Entity Recognition in Traditional Chinese Medicine Package Inserts Based on Large Language Models

Yali Wan[1,*], Deyi Xiong[2]
[1]School of Information Technology Engineering, Guangzhou College of Commerce, Guangzhou, Guangdong, China
[2]Chongqing Academy of Metrology and Quality Inspection, Chongqing, China
**Corresponding Author*

Abstract: Aiming at the problems in Traditional Chinese Medicine (TCM) package insert texts, such as conceptual ambiguity, professional terminology, and lack of large-scale annotated data, this paper proposes an entity recognition method based on Large Language Models (LLMs). This method selects DeepSeek-V3.2-exp as the base model and designs an iterative prompt optimization strategy. The basic prompt is constructed by defining the task and injecting domain knowledge. On this basis, Chain-of-Thought (CoT) reasoning rules are introduced to build a structured decision-making process to guide the model in performing entity recognition for TCM package inserts, effectively enhancing the model's ability to distinguish semantically ambiguous entities. Experimental results show that compared to the basic prompt, the optimized prompt incorporating CoT achieves a good improvement in the overall F1 score.

Keywords: Named Entity Recognition; LLM; Instruction Manual of Traditional Chinese Medicine

1. Introduction

With the development of information technology, various types of data stored in electronic text form are continuously accumulating in various fields. Among these, medical text big data, including drug package inserts and TCM package inserts, holds significant importance for extracting valuable data to support services like medical decision-making. Natural Language Processing (NLP) is a technology that uses computers to assist in analyzing and understanding textual data. Named Entity Recognition (NER) is one of the fundamental tasks in NLP, enabling the extraction of entities with specific meanings from specific textual big data. The purpose of entity recognition in TCM package inserts is to extract effective medical information from vast amounts of medical text data, accurately identifying specific types of named entities with medical application value, laying the foundation for constructing TCM knowledge graphs.

Currently, in the research history of Named Entity Recognition, it is mainly divided into three categories:

Rule-based NER methods. These methods primarily rely on manually compiled dictionaries, using string matching to extract structured information from entity texts. Anthony [1] et al. proposed a method that identifies proper nouns as candidate entities for annotation based on part-of-speech tagging results, while also considering the use of specific symbols and conjunctions. Through multiple manually compiled dictionaries, the system could handle three types of named entities: persons, locations, and organizations. Quimbaya [2] et al., addressing issues like low text quality, variable semantics, missing punctuation/vocabulary, and coarse terminology classification in electronic medical records, proposed an entity recognition system based on a domain dictionary.

Statistical machine learning-based NER methods. These methods mainly include Hidden Markov Models (HMM) [3], Decision Trees (DT) [4], Conditional Random Fields (CRF) [5], Maximum Entropy Models (ME), and Support Vector Machines (SVM). Feng [6] et al., targeting the lack of large-scale datasets in the cybersecurity domain and the difficulty of rule and dictionary-based methods in handling complex structured text, introduced a Conditional Random Field (CRF) extractor based on regular expressions and an entity dictionary, effectively improving the model's

recognition performance. Jiang [7] et al., addressing the issues of sparse entity attributes and fuzzy boundaries in NER, proposed context feature extraction based on information gain and a feature expansion method combining character clusters and lexicons, effectively integrated through a Maximum Entropy Model, ultimately enhancing the representation ability of named entity features.

Deep learning-based named entity recognition methods. These methods mainly include Recurrent Neural Networks (RNNs) and their variant networks, which are capable of processing textual information with contextual sequential dependencies. Gregoric [8] et al. proposed a novel architecture that deploys multiple independent Bidirectional Long Short-Term Memory (Bi-LSTM) units on the same input and employs an inter-model regularization term to enhance the diversity among the units. This architecture aims to reduce the total parameter count by distributing computations across multiple smaller LSTM units. Ma [9] et al., focusing on scenarios with scarce dataset resources, proposed a named entity recognition method based on an attention mechanism, Iterated Dilated Convolutional Neural Networks (IDCNN), and Conditional Random Fields (CRF). This method combines the characteristics of the attention mechanism, IDCNN, and CRF to construct a model for automatic entity information recognition. Zheng [10] et al., aiming to tackle the problems of complex entity naming and long entity recognition in specific domains, proposed a dual-channel BERT network combining CNN and BiLSTM. This model extracts features of named entities using CNN and captures contextual semantic order using BiLSTM.

Based on extensive research by scholars, Named Entity Recognition has gradually matured.

2. Methodology

This paper presents a method for entity recognition in TCM package inserts based on Large Language Models. DeepSeek-V3.2-exp is selected as the base model according to the dataset and the task's requirement for deep semantic understanding.

2.1 Model Selection

Addressing the problems of conceptual ambiguity, contextual dependency, and professional terminology in TCM package insert entity recognition requires the model to possess deep semantic understanding capabilities, rather than merely performing pattern matching. Additionally, publicly available entity naming datasets in the medical field are scarce, manual annotation costs are high, and training deep learning models lacks sufficient data volume. Therefore, this paper introduces Large Language Models to solve the problem of reliance on large amounts of annotated data. Leveraging their vast pre-trained knowledge, LLMs can improve the generalization ability of Named Entity Recognition under few-shot conditions.

The DeepSeek-V3.2-exp model incorporates a sparse attention mechanism, enabling better processing of Chinese idioms and professional terminology. Consequently, this model is well-suited to address the problems present in TCM package inserts. It demonstrates outstanding performance in tasks requiring multi-step reasoning and strict instruction adherence.

In summary, the Chinese language capability, powerful instruction following and reasoning skills, as well as the efficient architecture of DeepSeek-V3.2-exp, enable it to solve the problems of deep semantic understanding and complex rule application required by the TCM entity recognition task, hence it is selected as the base model in this paper.

2.2 Prompt Optimization Strategy

The entity recognition method in this paper guides the DeepSeek-V3.2-exp model to understand and complete the task of entity recognition in TCM package inserts through a designed initial prompt. This paper adopts an iterative prompt optimization strategy, specifically including the following 4 stages:

2.2.1 Basic Prompt Construction

First, the task of entity recognition in TCM package inserts is defined as: "identify all specified types of entities from the given TCM package insert and return them in a structured format." Subsequently, the model is provided with detailed definitions and examples of 13 entity types to supplement knowledge of professional terminology in the TCM domain.

2.2.2 Output Standardization and Constraints

This paper sets the JSON format as the output standard. Simultaneously, through explicit constraints (such as "Ensure entities truly exist in the text", "Return each entity independently",

etc.), it avoids the issue of inconsistent output formats from LLMs.

2.2.3 Optimized Prompt Construction

Based on the basic prompt (Prompt v1), this paper introduces Chain-of-Thought (CoT) reasoning rules to construct an optimized prompt (Prompt v2). This is the core optimization point of this strategy. The core idea of CoT is to force the model to demonstrate its internal reasoning steps before giving the final answer, thereby decomposing the complex entity classification problem into a sequential, transparent decision-making process. This paper integrates the "TCM Knowledge Graph Entity Comparison Table" and "judgment mnemonic" into the prompts. This rule essentially constitutes a decision tree discrimination method, which guides the model to perform the following sequential judgments when encountering candidate entities:

Domain Filtering: Does the vocabulary belong to the professional scope of TCM?

Attribute Discrimination: Does it describe the drug itself (such as name, ingredient, dosage form, property & flavor, efficacy)?

State Discrimination: If not a drug attribute, does it describe a health state (is it a specific disease, a single symptom, or a comprehensive syndrome)?

Associated Factor Discrimination: If none of the above, does it belong to related diet, population, or drug categories?

Through the aforementioned structured reasoning process, the model can more effectively distinguish semantically ambiguous entities. The final prompt is shown in Table 1.

Package Insert Text Example:

60 tablets/box. Non-prescription drug (Class A) Yunnan Baiyao Group Co., Ltd. 1. Avoid spicy, raw, cold, and greasy foods. 2. Contraindicated in those with weak spleen and stomach, difficulty in ingestion, vomiting, diarrhea, abdominal distension, loose stools, cough with profuse phlegm. 3. Not suitable for concurrent use with cold medications. 4. Not suitable to take with Veratrum Nigrum and its preparations during medication. 5. Patients with hypertension, diabetes, or those receiving other drug treatments should take under physician guidance.

Entity Definition Example:

"DRUG": "Drug (DRUG): The name of a traditional Chinese medicine, referring to substances used for preventing, treating, diagnosing diseases, and possessing rehabilitation and health care effects under the guidance of TCM theory. TCM mainly originates from natural medicines and their processed products, including plant medicines, animal medicines, mineral medicines, and some chemical and biological products. Examples: Wu Ji Bai Feng Wan, Xin Sheng Hua Granules, Liu Wei Di Huang Wan, Xiao Yao San".

Judgment Tip Example:

DRUG Drug: Complete traditional Chinese medicine or compound preparation, with fixed name and clinical use | Liu Wei Di Huang Wan, Wu Ji Bai Feng Wan | If it is a drug that can be directly purchased/used, it belongs to DRUG.

Judgment Mnemonic Example:

Drug-related: Specific drug vs DRUG; Medicinal material vs INGREDIENT; Category vs GROUP; Dosage form vs DOSAGE; Flavor vs TASTE; Efficacy vs EFFICACY.

Table 1. Entity Recognition Prompt Content

Component	Description
Final Prompt	Final Prompt Please identify all entities from the following TCM package insert and return the results in the specified format. Entity Type Definitions: {Entity Type Definitions} Package Insert Text Content: {Package Insert Text Content} Judgment Tips: {Judgment Tips} Judgment Mnemonic: {Judgment Mnemonic} JSON Format Requirement: { "entities": [{ "text": "Entity text", "label": "Entity type" }] } Notes: 1. Only return JSON format, do not include any other text. 2. Ensure all entities truly exist in the text. 3. Entity types must be one of the 13 types mentioned above. 4. Return each entity only once; do not put multiple entities within the same entity object.

2.2.4 Iterative Evaluation and Effect Verification

To quantify the effect of prompt optimization, this paper uses both the basic prompt (v1) and the optimized prompt incorporating CoT (v2) to perform entity recognition on the same test set, and compares their results from two dimensions:

category-specific performance and overall performance.

3. Experimental Results and Analysis

3.1 Dataset

The TCM package insert entity dataset comes from the publicly available Ali Tianchi dataset. This dataset contains a total of 1997 TCM package insert data entries. Entities include 13 types: Drug, Drug Ingredient, Disease, Symptom, Syndrome, Disease Group, Food, Food Group, Population, Drug Group, Drug Dosage Form, Drug Property & Taste, and TCM Efficacy.

3.2 Evaluation Metrics

Common evaluation metrics for Named Entity Recognition tasks include Precision, Recall, and F1 score. The specific formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Where TP represents predicted as positive and actual is also positive (True Positive); TN represents predicted as negative and actual is also negative (True Negative), FP represents predicted as positive but actual is negative (False Positive); FN represents predicted as negative but actual is positive (False Negative). The F1 score formula derived from Precision and Recall is as follows:

$$F = \frac{2PR}{P + R} \quad (3)$$

3.3 Experimental Results Analysis

During the testing phase, 100 samples were randomly selected from the dataset for testing experiments. The performance comparison of different entity categories under different prompt strategies is shown in Table 2 below, and the overall performance comparison of different prompt strategies is shown in Figure 1 below. Among them, Test Prediction 1 represents the test results of the initial prompt, and Test Prediction 2 represents the test results of the prompt optimization strategy after introducing CoT rules.

Table 2. Performance Comparison of Different Entity Categories under Two Prompt

Entity Category	Ground Truth Count	Test Prediction 1			Test Prediction 2		
		F1 Score	Precision	Recall	F1 Score	Precision	Recall
SYMPTOM	631	0.7335	0.7816	0.691	0.7574	0.8029	0.7167
DRUG_EFFICACY	350	0.764	0.7669	0.7612	0.7549	0.7886	0.7239
PERSON_GROUP	172	0.5238	0.4783	0.5789	0.5789	0.5789	0.5789
DRUG_INGREDIENT	160	0.8039	0.82	0.7885	0.8269	0.8269	0.8269
SYNDROME	132	0.5063	0.4255	0.625	0.5195	0.4444	0.625
DISEASE	122	0.4969	0.3419	0.9091	0.5419	0.3784	0.9545
DRUG_DOSAGE	99	0.6786	0.6786	0.6786	0.6786	0.6786	0.6786
FOOD_GROUP	92	0.5217	0.5	0.5455	0.5217	0.5	0.5455
DRUG_TASTE	78	0.6939	0.5862	0.85	0.7391	0.6538	0.85
DISEASE_GROUP	54	0.32	0.5	0.2353	0.3077	0.4444	0.2353
DRUG	37	0.5263	0.4	0.7692	0.5882	0.4762	0.7692
FOOD	17	0.5	0.6667	0.4	0.6	0.6	0.6
DRUG_GROUP	3	0.5714	0.5	0.6667	0.2857	0.25	0.3333

As can be seen from Table 2, for most entity categories in TCM package inserts, Test Prediction 2 outperforms Test Prediction 1 across all three metrics: F1 Score, Precision, and Recall. The F1 scores for SYMPTOM, DRUG_EFFICACY, and DRUG_INGREDIENT all exceeded 0.75 in Test Prediction 2. This indicates that the model achieves satisfactory recognition performance for symptoms, efficacy, and ingredients in TCM package inserts. Thus, it can be concluded that the DeepSeek-V3.2-exp-based entity recognition method proposed in this paper demonstrates effective performance on core entities (symptoms, efficacy, and ingredients) in TCM package inserts, validating the effectiveness of iteratively optimizing prompts by incorporating CoT reasoning rules.

As shown in Figure 1, compared to Test Prediction 1, Test Prediction 2 demonstrates overall performance improvements with increases of 2.87% in F1 Score, 0.82% in Precision, and 1.94% in Recall. These results indicate that the optimized prompt strategy

achieves better overall recognition effectiveness. This demonstrates that iteratively optimizing prompts by incorporating CoT reasoning rules can effectively enhance overall model performance, yielding higher F1 scores and recall rates.

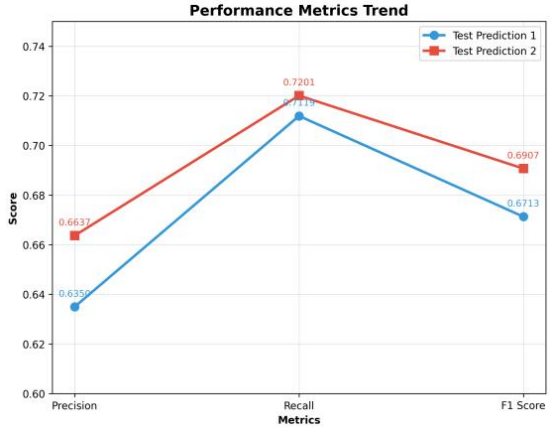

Figure 1. Overall Performance Comparison of Two Prompt Strategies

4. Conclusion

This paper proposes a large language model-based method for entity recognition in Traditional Chinese Medicine package inserts. Compared with traditional entity recognition approaches, this method effectively addresses challenges such as conceptual ambiguity and data sparsity in entity recognition tasks. Experimental results demonstrate that the iterative prompt optimization strategy improves model performance, thereby validating its advantages in both category-specific and overall entity recognition metrics.

Acknowledgments

This work was supported by the following projects:

1. Higher Education Teaching Reform Project of Guangzhou College of Commerce: "Teaching Reform and Research of the 'Digital Image Processing' Course Based on Project-Driven and Engineering Practice" (No. 2024JXGG42).
2. 2025 Guangdong Provincial Education Science Planning Project (Higher Education Specialization): "Restructuring the 'Digital Intelligence-Driven, Four Integrations and Five Progressions' Curriculum System and Innovating the Talent Development Model for Big Data Majors in New Business Studies" (No. 2025GXJK352).
3. "Artificial Intelligence + Data Collection Technology" Pilot Course of Guangzhou College of Commerce (No. 2024rgznsdkc06).
4. 2024 Quality Engineering Project "Data Collection Technology Course Teaching and Research Section" of Guangzhou College of Commerce (No. 2024ZLGC11).
5. First-Class Course "Data Collection Technology" of Guangzhou College of Commerce (No. 2020XJYLKC06).
6. 2024 Project of the Teaching Management Association of Guangdong Provincial Universities (No. GDZLGL2428).

References

[1] Anthony P, Alfred R, Leong L C, et al. A rule-based named-entity recognition for malay articles // International Conference on Advanced Data Mining & Applications.2013.
[2] Muñoz O M, Quimbaya A P, Sierra A, Gonzalez R A, García A A. Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach. Procedia Computer Science, 2016, 100: 55–61.
[3] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name. Machine Learning, 1999, 34:211-231.
[4] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003, 188-191.
[5] Lafferty J, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (ICML), 2001, 282-289.
[6] Feng Y, Jiang B, Wang L, et al. Cybersecurity named entity recognition using multi-modal ensemble learning. IEEE Access, 2020, 8, 63214-63224.
[7] Jiang W, Yi G, Wang X l. Improving Feature Extraction in Named Entity Recognition Based on Maximum Entropy Model // 2006 International Conference on Machine Learning and Cybernetics. Dalian, China: IEEE, 2006: 2630-2635.
[8] Ukov-Gregori A, Bachrach Y, Coope S. Named entity recognition withparallel

recurrent neural networks // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 69-74.

[9] Ma X, Yu R, Gao Co, Wei Z, Xia Y, Wang X, Liu H. Research on named entity recognition method of marine natural products based on attention mechanism. Frontiers in Chemistry, 2023, (11):958002.

[10] Zheng Z, Liu M, Weng Z. A Chinese BERT-based dual-channel named entity recognition method for solid rocket engines. Electronics, 2023, 12(3):752.

