



CMM-EmbedCluster: a clustering framework for Chinese materia medica based on large language model and Chinese materia medica property theory

Jiayi He^{a, b}, Jiadong Xie^b, Kongfa Hu^{b, c, d}, Haiyan Li^{a*}

a. Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

b. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023, China

c. Jiangsu Collaborative Innovation Center of Traditional Chinese Medicine in Prevention and Treatment of Tumor, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023, China

d. Jiangsu Engineering Research Center for Smart Traditional Chinese Medicine Health Services, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023, China

ARTICLE INFO

Article history

Received 17 December 2025

Accepted 16 April 2026

Available online 25 June 2026

Keywords

Large language model

Chinese materia medica property theory

Chinese materia medica clustering

Semantic representation

k-Means clustering

CMM-EmbedCluster

ABSTRACT

Objective This study proposes a clustering framework for Chinese materia medica (CMM) based on a large language model (LLM), aiming to explore potential compatibility patterns among CMMs from the semantic perspective of CMM property theory.

Methods First, a CMM property knowledge base was constructed based on *Chinese Materia Medica*, including 567 commonly used CMMs characterized by four properties, five flavors, and meridian tropism. Then, 49 CMMs derived from 10 prescriptions for Zangdu (脏毒, pathogenic toxins) recorded in *Waiké Zhengzong* (《外科正宗》, *Orthodox Manual of External Medicine*) and *Yangke Xinde Ji* (《疡科心得集》, *Collected Insights on Ulcer Medicine*) were selected as the experimental dataset. Five semantic representation methods—One-Hot, Word2Vec, Bidirectional Encoder Representations from Transformers (BERT), Beijing Academy of Artificial Intelligence General Embedding (BGE), and Qwen—were applied to encode CMM property information into vector representations. Subsequently, t-distributed Stochastic Neighbor Embedding (t-SNE) was used for nonlinear dimensionality reduction on high-dimensional semantic vectors, followed by *k*-means clustering ($k = 7$). Clustering performance was evaluated using the Silhouette Score (SS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI).

Results The Qwen-based clustering method, CMM-EmbedCluster, achieved the highest SS (0.607 4) and CHI (158.057 2), as well as the lowest DBI (0.499 5), indicating improved cluster separation and compactness compared with other methods. Visualization of CMM clustering results showed that the clusters were well separated in the low-dimensional space, with strong inter-cluster discrimination and high intra-cluster functional consistency. Further interpretability analysis of CMM clustering results revealed stable structural differences among clusters in terms of four properties, five flavors, and meridian tropism, forming functional partitions consistent with CMM property theory.

Conclusion CMM-EmbedCluster utilizes an LLM to achieve semantic-level representation and clustering of CMMs within the framework of CMM property theory, providing support for exploring potential compatibility patterns among CMMs from the perspective of CMM property semantics.

*Corresponding author: Haiyan Li, E-mail: li_haiyan01@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmcd.2026.05.010

Citation: HE JY, XIE JD, HU KF, et al. CMM-EmbedCluster: a clustering framework for Chinese materia medica based on large language model and Chinese materia medica property theory. *Digital Chinese Medicine*, 2026, 9(2): 278-289.

Copyright © 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

1 Introduction

Traditional Chinese medicine (TCM), guided by the principle of Bianzheng Lunzhi (辨证论治, syndrome differentiation and treatment), adopts a holistic approach and focuses on dynamic balance to achieve the prevention and treatment of diseases^[1]. As the core material foundation of TCM theory, Chinese materia medica (CMM) strongly relies on rational compatibility, which is key to optimizing therapeutic efficacy and reducing or regulating potential toxicity within the human body^[2]. Compatibility refers to the combination of two or more CMMs guided by specific theoretical principles, whereby synergistic effects such as mutual reinforcement and assistance are exploited to enhance therapeutic efficacy or reduce adverse effects^[3]. Over centuries of clinical practice, TCM physicians have accumulated extensive empirical knowledge and developed a systematic framework for CMM compatibility, most notably the Jun-Chen-Zuo-Shi (君臣佐使, sovereign-minister-assistant-courier) formula structure^[4]. Existing research predominantly relies on clinical records or prescription data, employing data mining techniques such as co-occurrence frequency, mutual information, and association rules to construct herb networks or perform clustering analysis, aiming to uncover underlying compatibility patterns^[5]. However, these approaches mainly focus on statistical co-occurrence in clinical text, often overlooking the intrinsic properties of CMMs, thus limiting their ability to reveal the semantic logic and underlying mechanisms of CMM compatibility.

The theory of CMM property is a foundational component of TCM, encompassing key categories such as the four properties (cold, hot, warm, and cool), five flavors (pungent, sweet, sour, bitter, and salty), ascending/descending and floating/sinking tendencies, and meridian tropism. Together, these characteristics describe the directional actions of CMMs and their correspondence to internal organs and meridians^[6]. This theory is fundamental not only to understanding pharmacological mechanisms but also to supporting research on CMM classification^[7] and efficacy prediction^[8]. In the field of TCM informatics, the quantitative and computational representation of CMM property theory has long been an important research focus. In early studies, these properties were often converted into binary codes or numerically quantified^[9], for example, by reducing the four properties, five flavors, and meridian tropism to fixed numerical values^[10]. Although such approaches enable the computational representation of CMM property information, they generally rely on manually defined rules and thus have limited capacity to capture the latent semantic relationships among CMM property concepts. With the development of machine learning methods, some studies have attempted to predict the cold-heat nature or functional categories of CMM using chemical component

features or multidimensional CMM property indicators, thereby improving the automation of CMM property classification. For instance, JIA et al.^[11] applied the *k*-nearest neighbors algorithm to predict the cold, hot, and neutral properties of compounds in CMM, demonstrating that chemical structure to some extent determines CMM property tendencies. However, these studies have mainly emphasized structural feature modeling, with relatively limited attention to the semantic expressive capacity of the CMM property texts themselves. In recent years, natural language processing techniques have gradually been introduced into CMM research, and study has used word embedding models or text representation methods to semantically model CMM terminology for tasks such as formula classification, knowledge extraction, and relation prediction^[12]. Nevertheless, existing research has largely focused on clinical texts or prescription data, while semantic modeling of the CMM property theory itself remains relatively limited, particularly in terms of systematic semantic representation methods for core concepts such as the four properties, five flavors, and meridian tropism. Therefore, developing representation methods capable of reflecting the intrinsic semantic structure of CMM property theory remains a key issue in intelligent CMM research.

Natural language processing technologies have advanced rapidly, and large language models (LLMs) in particular have demonstrated strong performance in semantic understanding and representation^[13]. Pretrained language models represented by Bidirectional Encoder Representations from Transformers (BERT) have made remarkable progress in the semantic modeling of medical texts. However, BERT and its variants, which are primarily based on bidirectional encoding architectures, focus on modeling contextual relationships between words, potentially limiting their ability to represent the highly abstract, metaphorical, and relation-oriented semantic structures inherent in CMM property theory. Built on large-scale neural networks trained on massive corpora, LLMs are capable of learning compositional, interactive, and emergent linguistic patterns, encoding rich semantic information within their parameter spaces^[14]. Recent applications of LLM-based methods have shown promising performance in tasks such as CMM prescription recommendation^[15], medical named entity recognition^[16], and drug interaction prediction^[17], offering new technical approaches for CMM semantic modeling. In parallel, studies on phenomics have incorporated molecular chemical structures and protein-protein interaction networks to construct feature representations for CMM^[18–20]. Nevertheless, research focusing on the semantic structure modeling of fundamental CMM theories remains relatively limited, and most previous studies have largely relied on manually constructed knowledge graphs for analysis^[21]. Incorporating LLM into the study of CMM

property theory can overcome the limitations of traditional feature representations and build semantic knowledge systems more aligned with the conceptual framework of CMM, thereby advancing intelligent TCM research.

In this study, we proposed a CMM clustering method that integrated LLM with CMM property theory. Unlike previous studies, the present study focused on representing the semantic structure of CMM property theory and proposed a semantic clustering framework for CMM based on LLM embeddings. By constructing a continuous semantic space from CMM property texts, this framework enables the automatic identification of functional structures and compatibility relationships among CMMs, thereby addressing the challenges of quantifying CMM property knowledge and uncovering its underlying semantic relationship.

2 Data and methods

2.1 Methodological framework

The workflow of the proposed CMM clustering approach is illustrated in Figure 1. The framework integrates CMM property theory with LLM-based semantic representation and consists of three sequential steps. First, CMM property information is extracted and structured into textual descriptions based on key attributes, including four properties, five flavors, and meridian tropism. This

information is derived from the textbook *Chinese Materia Medica*^[22], which serves as the source for the CMM property knowledge base. Second, the structured CMM property texts are encoded into vector representations using multiple semantic representation methods, including One-Hot encoding, Word2Vec, BERT, Beijing Academy of Artificial Intelligence General Embedding (BGE), and Qwen, enabling the transformation of symbolic CMM knowledge into a continuous semantic space. Third, dimensionality reduction is performed using t-distributed Stochastic Neighbor Embedding (t-SNE) to project high-dimensional semantic vectors into a low-dimensional space, followed by *k*-means clustering to identify groups of CMMs with similar functional characteristics. The optimal representation method is determined through comparative evaluation of clustering performance.

2.1.1 CMM property knowledge base To obtain standardized CMM property information, we constructed a CMM property knowledge base based on *Chinese Materia Medica*^[22], which includes 567 commonly used CMMs. This knowledge base was used to provide CMM property information for the CMMs included in the subsequent clustering dataset and served as the input for semantic representation. Examples are shown in Table 1.

2.1.2 Semantic representation In this study, the dataset for semantic representation was derived from prescriptions recorded in *Waike Zhengzong* (《外科正宗》),

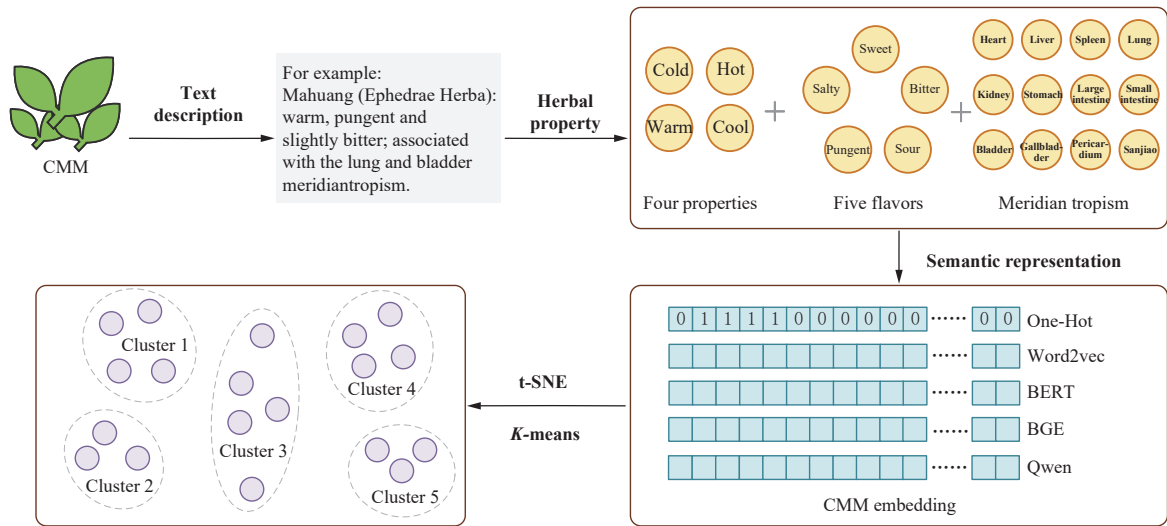


Figure 1 The workflow of the proposed CMM clustering approach

Table 1 Examples of CMM property knowledge

CMM	Property	Flavor	Meridian tropism
Mahuang (Ephedrae Herba)	Warm	Pungent, slightly bitter	Lung, bladder
Rougui (Cinnamomi Cortex)	Very hot	Pungent, sweet	Kidney, spleen, heart, liver
Zhiqiao (Aurantii Fructus)	Slightly cold	Bitter, pungent, sour	Spleen, stomach
Xiangfu (Cyperi Rhizoma)	Neutral	Pungent, slightly bitter, slightly sweet	Liver, spleen, Sanjiao
Shancigu (Cremastreae Pseudobulbus)	Cool	Sweet, slightly pungent	Liver, spleen
Pleiones Pseudobulbus)			

Orthodox Manual of External Medicine)^[23] and *Yangke Xinde Ji* (《疡科心得集》, *Collected Insights on Ulcer Medicine*)^[24]. Considering that the Zangdu (脏毒, pathogenic toxins) syndrome shares similarities with colorectal cancer in terms of anatomical location and clinical manifestations^[25], related prescriptions were selected as the study objects (Supplementary Table S1), with prescriptions 1 – 5 obtained from *Waike Zhengzong* and 6 – 10 from *Yangke Xinde Ji*.

All names of CMMs were standardized according to the *Pharmacopoeia of the People's Republic of China 2025*^[26]. For example, “Shanzhi” was standardized as “Zhizi (*Gardeniae Fructus*)”, and “Jiaozhen” as “Zaojiao-ci (*Gleditsiae Spina*)”. In addition, substances such as Chensha (*Cinnabaris*), Xiaoshi (*Nitrum*), and Jinbo (*Aurum*), primarily mineral or metallic preparations, were excluded due to the lack of standardized descriptions of their CMM properties and their absence from the *Chinese Materia Medica*^[22]. Ultimately, a total of 49 CMMs were included in this study.

To achieve semantic representation of CMM property theory, we employed five representation strategies: One-Hot encoding^[27], Word2Vec embeddings^[28], BERT^[29], BGE^[30], and Qwen text-embedding-v4^[31], an LLM-based embedding method. These approaches were used to construct multi-dimensional CMM feature vectors, ranging from shallow lexical representations to deep semantic embeddings. The resulting representations served as inputs for subsequent clustering analysis, enabling a comparative evaluation of their performance and the selection of the optimal representation method.

(i) One-Hot. One-Hot encoding is a fundamental word vectorization method that converts discrete tokens into computable numerical representations. Its core idea is to assign each unique token in the corpus to a separate dimension in a sparse vector, setting the corresponding dimension to 1 and all others to 0.

(ii) Word2Vec. Word2Vec is a distributed word representation model based on neural networks. Its core idea is to learn the semantic features of words by predicting the target word from its context (or vice versa). The model is grounded in the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts.

(iii) BERT. BERT is a bidirectional pretrained representation model based on the Transformer architecture. Using a self-attention mechanism, it captures both forward and backward contextual semantic information simultaneously, supporting deep semantic representation. Unlike static word embedding models such as Word2Vec, BERT dynamically generates context-aware word representations, effectively alleviating the semantic ambiguity of polysemous words. BERT is pretrained on two tasks: masked language modeling and next sentence prediction.

(iv) BGE. BGE is a general-purpose text embedding

model developed by the Beijing Academy of Artificial Intelligence, designed to generate high-quality semantic embeddings for tasks such as retrieval, clustering, classification, and semantic matching. Compared with traditional word embedding models, BGE is optimized on the basis of large-scale pretrained language models and shows clear advantages in semantic understanding and vector representation consistency.

(v) Qwen. To semantically model CMM property theory, this study adopted Qwen text-embedding-v4, a multilingual unified text embedding model developed by Alibaba Discovery, Adventure, Momentum and Outlook (DAMO) Academy based on the Qwen3 architecture. The model is capable of converting textual, visual, and video data into numerical vector representations. Internally, it employs a causal attention mechanism and appends an [EOS] token at the end of the input sequence. The hidden state corresponding to this token is then used as the final embedding vector.

To achieve semantic representation of CMM property information, this study directly concatenated the CMM property attributes of each CMM into structured text as model input. For example, the input text for Mahuang (*Ephedrae Herba*) was “warm, pungent, and slightly bitter; associated with the lung and bladder meridians”. No additional prompt design or contextual expansion was introduced to minimize the influence of human intervention on semantic representation.

2.1.3 CMMs clustering The high-dimensional semantic embeddings of the 49 CMMs were first reduced using t-SNE, followed by k -means clustering to group the CMMs. Given a set of herb embeddings $D = \{x_1, x_2, \dots, x_m\}$, the k -means algorithm partitions the data into clusters $C = \{C_1, C_2, \dots, C_k\}$ by minimizing the sum of squared errors (SSE):

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (1)$$

where $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ is the centroid of cluster C_i . This objective function measures the compactness of samples within each cluster; a smaller E value indicates higher intra-cluster similarity.

To evaluate the impact of different semantic representation approaches on CMM clustering performance, we conducted experiments using five combined pipelines: One-Hot + t-SNE + k -means, Word2Vec + t-SNE + k -means, BERT + t-SNE + k -means, BGE + t-SNE + k -means, and Qwen + t-SNE + k -means.

2.2 Experimental setup and parameter configuration

All experiments were conducted on a platform equipped with a 32-core 13th Gen Intel(R) Core(TM) i7-13620H

CPU (2.40 GHz). The software environment included scikit-learn version 1.7.2 and OpenAI version 2.2.0. The parameter settings for the five semantic representation approaches are described below. These approaches cover a spectrum of semantic modeling capabilities, from traditional statistical representations and shallow embeddings to pretrained language models and LLM embeddings. This design enabled a systematic evaluation of how different levels of semantic modeling affect CMM property representation and clustering performance. (i) One-hot: a 22-dimensional vector was constructed, covering CMM properties (cold, hot, warm, and cool), medicinal flavors (sour, bitter, sweet, pungent, and salty), and the 12 meridians. (ii) Word2Vec: word vectors were trained with a dimension of 128, a context window size of 5, a minimum word frequency threshold of 1, and 4 computation threads. (iii) BERT: the pretrained BERT-base-Chinese model was used for text vectorization, with a maximum sequence length of 128 tokens. The classification token [CLS] token vector from the last hidden layer was used as the sentence representation, yielding a 768-dimensional vector. (iv) BGE: the bge-large-zh model was used for semantic text encoding, generating 1024-dimensional sentence vectors with vector normalization. (v) Qwen: the text-embedding-v4 model was used, producing 1024-dimensional vectors encoded in float format.

In addition, the parameters for dimensionality reduction and clustering are as follows. (i) t-SNE: the perplexity parameter was set to 5, the learning rate to 100, and cosine similarity was used as the distance metric. (ii) k -Means: the number of clusters was set to 7, the k -means++ algorithm was used for centroid initialization, and the convergence tolerance was set to 1×10^{-3} .

2.3 Evaluation metrics

In clustering analysis, three commonly used evaluation metrics were adopted to assess and compare the clustering performance of different embedding methods: the Silhouette Score (SS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) [32].

(i) SS quantifies the similarity of each sample to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to 1 , where a value close to 1 indicates that the sample is well-matched to its own cluster, a value near -1 suggests a potential misclassification, and a value around 0 means the sample lies close to the decision boundary between two adjacent clusters. The formula is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

where $a(i)$ is the average intra-cluster distance of sample i , and $b(i)$ is the average nearest-cluster distance for

sample i . SS is obtained by averaging $S(i)$ over all samples.

(ii) DBI evaluates the ratio of within-cluster dispersion to between-cluster separation. A lower DBI value corresponds to superior clustering performance. The formula is:

$$DBI = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right) \quad (3)$$

where n is the number of clusters, s_i is the average distance between each point in cluster i and the cluster center C_i , and $d(c_i, c_j)$ is the distance between cluster centers i and j .

(iii) CHI quantifies the ratio of between-cluster dispersion to within-cluster dispersion. A higher CHI indicates more distinct and well-defined clusters. The formula is:

$$CHI = \frac{B(k)}{W(k)} \times \frac{n-k}{k-1} \quad (4)$$

where $B(k)$ is the inter-cluster dispersion, defined as the sum of squared distances between each cluster center and the global data center, and $W(k)$ is the intra-cluster dispersion, defined as the sum of squared distances between each sample and its cluster center. n is the total number of data points, and k is the number of clusters.

2.4 Interpretability analysis

To enhance the interpretability of the clustering results, a semantic property-based analysis was conducted. For each cluster, semantic attributes were encoded as binary features and averaged within the cluster to calculate the proportions of CMMs exhibiting each attribute, resulting in normalized frequencies ranging from 0 to 1 . A heatmap was then generated to visualize the distribution patterns of these semantic attributes across clusters. The clusters were arranged along the vertical axis and semantic properties along the horizontal axis, with color intensity representing the normalized frequency of each attribute within a cluster. This visualization provides an intuitive representation of the semantic differences and characteristic patterns across clusters.

3 Results

3.1 Comparative evaluation of semantic representation methods for CMM clustering

The results demonstrated that the proposed CMM-EmbedCluster framework based on an LLM (Qwen + t-SNE + k -means) achieved the best performance across all three evaluation metrics. Specifically, it attained the highest SS (0.6074), indicating strong intra-cluster compactness and consistency; the highest CHI (158.0572), reflecting strong

inter-cluster distinctness and separation; and the lowest DBI (0.499 5), further confirming the stability and rationality of the clustering structure. In contrast, One-Hot, Word2Vec, BERT, and BGE all yielded inferior results across the three metrics, suggesting that shallow or generic semantic embedding strategies are insufficient to capture the complex relational characteristics inherent in CMM property theory (Table 2).

Table 2 Comparison of clustering performance of five embedding methods for CMMs

Model	SS	DBI	CHI
One-Hot + t-SNE + k-means	0.523 5	0.598 6	114.465 5
Word2Vec + t-SNE + k-means	0.347 4	0.828 6	49.966 1
Bert + t-SNE + k-means	0.490 5	0.627 9	101.192 8
BGE + t-SNE + k-means	0.488 4	0.712 5	107.257 3
Qwen + t-SNE + k-means (CMM-EmbedCluster)	0.607 4	0.499 5	158.057 2

3.2 Visualization of CMMs clustering

Using t-SNE for dimensionality reduction of CMM property semantic vectors followed by k-means clustering, a total of 49 CMMs were divided into seven clusters (Figure 2). The clusters are well separated in the low-dimensional space, indicating strong inter-cluster discrimination. Moreover, CMMs within each cluster exhibit consistent pharmacological characteristics, suggesting that the clustering results effectively captured the underlying functional structure of CMMs. The detailed composition of each cluster is presented in Table 3.

3.3 Interpretability of CMMs clustering

To investigate the semantic characteristics of different clusters and their structural differences in the embedding

space, we conducted an interpretability analysis focusing on the distribution patterns of CMM properties, medicinal flavors, and meridian tropisms within each cluster (Figure 3). The results showed distinct and well-structured differences among clusters across multiple CMM property attributes, suggesting that the LLM-based embedding space encodes rich latent semantic structures consistent with CMM property theory.

3.3.1 Four properties pattern From the perspective of CMM properties distribution, Clusters 0, 2, 4, and 6 were dominated by warm-nature CMMs, whereas Clusters 1, 3, and 5 were mainly composed of cold-nature CMMs, showing a clear cold-heat differentiation pattern. This finding suggested that, during semantic representation, the LLM automatically formed a latent semantic axis reflecting cold-heat attributes, such that CMMs with similar cold-heat properties exhibited higher vector similarity in the embedding space.

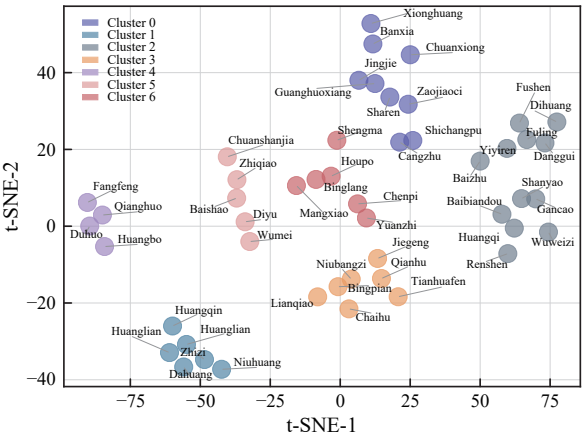


Figure 2 Visualization of CMMs clustering based on t-SNE and k-means using Qwen embeddings

Table 3 Composition of each cluster in CMMs clustering results

Cluster ID	Number of CMMs	Composition
Cluster 0	9	Jingjie (Schizonepetae Herba), Guanghuoxiang (Pogostemonis Herba), Cangzhu (Atractylodis Rhizoma), Sharen (Amomi Fructus), Chuanxiong (Chuanxiong Rhizoma), Banxia (Pinelliae Rhizoma), Zaojiaoci (Gleditsiae Spina), Shichangpu (Acori Tatarinowii Rhizoma), Xionghuang (Realgar)
Cluster 1	6	Zhizi (Gardeniae Fructus), Huangqin (Scutellariae Radix), Huanglian (Coptidis Rhizoma), Huhuanglian (Picrorhizae Rhizoma), Dahuang (Rhei Radix et Rhizoma), Niu Huang (Bovis Calculus)
Cluster 2	12	Shengdihuang (Rehmanniae Radix), Fuling (Poria), Fushen (Poria cum Radix Pini), Yiyiren (Coicis Semen), Renshen (Ginseng Radix et Rhizoma), Huangqi (Astragali Radix), Baizhu (Atractylodis Macrocephalae Rhizoma), Shanyao (Dioscoreae Rhizoma), Baibiandou (Lalab Semen Album), Gancao (Glycyrrhizae Radix et Rhizoma), Danggui (Angelicae Sinensis Radix), Wuweizi (Schisandrae Chinensis Fructus)
Cluster 3	7	Niubangzi (Arctii Fructus), Chaihu (Bupleuri Radix), Tianhuafen (Trichosanthis Radix), Lianqiao (Forsythiae Fructus), Qianhu (Peucedani Radix), Jiegeng (Platycodonis Radix), Bingpian (Borneolum Syntheticum)
Cluster 4	4	Fangfeng (Saposhnikoviae Radix), Qianghuo (Notopterygii Rhizoma et Radix), Huangbo (Phellodendri Chinensis Cortex), Duhuo (Angelicae Pubescentis Radix)
Cluster 5	5	Zhiqiao (Aurantii Fructus), Diyu (Sanguisorbae Radix), Chuanshanjia (Manis Squama), Baishao (Paeoniae Radix Alba), Wumei (Mume Fructus)
Cluster 6	6	Shengma (Cimicifugae Rhizoma), Mangxiao (Natrii Sulfas), Houpo (Magnoliae Officinalis Cortex), Chenpi (Citri Reticulatae Pericarpium), Binglang (Arecae Semen), Yuanzhi (Polygalae Radix)

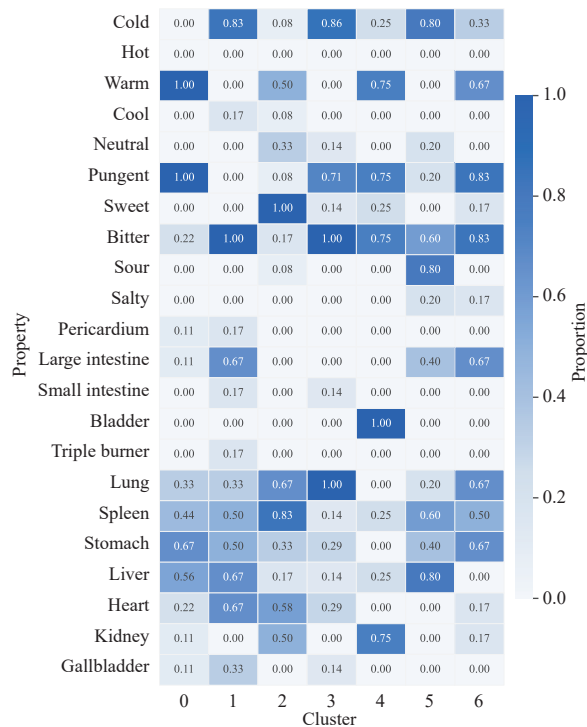


Figure 3 Heatmap of the distribution of CMM properties across seven clusters

3.3.2 Five flavors pattern In terms of five flavors distribution, the clusters also exhibited distinct semantic structures. Overall, different clusters showed clear enrichment patterns for major medicinal flavors, including bitter, sweet, and pungent, suggesting that flavor attributes constituted an important dimension affecting semantic proximity. Specifically, Clusters 1 and 3 contained a relatively high proportion of bitter-flavored CMMs, corresponding to functional semantics such as clearing heat, draining fire, drying dampness, and detoxifying; Clusters 2 and 5 were enriched with sweet-flavored CMMs, reflecting tendencies toward tonifying and regulating effects. Clusters 0 and 6 showed higher proportions of pungent-flavored CMMs, indicating functional orientations such as promoting Qi movement and exterior dispersion. The remaining clusters exhibited mixed distributions of multiple flavors. These results indicated that LLM embeddings simultaneously encoded the functional tendencies of CMMs, such as pungent dispersing, bitter descending, and sweet tonifying, and formed corresponding semantic similarity structures in the vector space.

3.3.3 Meridian tropism pattern With respect to meridian tropism distribution, each cluster showed a distinct affinity for specific Zang-Fu organs. Cluster 0 was mainly enriched in CMMs targeting the stomach and liver meridians; Cluster 1 primarily comprised CMMs acting on the large intestine, liver, and heart meridians; Cluster 2 was dominated by CMMs associated with the lung, spleen, and heart meridians; Cluster 3 was chiefly associated with

CMMs targeting the lung meridian; Cluster 4 mainly involved CMMs acting on the bladder and kidney meridians; Cluster 5 was highly concentrated in CMMs of the spleen and liver meridians; and Cluster 6 primarily involved CMMs corresponding to the large intestine, lung, and stomach meridians. These findings suggested that the LLM not only captured the nature and functional tendencies of CMMs, but also encoded information related to their therapeutic target sites within the semantic space, enabling CMMs with analogous meridian tropisms to form locally aggregated structural clusters.

The differences observed among clusters in terms of CMM properties, medicinal flavors, and meridian tropisms indicated that the semantic vectors generated by the LLM contained multidimensional latent semantic structures reflecting thermal attributes, functional tendencies, and Zang-Fu organs associations. Together, these semantic features governed the similarity relationships among CMMs in the embedding space, enabling the emergence of clustering patterns consistent with TCM theory under unsupervised learning conditions.

4 Discussion

This study proposes CMM-EmbedCluster, a framework that leverages LLM-derived embeddings to represent CMM properties in a continuous semantic space, enabling the identification of latent functional structures. Its key contribution lies in applying LLM-based semantic representations to clustering analysis and compatibility interpretation of CMMs. This work introduces a novel computational paradigm for CMM property theory.

4.1 Semantic representation-driven reconstruction of functional structures in CMM

The comparative results demonstrated that CMM-EmbedCluster outperformed clustering frameworks based on One-Hot, Word2Vec, BERT, and BGE in terms of SS, DBI, and CHI, suggesting that LLM-based representations more effectively capture the latent semantic structure of CMM property theory. These findings suggested that, although CMM property descriptions appeared as discrete textual expressions, they were underpinned by a systematic and logically structured knowledge system. Leveraging modern semantic modeling approaches such as the Qwen text-embedding-v4 model [31], CMM property knowledge can be encoded into high-dimensional vectors that capture latent semantic relationships, enabling structured and quantitative representation [33]. Combined with the clustering outcomes and semantic attribute analysis, the CMM clusters generated by the model exhibited stable structural differences in terms of four properties, five flavors, and meridian tropism, suggesting that the CMM-EmbedCluster embedding space contained

latent semantic dimensions corresponding to thermal properties, action tendencies, and Zang-Fu organs associations. These multidimensional semantic features jointly determined the similarity relationships among CMMs in the vector space, yielding a consistent clustering structure under unsupervised conditions.

At the functional level, the clustering results exhibited clear functional partitioning that was highly consistent with CMM property theory. Specifically, Clusters 0 and 6 primarily consisted of CMMs that regulate Qi, promote Qi movement, and eliminate dampness, corresponding to therapeutic actions such as regulating Qi flow and invigorating the spleen to resolve dampness. Clusters 1, 3, and 4 were enriched with CMMs possessing effects of clearing heat, purging fire, detoxifying, dispelling wind, and relieving itching, forming a functional group characterized by heat-clearing, detoxification, and wind-dispelling properties. Clusters 2 and 5 were mainly composed of tonifying and astringent CMMs, reflecting functions related to replenishing Qi and blood as well as astringent hemostasis. Overall, the clustering structure was largely consistent with classical TCM theoretical principles described in *Chinese Materia Medica* [22], including the classifications of cold-hot and warm-cool properties, as well as functional tendencies such as ascending-descending and tonifying-reducing. These findings suggest that clustering results based on semantic embeddings can, to some extent, reflect the intrinsic functional structure. Notably, this structured pattern was not imposed by predefined functional classification rules, but instead emerged naturally from the inherent semantic organization captured in the embedding space.

4.2 Semantic mapping of modern prescriptions and validation of theoretical consistency

To further validate the correspondence between the clustering structure and the composition of modern clinical prescriptions, two representative formulas commonly used in the contemporary treatment of colorectal cancer, Xianlian Jiedu Fang (仙连解毒方) [34] and Shenbai Jiedu Fang (参白解毒方) [35], were selected for qualitative analysis. Xianlian Jiedu Fang is composed of Xianhecao (*Agrimoniae Herba*), Huanglian (*Coptidis Rhizoma*), Kushen (*Sophorae Flavescentis Radix*), Yiyiren (*Coicis Semen*), Sanleng (*Sparganii Rhizoma*), Ezhu (*Curcumae Rhizoma*), Huangqi (*Astragali Radix*), and Baizhu (*Atractylodis Macrocephalae Rhizoma*). The CMMs in this formula were mainly distributed across Clusters 1, 2, and 5. This reflects its therapeutic functions of clearing heat and detoxifying, strengthening the spleen and eliminating dampness, as well as promoting blood circulation and resolving stasis, which aligned well with the functional groups identified by the model: heat-clearing and dampness-eliminating + Qi and blood tonification +

blood-activating and stasis-resolving. Similarly, Shenbai Jiedu Fang, consisting of Baihuasheshecao (*Hedyotis Diffusae Herba*), Kushen (*Sophorae Flavescentis Radix*), Dangshen (*Codonopsis Radix*), Baizhu (*Atractylodis Macrocephalae Rhizoma*), Yiyiren (*Coicis Semen*), Huanglian (*Coptidis Rhizoma*), Paojiang (*Zingiberis Rhizoma Preparatum*), and Wumei (*Mume Fructus*), also mapped predominantly to Clusters 1, 2, and 5, emphasizing a therapeutic strategy that integrated detoxification with spleen-Qi tonification, consistent with the modern oncological principle of “reinforcing healthy Qi while eliminating pathogenic factors”.

Both formulas were developed by Professor Haibo Cheng's team based on extensive clinical experience [36–38]. Their formulation principles share structural similarities in CMM properties and actions with the medicinal systems described in the Zangdu Theory of *Waike Zhengzong* and the “Discussion on Intestinal Wind and Viscera Toxin” in *Yangke Xinde Ji*. Specifically, the former is centered on heat-clearing and detoxifying CMMs such as Xianhecao (*Agrimoniae Herba*), Huanglian (*Coptidis Rhizoma*), and Kushen (*Sophorae Flavescentis Radix*), supported by spleen-fortifying and blood-activating CMMs like Yiyiren (*Coicis Semen*) and Ezhu (*Curcumae Rhizoma*). In comparison, Shenbai Jiedu Fang incorporates tonifying CMMs such as Dangshen (*Codonopsis Radix*) and Baizhu (*Atractylodis Macrocephalae Rhizoma*) on the basis of detoxification, highlighting the therapeutic strategy of “detoxification and reinforcement”.

This convergence between classical and modern CMM patterns suggested that the clustering model built on semantic representations not only reflected the theoretical structure of CMM formulas, but also revealed potential alignment between ancient pharmacological systems and contemporary clinical practices. Therefore, the findings of this study provided novel semantic evidence for interpreting the theoretical origins of modern empirical formulas, indicating that formulas emphasizing “detoxification and reinforcement” may represent a clinical evolution and innovative application of the traditional Zangdu system. This further supported the value of the proposed semantic-based clustering approach as an effective bridge connecting classical TCM theory with modern clinical practice.

4.3 Advantages and application potential of LLM-based semantic modeling in CMM

From a methodological perspective, LLM-based semantic representation presents multiple distinct advantages. The proposed CMM-EmbedCluster framework effectively captures contextual dependencies and latent semantic relationships in CMM property texts, thereby outperforming traditional methods such as One-Hot and Word2Vec in terms of semantic representation and

structural modeling. Particularly for abstract core concepts in CMMs, such as Yin-Yang and cold-heat, reinforcing-draining and ascending-descending, and meridian tropism, LLMs can encode the hidden logical structures into continuous vector spaces, enabling herb similarities to better reflect the theoretical relationships defined under the classical CMMs system. From this standpoint, the present study provided a novel technical paradigm for the digital and structural representation of CMM property knowledge. In addition, functional herb groups identified through semantic clustering may serve as candidate structural units for optimizing CMMs compatibility, and provide a data-driven foundation for designing new formulas. For example, the functional similarity of CMMs within the same cluster may support the screening of CMM substitutes, while cross-cluster combinations may help uncover potential synergistic interactions^[39], thereby assisting the optimization of clinical prescriptions and the development of novel formulas. This approach may also facilitate the discovery of latent structural patterns in classical formulas, providing further support for the systematic representation of CMMs knowledge and intelligent decision-making^[40].

4.4 Limitations and future prospects

Nevertheless, several limitations exist in the present study. First, the CMM property knowledge base was primarily constructed from the textbook *Chinese Materia Medica*^[22], representing a relatively single information source. Richer textual descriptions from TCM classics such as *Bencao Gangmu* (《本草纲目》, *Compendium of Materia Medica*), *Shanghan Lun* (《伤寒论》, *Treatise on Cold Damage*), or *Mingyi Bielu* (《名医别录》, *Supplementary Records of Famous Physicians*) have not yet been fully integrated, which may limit the diversity and depth of the learned semantic space. Second, the current framework only employed textual features related to CMM properties without incorporating multi-modal data including chemical constituents, molecular targets, metabolic pathways, or modern pharmacological mechanisms. As such, the resulting clusters primarily reflected classical theoretical frameworks rather than comprehensive pharmacodynamic mechanisms. Third, the number of CMMs included in this study was relatively limited (49 in total), which was sufficient for methodological validation but necessitated larger-scale datasets to further evaluate the framework's robustness and generalizability.

Future research can be advanced in the following directions. First, integrate CMM omics data resources, such as Shennong Alpha^[41], HERB^[42], and SymMap^[43], to construct a multimodal feature system that supports joint modeling of CMM properties, chemical components, molecular targets, and disease-related mechanisms. Second, apply graph neural networks^[44] to construct

graph-structured relationships among medicinal CMMs, thereby further enhancing the biological interpretability of the framework. Third, explore retrieval-augmented generation techniques^[45] to build reasoning frameworks for CMM formula compatibility, promoting deep integration between CMM knowledge graphs and LLMs. These approaches are expected to improve both the biological interpretability and intelligent recommendation capability of the system. They will help establish a more systematic and interpretable intelligent framework for CMM formula design, while providing methodological support for the modernization of CMM and innovative formula development.

5 Conclusion

This study proposed an LLM-based semantic clustering framework (CMM-EmbedCluster) to reconstruct the latent functional structure of CMMs by integrating textual information on four properties, five flavors, and meridian tropism. The results demonstrated that deep semantic representations outperformed traditional embedding methods in clustering compactness and inter-cluster separability. Under unsupervised conditions, the framework successfully identified functionally coherent CMM groups that showed strong consistency with classical CMM theory as well as modern prescription compatibility patterns. These findings suggested that the implicit semantic structure underlying CMM properties was computationally tractable and was effectively captured using LLM, providing a novel technical paradigm for the digital modeling of CMMs knowledge. From an application perspective, the functionally coherent clusters derived from semantic representations may serve as structural units for prescription optimization and novel formula design, offering methodological support for intelligent decision-making and knowledge discovery in CMMs research and practice.

Fundings

Frontier Technologies Research and Development Program of Jiangsu (BF2025076), Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (CI2021B002), and National Natural Science Foundation of China (82575255).

Acknowledgements

The authors extend appreciation to China Academy of Chinese Medical Sciences and Nanjing University of Chinese Medicine.

Author contributions

Jiayi He: conceptualization, data curation, formal analysis, and writing – original draft. Jiadong Xie: methodology,

validation, and writing – original draft. Kongfa Hu: investigation, supervision, and writing – review & editing. Haiyan Li: project administration, funding acquisition, and writing – review & editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

The authors declare no conflict of interest.

References

- [1] XUE XJ, LIU C, WANG YM, et al. Status quo and development strategy of traditional Chinese medicine in the new era. *Strategic Study of CAE*, 2023, 25(5): 11–20.
- [2] CHENG HB, ZHANG L, FU Y, et al. 2023 major scientific issues, engineering challenges and industrial technology problems. *Journal of Traditional Chinese Medicine*, 2023, 64(14): 1405–1421.
- [3] ZHANG LL, ZHOU XP. Discussion on the heterogeneous phase system from the perspective of seven emotions and the compatibility theory of monarch, minister, assistant and guide. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2022, 37(2): 655–658.
- [4] SHAO YY, ZHANG CN, WANG J, et al. Current status and development considerations of detoxification research on toxic traditional Chinese medicine compatibility. *Chinese Traditional and Herbal Drugs*, 2025, 56(4): 1394–1403.
- [5] DAN WC, ZHAO GZ, HE QY, et al. Analysis and prospects of common problems in clinical data mining of traditional Chinese medicine prescriptions. *China Journal of Chinese Materia Medica*, 2023, 48(17): 4812–4818.
- [6] HAO YL, CHEN MR, LIU XH. Understanding theory of property of Chinese materia medica based on Xiang thinking. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2020, 35(3): 1230–1232.
- [7] TANG YC, HAO WY, LI P. Discussion on meaning and classification of Ruanjian Sanjie Drugs from angle of five flavors in Chinese medicines. *Journal of Basic Chinese Medicine*, 2023, 29(3): 450–451, 456.
- [8] CHEN LH, XIAO FL, HUANG SY, et al. Research ideas and innovation development trends of quality evaluation of Chinese materia medica. *Chinese Traditional and Herbal Drugs*, 2021, 52(9): 2541–2547.
- [9] ZOU HB. Quantitative-index establishment of multiple-dimension herbal characters of traditional Chinese compound formula and dual index ratio sequence analysis. *Chinese Traditional Patent Medicine*, 2012, 34(4): 714–718.
- [10] DENG L, DING CS, HUANG XD, et al. Quantitative study on medicinal properties of traditional Chinese medicine based on BP neural network. *Chinese Traditional and Herbal Drugs*, 2020, 51(16): 4277–4283.
- [11] JIA RH, WEI GH, ZHAO WH, et al. Study on the prediction of cold-hot-neutral natures of Chinese herbs compounds based on K-nearest neighbor algorithm. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2023, 38(4): 1522–1525.
- [12] ZHANG WB, DANG HB, BAO ZS, et al. LAMGCN: traditional Chinese medicine herb recommendation via LSTMs with attention mechanisms and graph convolutional networks. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025, 24(3): 1–22.
- [13] MIN BN, ROSS H, SULEM E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Computing Surveys*, 2024, 56(2): 1–40.
- [14] YANG X, CHEN AK, POURNEJATIAN N, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 2022, 5(1): 194.
- [15] TIAN HY, YANG K, DONG X, et al. TCMLLM-PR: evaluation of large language models for prescription recommendation in traditional Chinese medicine. *Digital Chinese Medicine*, 2024, 7(4): 343–355.
- [16] WANG L, MA YY, BI WS, et al. An entity extraction pipeline for medical text records using large language models: analytical study. *Journal of Medical Internet Research*, 2024, 26: e54580.
- [17] LI TH, SHETTY S, KAMATH A, et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *NPJ Digital Medicine*, 2024, 7(1): 40.
- [18] XU M, SHAO XY, GUO DG, et al. Phenomics of traditional Chinese medicine 2.0: the integration with digital medicine. *Digital Chinese Medicine*, 2025, 8(3): 282–299.
- [19] YUAN SS, ZHOU ZC, JIN XY, et al. Enhancing herbal medicine-drug interaction prediction using large language models. *IEEE Journal of Biomedical and Health Informatics*, 2025, 29(10): 6971–6981.
- [20] HE JY, XIE JD, HU CJ, et al. Research on clustering method for Chinese herbal medicine based on graph neural network. *Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology*, 2024, 26(11): 2988–2995.
- [21] DUAN PB, YANG K, SU X, et al. HTINet2: herb–target prediction via knowledge graph embedding and residual-like graph neural network. *Briefings in Bioinformatics*, 2024, 25(5): bbae414.
- [22] ZHONG GS. *Chinese Materia Medica*. Beijing: China Traditional Chinese Medicine Press, 2016.
- [23] CHEN SG. *Waike Zhengzong (Orthodox Manual of External Medicine)*. Beijing: China Traditional Chinese Medicine Press, 2007.
- [24] GAO BJ. *Yangke Xinde Ji (Collected Insights on Ulcer Medicine)*. Beijing: People's Medical Publishing House, 2006.
- [25] YANG YF, TANG LL, SUN LY, et al. Guidelines for psychological rehabilitation intervention combined integrated traditional Chinese and Western medicine after radical surgery for early and middle stage colorectal cancer. *Chinese Journal of Experimental Traditional Medical Formulae*, 2023, 29(21): 10–23.
- [26] Pharmacopoeia Commission of the People's Republic of China. *Pharmacopoeia of the People's Republic of China*. Beijing: China Medical Science Press, 2025.
- [27] MURPHY KP. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT press, 2012.
- [28] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013, 26. doi: 10.48550/arXiv.1310.4546.
- [29] DEVLIN J, CHANG MW, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American*

- Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171–4186.
- [30] XIAO ST, LIU Z, ZHANG PT, et al. C-pack: packed resources for general Chinese embeddings. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024: 641–649.
 - [31] ZHANG Y, LI M, LONG D, et al. Qwen3 embedding: advancing text embedding and reranking through foundation models. *arXiv*, 2026. doi: 10.48550/arXiv.2506.05176.
 - [32] XU R, WUNSCH D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 2005, 16(3): 645–678.
 - [33] QU XL, TIAN ZW, CUI JM, et al. A review of knowledge graph in traditional Chinese medicine: analysis, construction, application and prospects. *Computers, Materials & Continua*, 2024, 81(3): 3583–3616.
 - [34] ZHANG WC, CHENG HB, SHEN WX, et al. Effect of Xianlian Jiedu Formula on intestinal immune cells in colorectal cancer mice with dampness-heat stasis toxin syndrome based on single-cell transcriptome sequencing. *Journal of Traditional Chinese Medicine*, 2023, 64(12): 1263–1272.
 - [35] LIU JR, HUANG M, FAN MM, et al. Shenbai Jiedu Prescription inhibits proliferation of colorectal cancer cells by regulating pten/PI3K/AKT signaling pathway. *Chinese Journal of Experimental Traditional Medical Formula*, 2022, 28(14): 36–43.
 - [36] YUAN WC, ZHOU HG, LI WT, et al. Current status, advantages and strategies of TCM in the prevention and treatment of malignant tumors. *Journal of Nanjing University of Traditional Chinese Medicine*, 2025, 41(6): 707–720.
 - [37] XU YR, TAO LHP, QIAN JY, et al. Efficacy mechanism of Xianlian Jiedu Prescription against colorectal cancer recurrence via regulating angiogenesis. *Chinese Journal of Experimental Traditional Medical Formulae*, 2025, 31(6): 79–87.
 - [38] HUANG M, FAN MM, CHENG HB, et al. Inhibition effects of Shenbai Jiedu Formula on proliferation and migration of colorectal cancer cells based on Wnt/ β -catenin signaling pathway. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2023, 38(2): 789–794.
 - [39] HUA R, DONG X, WEI Y, et al. Lingdan: enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models. *Journal of the American Medical Informatics Association*, 2024, 31(9): 2019–2029.
 - [40] WANG XY, SUN XH, YANG L, et al. Reinforcement learning for LLM-based explainable TCM prescription recommendation with implicit preferences from small language models. *Chinese Medicine*, 2025, 20: 193.
 - [41] YANG ZJ, YIN YJ, KONG CJ, et al. ShennongAlpha: an AI-driven sharing and collaboration platform for intelligent curation, acquisition, and translation of natural medicinal material knowledge. *Cell Discovery*, 2025, 11(1): 32.
 - [42] GAO K, LIU L, LEI SS, et al. HERB 2.0: an updated database integrating clinical and experimental evidence for traditional Chinese medicine. *Nucleic Acids Research*, 2025, 53(D1): D1404–D1414.
 - [43] WU Y, ZHANG FL, YANG K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Research*, 2019, 47(D1): D1110–D1117.
 - [44] ZHANG S, TONG HH, XU JJ, et al. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 2019, 6(1): 11.
 - [45] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020, 33: 9459–9474.

(Editor-in-Charge Siyi Wei)

CMM-EmbedCluster: 一种融合大语言模型与中药药性理论的中药聚类框架

何佳怡^{a,b}, 谢佳东^b, 胡孔法^{b,c,d}, 李海燕^{a*}

a. 中国中医科学院中医药信息研究所, 北京 100700, 中国

b. 南京中医药大学人工智能与信息技术学院, 江苏南京 210023, 中国

c. 南京中医药大学江苏省中医药防治肿瘤协同创新中心, 江苏南京 210023, 中国

d. 南京中医药大学江苏省智慧中医药健康服务工程研究中心, 江苏南京 210023, 中国

【摘要】目的 本研究提出一种基于大语言模型的中药聚类框架, 旨在从中药药性理论的语义层面挖掘中药间的潜在配伍规律。**方法** 首先, 基于《中药学》建立包含四气、五味及归经的 567 个中药的药性知识数据库。其次, 基于《外科正宗》和《疡科心得集》中记载的 10 首脏毒相关方剂, 提取其中 49 味中药作为实验数据集, 采用 One-Hot、Word2Vec、双向 Transformer 编码表示 (BERT)、北京智源人工智能研究院通用嵌入 (BGE)、Qwen 五种语义表示方法对药性知识进行向量化表示。最后, 利用 t 分布随机近邻嵌入 (t-SNE) 算法对高维语义向量进行非线性降维, 并基于 k -means 算法 ($k=7$) 完成聚类分析。采用轮廓系数 (SS)、戴维斯-博尔丁指数 (DBI) 和卡林斯基-哈拉巴斯指数 (CHI) 对聚类性能进行评价。**结果** 基于 Qwen 的聚类方法 CMM-EmbedCluster 取得了最高的 SS (0.6074) 和 CHI (158.0572), 以及最低的 DBI (0.4995), 表明其在类间分离度和类内紧密性方面优于其他方法。中药聚类结果的可视化显示, 各聚类在低维空间中具有良好的分离性, 类间区分度较高, 且类内中药在功能属性上表现出较高一致性。进一步的聚类结果可解释性分析表明, 不同聚类在四气、五味及归经等方面呈现出稳定的结构性差异, 形成与中药药性理论相一致的功能分区特征。**结论** CMM-EmbedCluster 有效融合大语言模型与中药药性理论, 实现中药的语义级表示与聚类, 为从中药药性语义角度探索中药间潜在配伍规律提供了支持。

【关键词】 大语言模型; 中药药性理论; 中药聚类; 语义表征; k -means 聚类; CMM-EmbedCluster