approaches for biomedical terminology normalization. The evaluation metric is accuracy, "†" denotes that the method uses biomedical PLMs, and "∗" denotes our proposed approach or module.

We selected GPT-3.5-turbo-1106 (OpenAI, 2023) as the foundational LLM for the agents. In the demonstration selection module, we used the 10 nearest-neighbor examples for each mention. Following prior work (Liang et al., 2021) and considering the suitability, during the candidate grouping step, we divided the 200 candidates into 4 default groups. In the "Top-K Ranking" module, we selected the top 10 terms as input candidates for the re-ranking module. For LLM inference, the temperature was set to 0, with a seed value of 42.

4.3 Evaluation and Analysis

For Knowledge-Enhanced Retrieval, we conducted experiments to prove the importance of the knowledge card for the embedding-based retrieval, and the evaluation metric is the Hit Rate, denoted as "HR@n", where n represents the number of candidate terms retrieved, which measures the proportion of the top n results that contain the correct answer. Mathematically, HR@n is defined as:

$$\text{HR@n} = \frac{1}{|M|} \sum_{m \in M, c \in C} \mathbb{I}(m, c)$$

where C is the selected candidates list of input mention list M, and $\mathbb{I}$ is an indicator function that returns 1 if there is the correct term in c for a mention m, and 0 otherwise. The results are displayed in the Table 1. We also compared the effect of RAG on the quality of knowledge cards.

In the recall phase, results across all datasets indicate that using both mentions and the term name, along with the knowledge card, yields a higher hit rate compared to using only the term name. The introduction of knowledge cards enhances the retrieval process by incorporating additional information and context, helping refine the candidate set and improve the recall rate. Additionally, the use of RAG enhances performance by reducing inaccuracies and increasing the reliability of the information on the knowledge cards.

When considering our method as an unsupervised term normalization approach, as shown in the top half of Table 2, we focus only on the term with the highest score. Even then, the results with knowledge cards outperform those of traditional BM25 and TF-IDF models, as well as those of the use of advanced embedding models alone.

These improvements suggest that the introduction of knowledge cards significantly enhances the retrieval process by providing additional context and refining the embedded vectors to capture more specific semantics. This, in turn, improves the identification of semantically similar terms. The integration of RAG technology further stabilizes performance by reducing the effects of LLM hallucinations and knowledge ambiguity through the inclusion of reliable references. We also give a practical case in Appendix Figure A4 to demonstrate the impact of whether or not to perform RAG

Setting	Top-K Ranking (HR@10)	Multi-Persona Re-ranking (Acc)
Ours	97.36	90.84
w/o Knowledge-Enhanced Retrieval	96.20	90.64
w/o CoT Instructions	93.64	84.92
w/o Demonstration Examples	76.96	58.40
w/o Grouping	96.56	90.72
w/o Expanded Demonstration Candidates	93.04	87.88
w/o Multi-Persona Re-ranking	-	89.84
w/o Ensemble(A_c/A_o/A_i)	-	90.76 / 90.56 / 90.80

Table 3: Ablation Experiments for Multi-Agent Collaboration Ranking Modules.

on the quality of the generated knowledge cards in a more intuitive way.

However, we also observed that advanced embedding models perform exceptionally well. When a larger number of candidates (e.g. 200) are considered, the difference between using or not using knowledge cards becomes less pronounced, suggesting that these advanced models are learning richer semantics from extensive data.

For Multi-Agent Collaboration Ranking, while we proposed a training-free terminology normalization framework, we still leverage demonstration examples from the training set to enable the LLM to perform the task through in-context learning. Therefore, we compare our approach to supervised methods using the same datasets. Additionally, given the strong medical capabilities of large models, we also include comparisons with biomedical pre-trained language models.

The evaluation metric for the final normalization result is accuracy, defined as the percentage of samples where the selected term exactly matches the correct normalized term. The results are presented in the bottom half of Table 2. To assess the contribution of each module to the final outcome, we performed ablation experiments on the SMM4H-17 dataset, which features the most extensive standard terminology base and the greatest variety of semantic types. The detailed findings are shown in Table 3. We also validate the performance impact of the multi-persona ranking module on all three datasets, as shown in Appendix Table 5. Furthermore, to validate the generalizability of our framework, we performed comparative experiments using four different LLM foundations, as shown in Table 4.

Our proposed method significantly outperforms models fine-tuned on individual datasets, which were designed to provide demonstration examples for in-context learning without requiring parameter fine-tuning. The ablation experiments confirm that each of our proposed modules contributes positively to the final performance. Key contributors include high-quality demonstrations, specifically designed CoT instructions, the expanded candidate terms supplemented by demonstration examples, and the Multi-Persona module. These results underscore the importance of supervised signals in guiding LLM agents. The introduction of medical persona agents enhances accuracy, demonstrating that agents with different personas indeed produce distinct outputs when completing tasks. This variation in analysis perspectives can also be observed in the output inference process.

As the context lengths supported by advanced LLMs have increased and their reasoning capabilities have improved, grouping and ensemble strategies have become minor yet effective enhancements to the robustness. Additionally, the frame-

LLM	A_t	A_c	A_o	A_i	Ensemble
llama3-8b	86.24	85.64	88.52	85.88	87.16
GPT-3.5 1106	89.84	90.76	90.56	90.80	90.84
llama3-70b	89.56	89.96	90.52	90.64	90.88
GPT-4o 0513	**91.84**	**91.68**	**91.76**	**92.04**	**92.12**

Table 4: Comparison of results on SMM4H-17 using different foundational LLMs for the agents, where A_t represents the Term Expert Agent, A_c the Clinical Doctor Agent (results from Top-K ranking), A_o the Outpatient Doctor Agent, and A_i the Internet Doctor Agent.

Dataset	A_t	A_c	A_o	A_i	Ensemble
AskApatient	87.32	87.51	87.83	88.19	**88.54**
TwADR-L	52.82	50.54	**53.19**	51.13	52.28
SMM4H-17	89.84	90.76	90.56	90.80	**90.84**

Table 5: Comparison of results on different datasets using GPT-3.5 as foundational LLM for the agents, where A_t represents the Term Expert Agent, A_c the Clinical Doctor Agent (results from Top-K ranking), A_o the Outpatient Doctor Agent, and A_i the Internet Doctor Agent.

work demonstrates strong generalization capabilities and future potential, with results improving as LLM capabilities advance as shown in Table 4.

5 Conclusion

In this paper, we propose a training-free, LLM-based multi-agent collaboration framework for biomedical normalization tasks in social media, featuring two key components: Knowledge-Enhanced Retrieval and Multi-Agent Collaboration Ranking.

For Knowledge-Enhanced Retrieval, we tackle the ambiguity of short texts by expanding mentions and terms using a terminology expert agent. This agent leverages a search engine tool in combination with UMLS to generate knowledge cards, resulting in more informative vector representations during retrieval. This approach improves accuracy and hit rates across various datasets without the need for additional training of a supervised recall model. The agent's use of a tool is guided by RAG techniques to obtain high-quality knowledge cards and minimize hallucinations.

In Multi-Agent Collaboration Ranking, we leverage the reasoning capabilities of LLM agents to enhance performance in ranking candidate terms. The terminology expert agent conducts a Top K ranking task using a comprehensive prompt to narrow down the candidate terms. We then refine the prompt and introduce three medical persona agents: a clinical doctor, an outpatient doctor, and an internet doctor. These agents collaborate to achieve more accurate term normalization results.

Extensive experiments on this framework show that all our proposed modules are effective. Notably, our untrained framework achieves performance on par with state-of-the-art methods.

6 Limitations

First, the knowledge cards generated by the terminology expert agent provide only a vague description of the mentions or terms rather than precise, structured knowledge, even with RAG and a specialized knowledge base. Future research can explore this interaction with LLM to distill more fine-grained knowledge.

Secondly, we found that some model outputs failed the format check during the ranking process using the large model. This might indicate that the model could not find the current candidates' answers. We addressed this issue by choosing a more relaxed temperature setting, such as 0.5, which might have led to incorrect answers. However, using dynamic candidates could be a better solution. This also suggests that multiple rounds of interaction with the LLM could further improve task accuracy. Moreover, we cannot entirely eliminate randomness of non-privatised deployment LLMs such as GPT-3.5 and GPT4o even with the temperature set to 0 and fixed seeds provided.

Finally, we propose a training-free multi-agent collaboration framework for normalization tasks in social media. Although our approach eliminates the need for repetitive training, it does involve the computational costs associated with LLM inference. Despite this, exploring LLM methods remains valuable due to their strong multitasking capabilities, as LLMs excel at completing various tasks based on instructions. Consequently, it is practical to design different modules for different tasks that share the same LLM, a common practice in industry. Additionally, we conducted experiments on a comprehensive and feasible benchmark containing multiple term types. However, most mentions were sourced from social media, and the lack of clinically extracted mentions warrants further investigation.

7 Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback. Our thanks also go to the Chairs and the organizing staff for their dedicated efforts in facilitating this work. This work was supported by the National Key Research and Development Program of China (Grant 2021YFC2701801).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Maksim Belousov, William G Dixon, and Goran Nenadic. 2017. Using an ensemble of linear and deep learning models in the smm4h 2017 medical concept normalisation task. In *SMM4H@ AMIA*, pages 54–58.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

DuckDuckGo. 2008. Duckduckgo.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Sifei Han, Tung Tran, Anthony Rios, and Ramakanth Kavuluru. 2017. Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter. In *SMM4H@ AMIA*, pages 49–53.

Zongcheng Ji, Qiang Wei, and Hua Xu. 2020. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269.

Zongcheng Ji, Tian Xia, Mei Han, and Jing Xiao. 2021. A neural transition-based joint model for disease named entity recognition and normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2819–2827, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2022. Improving candidate retrieval with entity profile generation for Wikidata entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3696–3711, Dublin, Ireland. Association for Computational Linguistics.

André Leal, Bruno Martins, and Francisco M Couto. 2015. Ulisboa: Recognition and normalization of medical concepts. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 406–411.

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

Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2016. Audis: an automatic crf-enhanced disease normalization in biomedical text. *Database*, 2016:baw091.

Kahyun Lee and Özlem Uzuner. 2020. Normalizing adverse events using recurrent neural networks with attention. *AMIA Summits on Translational Science Proceedings*, 2020:345.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18:79–86.

Ming Liang, Kui Xue, Qi Ye, and Tong Ruan. 2021. A combined recall and rank framework with online negative sampling for chinese procedure terminology normalization. *Bioinformatics*, 37(20):3610–3617.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

Yijia Liu, Bin Ji, Jie Yu, Yusong Tan, Jun Ma, and Qingbo Wu. 2020. An advanced icd-9 terminology standardization method based on bert and text similarity. In *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1868–1879. Springer.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: a comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92:103132.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, 49:1239–1256.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2022. New and improved embedding model. Technical report.

OpenAI. 2023. New models and developer products announced at devday. Technical report.

OpenAI. 2024. New embedding models and api updates. Technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Petros Papadopoulos, Mario Soflano, Yaelle Chaudy, Wilson Adejo, and Thomas M Connolly. 2022. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology*, 12(4):713–727.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Patrick Ruch, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. 2008. Automatic medical encoding with snomed categories. In *BMC medical informatics and decision making*, volume 8, pages 1–8. BioMed Central.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Guergana K Savova, Anni R Coden, Igor L Sominsky, Rie Johnson, Philip V Ogren, Piet C De Groen, and Christopher G Chute. 2008. Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41(6):1088–1100.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Xuhui Sui, Kehui Song, Baohang Zhou, Ying Zhang, and Xiaojie Yuan. 2022. A multi-task learning framework for chinese medical procedure entity normalization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8337–8341. IEEE.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, et al. 2020. A study of entity-linking methods for normalizing chinese diagnosis and procedure terms to icd codes. *Journal of biomedical informatics*, 105:103418.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Dongfang Xu, Zeyu Zhang, and Steven Bethard. 2020. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8452–8464.

Hua Xu, Dina Demner Fushman, Na Hong, and Kalpana Raja. 2024. Medical concept normalization. In *Natural Language Processing in Biomedicine: A Practical Guide*, pages 137–164. Springer.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving biomedical entity linking with cross-entity interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13869–13877.

Jinghui Yan, Yining Wang, Lu Xiang, Yu Zhou, and Chengqing Zong. 2020. A knowledge-driven generative model for multi-implication chinese medical procedure entity normalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1490–1499.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Baohang Zhou, Xiangrui Cai, Ying Zhang, Wenya Guo, and Xiaojie Yuan. 2021a. Mtaal: multi-task adversarial active learning for medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14586–14593.

Baohang Zhou, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. 2021b. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6214–6224, Online. Association for Computational Linguistics.

A Supplementary materials

From Figure A1 to Figure A3, we present the system prompts initialized by the four specific agents: the terminology expert agent, the clinical doctor agent, the outpatient doctor agent, and the internet doctor agent. These figures also detail the prompts for three key tasks: Knowledge Card Generation, Top-K Ranking, and Multi-Persona Re-ranking.

In Figure A4, we show a specific case to demonstrate more visually the impact of whether or not to perform RAG on the quality of the generated knowledge cards.

system:
You are a Terminology Expert Agent, assisting in the management and standardization of terminology across various fields. They help ensure consistency and accuracy in the use of terms by analyzing data, researching terminology usage, and coordinating with subject matter experts. This role involves the creation and maintenance of glossaries, dictionaries, and knowledge bases to support clear and effective communication.

user:
Please write a short and clear knowledge card for the input term based on the provided references and your knowledge of this term.

The knowledge card of this term should mention:
(1) the type of this term (e.g., disease, structure, drug, etc.).
(2) the description, explanation, and meaning of this term.

Output in the following JSON format:

```
{
        "term": "xxx",
        "knowledge card": "xxx",
}
```

References:

{reference}

Input term: {term}

Knowledge Card:

Figure A1: The specific prompt for knowledge card generation, used in the knowledge distillation step of the Knowledge-Enhanced Retrieval.

system:
You are a Terminology Expert Agent, assisting in the management and standardization of terminology across various fields. They help ensure consistency and accuracy in the use of terms by analyzing data, researching terminology usage, and coordinating with subject matter experts. This role involves the creation and maintenance of glossaries, dictionaries, and knowledge bases to support clear and effective communication. You are asked to rank the input terms based on their semantic similarity to the meaning of the input mention. The more semantically similar, the higher the ranking. Note that mentions are often written in an informal way and terms are written in a relatively formal way.

user:
I will provide you with several candidate terms, your task is to output the most relevant topk terms after your ranking, in this task k is set to 10.

I have also provided some examples of mention with its corresponding standard term annotated by experts and some special cases.
[Example]:
{example}

[Two Special Cases]:
1. If the mention input is the same as a term, this term should be put at the top of the ranking topk_list.
2. If the mention in the examples are the same as the input mention, the corresponding term in the example should be put at the top of the ranking topk_list.

Follow the steps below for step-by-step reasoning:
1. Summarize the correspondence between mentions and terms from examples as the ranking reference.
2. Analyze the meaning of the input mention or the state it describes.
3. Give the basis for this ranking.
4. Rank the candidate list and select the topk terms according to the task objectives.
5. Final check: Determine if there are any special cases I mentioned before, if so, correct the ranking result.

Please follow the above reasoning steps for the task input and then output the reasoning process and and the selected topk terms in the follow JSON format::
{
"reasoning_process": 1.xxx, 2.xxx, ...,
”ranking_result": [term1,term2,...] ,
}

[Task Input]:
mention:
{mention}

List of candidate terms:
{cand}

[Task Output]:

Figure A2: The specific prompt for “Top-K Ranking” task.

system:

- You are a Clinical Doctor Agent, assisting in managing patient diagnoses and treatment processes. You may handle data analysis, medical records management, and patient follow-ups, ensuring that the clinician can focus on delivering high-quality healthcare.
- You are an Outpatient Doctor Agent, helping manage daily outpatient operations, including appointment scheduling, patient reception, and basic medical examinations. You ensure that the outpatient process runs smoothly, allowing the doctor to efficiently see more patients.
- You are an Agent of Internet Doctor, supporting online healthcare services by assisting with remote consultations, patient inquiries, and health management. You may also help schedule virtual meetings, manage online patient records, and provide technical support.

You are asked to rank the input terms based on their semantic similarity to the meaning of the input mention. The more semantically similar, the higher the ranking. Note that mentions are often written in an informal way and terms are written in a relatively formal way.

user:
I will provide you with several candidates, your task is to find the term that is closest to its meaning or to the state it describes for the input mention as its standard term from the input candidates, and then re-rank candidate list according to the task objectives.

I have also provided some examples of mention with its corresponding standard term annotated by experts and some special cases.
[Example]:
{example}

[Three Special cases]:
1. If the mention input is exactly the same as one term, this term should be put at the top of the ranking result list.
2. If the mention in the examples is exactly the same as the input mention, the corresponding term in the example should be put at the top of the ranking result list.
3. If more than one standard terms are selected the annotation preferences and habits of the experts should be considered in ranking.

Follow the steps below to reason about the task input step by step, giving details of the process at each step::
1. Summarize the correspondence between mentions and terms and the annotation preferences and habits of experts from examples as the ranking reference.
2. Analyze the meaning of the input mention or the state it describes.
3. Give the basis for this ranking.
4. Rank the selected terms according to the task objectives.
5. Final check: Determine if there are any special cases I mentioned before, if so, correct the ranking result.

Please follow the above reasoning steps for the task input and then output the reasoning process and ranking result in format as follows, note that the ranking result is in JSON format::

```
{
        "reasoning_process": 1.xxx, 2.xxx, ...,
        "ranking_result": [term1, term2, …]
}
```

[Task Input]:
mention:
{mention}

List of candidate terms:
{cand}

[Task Output]:

Figure A3: The specific prompt for “Multi-Persona Re-ranking” module.

Term: "Abdominoplasty".

Knowledge Card without RAG: Type: Procedure, Description: Abdominoplasty, also known as a tummy tuck, is a surgical procedure to remove excess skin and fat from the abdomen and tighten the abdominal muscles. It is often performed after significant weight loss or pregnancy to improve the appearance of the abdomen.

Knowledge Card with RAG: Abdominoplasty is a surgical procedure that involves the removal of excess abdominal skin and fat, as well as the tightening of the abdominal muscles. It is also known as a tummy tuck. This procedure may also include the re-creation of the umbilicus and the removal of intra-abdominal fat. Abdominoplasty is typically performed for cosmetic reasons to achieve a flatter and more toned abdomen.

Figure A4: A specific case to demonstrate more visually the impact of whether or not to perform RAG on the quality of the generated knowledge cards.