A Novel LLM-based Framework for Biomedical Terminology Normalization via Multi-Agent Collaboration

Anonymous ACL submission

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 enhancement 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 biomedical terminology normalization agent collaboration framework. By leveraging the reasoning capabilities of LLM, our framework conducts more sophisticated ranking and re-ranking processes with the collaboration of different LLM agents. Experimental results across multiple datasets indicate that our approach exhibits competitive performance.

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).

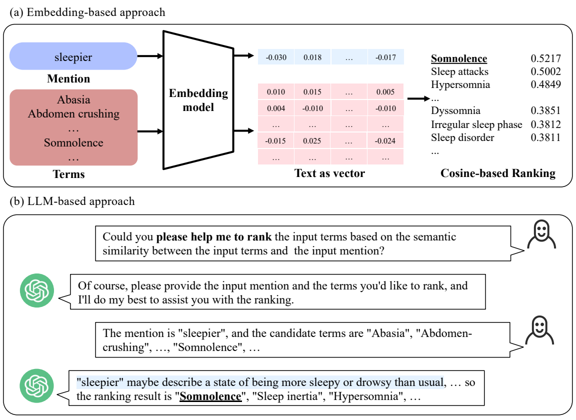

Figure 1: Comparison of Embedding-based Approach and LLM-based approach for Terminology Normalization Tasks.

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 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, include examples 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 elaborate on a format for knowledge acquisition called a "knowledge card". This format utilizes knowledge and expands on the names of mentions or terms through knowledge distillation from LLM. 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 terminology normalization that utilizes advanced embedding models and LLMs to acquire the candidate terms via Knowledge-Enhanced Retrieval and obtain the final standard terms through ranking with demonstration and chain-of-thought using an LLM.

- We propose a knowledge expansion 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.

- We employ prompt engineering techniques such as chain-of-thought instructions and demonstration selection to develop a workflow for ranking with multi-agent collaboration. Utilizing the Divide-and-Conquer algorithm's concept, the "Top-K Ranking" module further refines the list of candidate terms. Additionally, by aggregating the ranking conclusions of different agents, we further improve the performance of the re-ranking stage.

2 Related Work

2.1 Biomedical Terminology Normalization

Biomedical term normalization is one of the fundamental tasks within biomedical natural language processing (Leaman et al., 2013; Ji et al., 2020; Li et al., 2017), aiming at finding standard terms for various clinical statements.

Early approaches for clinical term normalization involve using dictionaries for lookup (Lee et al., 2016) or employing heuristic search methods based on string matching (Leal et al., 2015), which incurred significant manual effort. With the advancement of Artificial Intelligence, methods such as Machine Learning and Deep Learning emerge (Savova et al., 2008; Sui et al., 2022; Zhou et al., 2021b; Ji et al., 2021; Zhou et al., 2021a).

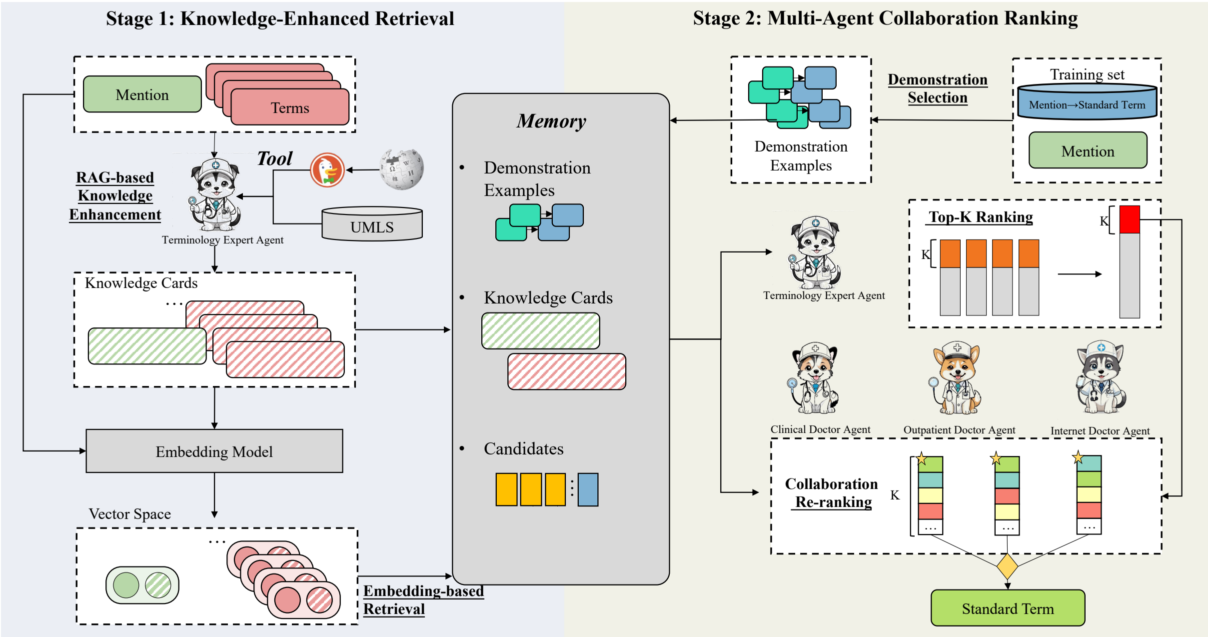

Figure 2: The proposed framework. The left side is the Knowledge-Enhanced Retrieval stage, and the right side shows the LLM-based Multi-Agent Collaboration Ranking flow.

Due to the massive scale of the knowledge base, it becomes challenging to rank the entire standard terminology base directly. It is vital to recall some semantically related candidate terms for subsequent ranking. Therefore, the two-stage clinical term normalization tasks consist of two main steps: recall and rank. For instance, Liang et al. (2021) proposed a framework based on "recall, rank, and fusion," and introduced a model-based online negative sampling strategy in the recall stage. Xu et al. (2020) also proposed an architecture that includes a candidate generator and a list-wise ranker based on BERT.

The recall module can be traditional models such as Elastic Search, BM25, and TF-IDF, while vector-based text semantic similarity has become mainstream. Ji et al. (2020) was the first to use the BM25 scores as the recall evaluation. Liu et al. (2020) provided an ABTSBM method for ICD-9-CM3 terminology normalization. The N-gram algorithm was applied to generate a standard candidate terminology set. Niu et al. (2019) presented a multi-task character-level attentional network that learned character structure features. Yan et al. (2020) suggested a generative sequence framework to generate all the corresponding candidate medical procedure entities directly and adopt prefix tree decoding to avoid producing unrealistic results.

The ranking module is usually a scoring or classification model incorporating various features to find the standard term corresponding to a few candidates' mentions. For example, Leaman et al. (2013) proposed a linear pair-wise model for representing medical terms, ranking standard terminologies based on the similarity between vectors, and devising strategies for choosing negative samples in the training process. In addition, many studies regard normalization tasks as a classification problem. Liu et al. (2020) use the BERT-based classification model to classify the correct standard terminology. Ji et al. (2020) fine-tuned the existing BERT models as well.

2.2 Leveraging Large Language Models

Recently, pretrained language models (Radford et al., 2018; Kenton and Toutanova, 2019) have shown promising improvements across many NLP tasks. Motivated by the finding that model scaling enhances the model capacity (Kaplan et al., 2020), researchers have further explored the scaling effect by scaling up the parameters to a larger size (Ouyang et al., 2022). With parameter scaling, LLMs exhibit unique and powerful abilities that enable multiple ways to leverage LLMs for accomplishing downstream tasks.

The concept of In-Context Learning (ICL) was rigorously introduced by GPT-3 (Brown et al., 2020). This framework posits that once the LLM is given natural language instructions and multiple task demonstrations, it can generate the expected

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.

output of a test instance by completing the word order of the input text (prompt) without additional training or gradient updates (Zhao et al., 2023). For instance, designing appropriate prompts makes it possible to leverage LLMs for knowledge acquisition.Nori et al. (2023) examines the impact of various prompting techniques on LLM performance in medicine, including chain-of-thought, kNN demonstration examples, and model output ensemble, which enhance the specialist capabilities of LLMs. RankGPT Sun et al. (2023) explores using large models to solve document ranking issues and investigate new paradigms for this task.

Retrieval-Augmented Generation (RAG) represents another pivotal and effective development of LLM technique (Lewis et al., 2020; Gao et al., 2023; Asai et al., 2023) that enhances the accuracy and expertise of large model responses. It retrieves relevant reference information related to the user's query and passes it to the LLM, thereby mitigating the problem of hallucination (Tonmoy et al., 2024).

Besides these, LLM agents are autonomous systems (Wang et al., 2024; Guo et al., 2024; Zhao et al., 2024) powered by advanced language models. These agents are assigned different roles and use their natural language processing capabilities to interact, make decisions, and perform tasks across various domains. For example, some researchers use multi-agent debate (Chan et al., 2023) to conduct detailed and automated performance evaluations of systems.

3 Method

We outline the comprehensiveness 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 using an agent and recalls high-quality candidate terms. The 'Multi-Agent Collaboration Ranking' stage includes the 'Top-K Ranking' module and the 'Collaboration Re-ranking' module, 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 Knowledge Enhancement

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

Initially, we introduce a terminology expert agent, construct a seed task, and manually craft a prompt. Specifically, we configure the agent as a terminology expert, define explicit task objectives and output formats for generating knowledge cards, and provide several reference dimensions. For instance, for a medicine term, the knowledge card contains pertinent details such as its definition description, active ingredient, content specification, dosage form, etc.

Meanwhile, we integrated a search engine as a tool for the agent and enhanced it with knowledge from a specialized terminology base to improve the quality of the generated knowledge cards. Additionally, the prompt includes some chain-of-thought instructions, which require the LLM to analyze the type of input mentions or terms, then refer to some dimensions given to determine the dimensions of this knowledge card, and finally output the specific content of the knowledge card. The specific prompt content is displayed 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 is encoded as $\hat{\mathbf{m}}$ through the same operation. The cosine similarities between the mention m and every standard term t in the entire terminology base are used as measures, some standard terms with high similarities to the mention m are selected and added to a candidate set C, and we select the term with the highest score as the standard term.

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 Memory for Multi-Agent

Memory is where multi-agent interactions converge. In this framework, memory includes the knowledge cards and recalled candidate terms generated in "Knowledge-Enhanced Retrieval" stage, as well as some demonstration examples related to input mentions.

Demonstration Selection. Demonstrations have proven very effective information for LLM to conduct in-context learning to accomplish tasks. so we designed a demonstration selection module to find higher-quality demonstration examples from the training data based on the k-nearest neighbors algorithm. By employing the knowledge-enhanced retrieval between the input mention and the mentions in training data, based on the input mention m, we find the appropriate demonstration examples E from the training set D. The specific algorithm flow is shown in Algorithm 3.

3.2.2 Agent Initialization

In addition to the terminology expert agent mentioned above, we introduced three more agents: a clinical doctor agent, an internet doctor agent, and an outpatient doctor agent. During the ranking phase, these agents are assigned different roles via system prompts to focus the capabilities of the LLM on various biomedical perspectives. They then process content prompts to complete tasks, including the following items. Specific prompt content we provide in the appendix A.

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

Demonstrations and knowledge Card of mention. Valid prior knowledge comes from memory that can help the agent find evidence and clues.

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 an unnecessary part to realize a more automated and controllable algorithm process, we let the agent's output in JSON format so that it is accessible to extract the conclusions and contents we want to obtain.

The task input 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, we use a balanced grouping strategy that randomly assigns candidates C to groups G according to their cosine scores. 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 $\tilde{G}$.

3.2.3 Ranking and Re-ranking

The specific ranking procedure lets the term expert agent complete a "Top-K Ranking" task. The objective here is to further refine the list of candidate terms, reducing their number to K, where K represents a relatively small value. Subsequently, the "Collaboration Re-ranking" module re-ranks these terms and selects the most suitable standard term corresponding to the mention by three medical persona agents. The specific algorithm flow is shown in Algorithm 2.

Algorithm 2: Algorithm of LLM-based Ranking

```
Input: given mention m,
candidate terms set C,
Term Expert Agent A_t,
Clinical Doctor Agent A_c,
Outpatient Doctor Agent A_o
Internet Doctor Agent A_i
Output: nomalized result s
candidateGrouping(C) → g ∈ G;
addDemocandidate(G) → g̃ ∈ G̃;
foreach g̃ in G̃ do
    A_t: topkRanking(m, g̃) → v ∈ V;
end
A_t: topkRanking(m, V) → C̃;
A_c: re-ranking(m, C̃) → r_c ∈ R;
A_o: re-ranking(m, C̃) → r_o ∈ R;
A_i: re-ranking(m, C̃) → r_i ∈ R;
ensemble(R) → s;
```

Top-K Ranking. Applying the divide-and-conquer algorithm, the term expert agent A_t finds the top K terms from each group, individually combines the answers, and then finds the top K terms v again from the new combination candidate set V. The final result is a set $\tilde{C}$ with only a few candidate terms.

Collaboration Re-ranking. To find the most appropriate term from a smaller set of candidate terms $\tilde{C}$ as the standard term corresponding to the mention, we delete the constraint of finding K terms in the ranking prompt and change it to filtering out the relevant terms and then re-ranking them. Each of the three medical persona agents, A_c, A_o, A_i, provides its own opinion, and the final answer s is then determined through ensemble learning.

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 handpicked 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 our Embedding model, and we 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).

For the Agents, we chose gpt-3.5-turbo-1106 (OpenAI, 2023) as the basic LLM. In the demonstration selection module, we chose 10

[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)	64.94	35.18	45.48
text-embedding-3-large (OpenAI, 2024)	69.31	38.68	55.92
$*$ text-embedding-ada-002 + KnowledgeCard	72.95	39.38	64.28
$*$ text-embedding-3-large + KnowledgeCard	**74.07**	**42.47**	**64.40**
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 + BERT-rank + ST-reg (Xu et al., 2020)	87.46	47.02	88.24
$*$ Ours	**88.54**	**52.28**	**90.84**

Table 2: Comparison of different approaches for biomedical terminology normalization. The evaluation metric is accuracy, and the "$*$" denotes our proposed approach or module.

nearest-neighbor examples for each mention. In the candidates grouping step, we divided the 200 candidates into 4 groups by default, and in the "Top-K Ranking" module, we finally chose the top 10 terms as input candidates for the re-ranking module. The temperature for LLM inference is set to 0, and the seed is set to 42.

4.3 Evaluation of Knowledge-Enhanced Retrieval

We conducted experiments to prove the importance of the knowledge card for the embedding-based retrieval stage, and the evaluation metric is the Hit Rate, denoted as "HR@num", which means the ratio of samples in which the candidates contain the corresponding normalized term, where "num" represents the number of candidates to be retrieved, the results are displayed in the Table 1. We also compared the effect of RAG on the quality of knowledge cards in it. Additionally, in the demonstration selection module, as mentioned above, we used the same retrieval technique to select the demonstration examples, and we show the corresponding effect in the Appendix Table A1.

In the recall phase, the results of all three datasets specify that the use of both mentions and the name of the term, as well as the knowledge card, will result in a higher hit rate than the use of only the name in general. Introducing knowledge cards enhances the retrieval process by incorporating additional information and context. This additional knowledge helps refine the candidate set and improves the recall rate, and RAG further improves performance, alleviates some of the illusions, and makes the information on the knowledge cards more accurate.

Meanwhile, when we consider it as an unsupervised term normalization method directly in the top half of Table 2, we only consider the term with the highest scores, and we still notice that the results after using the knowledge cards are much better than the traditional BM25 model and TF-IDF model, as well as better than just using the advanced embedding model.

These improvements indicate that the introduction of knowledge cards can enhance the retrieval process by integrating additional information and context. This additional knowledge helps the embedded vectors have more specific semantics, helping to find terms with the same semantics.

However, we have also noticed the superior performance of advanced embedding models, and it can be noted that when we select a more significant number of candidates (e.g., 200), the difference between whether or not to use the knowledge card is not so significant, suggesting that these advanced models are learning richer semantics from a large amount of data. In addition, in our demonstration selection experiments, we found that on the TwADR-L and SMM4H-17 datasets, sometimes the results are better without using the knowledge card instead, as we will discuss in the Limitation Section 6.

Setting	SMM4H-17
Top-K Ranking	HR@10
Ours	97.36
w/o Knowledge-Enhanced Retrieval	96.20
w/ Knowledge-Enhanced Retrieval	
w/o CoT Instructions	93.64
w/o Demonstration Examples	76.96
w/o Grouping	96.56
w/ Grouping	
w/o Balanced Grouping	97.12
w/o Expanded Candidates	93.04
Term Selection	Acc
Ours	90.84
w/o Knowledge-Enhanced Retrieval	90.64
w/ Knowledge-Enhanced Retrieval	
w/o CoT Instructions	84.92
w/o Demonstration Examples	58.40
w/o Grouping	90.72
w/ Grouping	
w/o Balanced Grouping	90.52
w/o Expanded Candidates	87.88
w/o Collaboration Re-ranking	89.84
w/o Ensemble(A_c/A_o/A_i)	90.76/ 90.56 /90.80

Table 3: Ablation experiments to validate the effectiveness of individual modules, the indentation indicates the subordination between the different settings.

4.4 Evaluation of Multi-Agent Collaboration Ranking

Although we proposed a training-free terminology normalization framework, we still use the demonstration examples from the training set to enable the LLM to accomplish the task through in-context learning. Therefore, we compare our approach to supervised methods using the same datasets.

The evaluation metric of the final normalization result is the accuracy score, which denotes the percentage of samples where the selected term is the correct normalized term. The bottom half of Table 2 presents the accuracy scores of the introduced methods compared to our proposed model. Meanwhile, to study the contribution of each module to the final result, we conducted ablation experiments on the SMM4H-17 dataset, which has the most extensive standard terminology base and the most significant number of semantic types. The specific results are displayed in Table 3.

Our proposed method significantly improves over models that have been fine-tuned on individual datasets, which were only intended to provide demonstration examples for in-context learning without requiring parameter fine-tuning. The ablation experiments demonstrate that all of our proposed modules positively contribute to the final performance. The primary contributors are the high-quality demonstrations, the specifically designed CoT instructions, the expanded candidate terms supplemented by the demonstration examples, and the collaborative re-ranking module. It is evident that supervised signals are crucial for informing the LLM agents. Introducing medical persona agents yields more accurate results as different agents reason to different conclusions and can complement each other. As the context lengths supported by current advanced LLMs have increased and their logical reasoning capabilities have improved, grouping and ensemble strategies have proven minor yet effective enhancements to the system's robustness.

5 Conclusion

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

For Knowledge-Enhanced Retrieval, to address the ambiguity caused by short texts, we expand mentions and terms using a terminology expert agent. This agent uses a search engine tool combined with UMLS to generate knowledge cards, providing more informative vector representations during retrieval. This improves the accuracy and hit rate across various datasets without the additional training of a supervised recall model. The agent's use of a tool follows an RAG technique to obtain high-quality knowledge cards and to minimize hallucinations

For Multi-Agent Collaboration Ranking, we leverage the reasoning capabilities of the LLM agents to rank and re-rank the candidate terms further to improve performance. By using a very comprehensive and effective prompt, the terminology expert agent is able to narrow down the list of candidate terms by completing the Top-K ranking task. Then, we modify the prompt and introduce three medical persona agents: a clinical doctor agent, an outpatient doctor agent, and an internet doctor agent. These agents collaboratively reason to achieve more precise term normalization results.

With extensive experiments on the framework, experimental results demonstrate that all our proposed modules are effective. Remarkably, our untrained framework achieves the same level of performance as the state-of-the-art methods.

6 Limitations

First, we observed that the knowledge cards negatively impacted the demonstration selection experiments. This was due to calculating the semantic similarity between mentions during example selection, which differs from the similarity between mentions and terms. Mentions often have slight character differences but are not significantly distinct overall, especially given the high repetition rate of mentions in the SMM4H-17 dataset. Consequently, 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.

Finally, we propose a training-free multi-agent collaboration framework to accomplish the task, using advanced LLMs such as ChatGPT as agents. However, we cannot entirely eliminate randomness even with the temperature set to 0 and fixed seeds provided.

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.

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.

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.

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.

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.

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

Algorithm 3: 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$(m, D, \hat{\mathbf{m}}, \hat{\mathbf{D}}) \rightarrow E$;

Dataset	De-dup	NAME	KC	HR@1	HR@5	HR@10	HR@20	HR@50	HR@100	HR@200
AskPatient	✗	✓	✗	82.68	93.98	96.21	97.65	98.95	99.51	99.71
		✓	✓	**84.30**	**94.26**	**96.36**	**97.67**	**99.00**	**99.56**	**99.81**
	✓	✓	✗	70.92	89.65	93.47	95.96	98.19	99.15	99.50
		✓	✓	**73.67**	**90.10**	**93.75**	**96.01**	**98.27**	**99.24**	**99.68**
TwADR-L	✗	✓	✗	**41.20**	73.70	81.87	87.83	**93.11**	95.79	**97.93**
		✓	✓	40.06	**74.47**	**82.72**	**88.30**	93.04	**96.10**	97.63
	✓	✓	✗	**25.85**	**60.18**	71.54	80.92	89.00	93.33	**96.69**
		✓	✓	23.40	59.60	**72.47**	**81.27**	**88.87**	**93.75**	96.19
SMM4H-17	✗	✓	✗	**89.68**	**94.96**	**96.20**	**97.16**	**97.72**	**97.88**	**98.08**
		✓	✓	89.48	94.72	96.08	96.80	97.36	97.84	98.00
	✓	✓	✗	**68.95**	**84.84**	**88.56**	**91.46**	**93.14**	**93.62**	**94.22**
		✓	✓	68.35	84.12	88.21	90.37	92.06	93.50	93.98

Table A1: The Demonstration Selection experiment, where "De-dup" denotes deduplication, meaning that I remove samples in the test set that duplicate mentions in the training set, "NAME" denotes the names of mentions and terms used in retrieval, "KC" denotes the knowledge cards (with RAG) used in retrieval, "HR@num" denotes the hit rate of the terms of examples containing the standard term corresponding to the input mention, and "num" denotes the number of examples recalled.

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:
You are asked to play the role of a doctor and you need to help me with a knowledge card generation task based on your medical knowledge.
For knowledge Card Generation, please recognize the medical terms in the input (e.g., disease, symptom, procedure, medication) and generate a knowledge card for them.
Please decide on the content of the knowledge card based on your medical knowledge, but it must include definitional descriptions and I will give you some references for common terminology type content. Knowledge card content needs to be exported item by item.

Knowledge Card Content Dimension Reference:
Disease diagnosis terms can contain dimensions such as definition description, etiology, pathology, site, disease type, and clinical manifestations (e.g., symptoms, characteristics, classification, gender, age, acute chronic, onset time).
Symptom terms may contain dimensions such as definition description, cause, classification, site, characteristics, and associated diseases.
Surgical operation terms may contain dimensions such as definition description, surgical technique, target site, surgical approach, and nature of the surgical condition, etc.
Medicine terms can contain dimensions such as definition description, active ingredient, content specification, dosage form, etc.

Requirements:
1. be as detailed as possible, consistent with medical knowledge, not made up, unrecognized term types and dimensions need not be output.
2. do not refuse to answer, output relevant medical knowledge as much as possible.
3. indicate the type of terminology, if possible
4. do not engage in explanations and politeness.
5. do not make additional summaries.

Input:
{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, ...,
"topk_list": [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 “Collaboration Re-ranking” module.