



Article

# Dynamic Seed Topic Construction and LLM-Driven Multi-Topic Identification for Social Q&A Platforms

Ying Zhao <sup>1,\*</sup>, Xiurui Yang <sup>1</sup> , Tian Qiang <sup>1</sup> and Luoming Liang <sup>2</sup>

<sup>1</sup> School of Public Administration, Sichuan University, Chengdu 610065, China

<sup>2</sup> Business School, Sichuan University, Chengdu 610065, China

\* Correspondence: zhaoying@scu.edu.cn

## Abstract

Social Q&A platforms produce short, noisy, and highly diverse user questions, making coarse topic labels insufficient for accurate information organization. This study proposes a dynamic seed topic construction and multi-demand topic identification framework for health science popularization questions on the Zhihu platform. Using 3529 cleaned questions derived from 2011 to 2024, the framework combines BERTopic clustering, LLM-based topic naming, and a TOP-K cross-filtering update strategy that integrates semantic similarity and frequency. The LLM then performs topic identification, optimization, and assignment, and a co-occurrence network analyzes topic associations. The method generates 26 initial seed topics, which expand to 51 topic instances and are subsequently optimized to 36 topics. Compared with LDA, BERTopic, and LLM-based baselines, our framework achieved the best topic coherence and the lowest topic similarity while maintaining good topic diversity. Human evaluation by 17 volunteers on 30 questions showed that the model performed at least as well as user-assigned tags across relevance, comprehensiveness, clarity, accuracy, and readability, with clearer advantages in relevance and accuracy. These results suggest that the proposed framework can identify fine-grained, interpretable, and strongly associated information-demand topics for social Q&A platforms.

**Keywords:** social Q&A platforms; topic modeling; large language models; dynamic seed topics; BERTopic

## 1. Introduction

Social question-and-answer (Q&A) platforms now accumulate large volumes of user-generated content (UGC), and the fragmentation of this content makes information organization difficult [1,2]. Users often post short and open-ended questions that contain diverse information demands from daily life [3]. This problem is more salient in healthcare contexts. Users are increasingly expressing their need for health information through mobile applications and online platforms, where they seek information, emotional support, and practical assistance beyond traditional medical care [4,5]. On these platforms, topic aggregation mainly relies on simple topic classification and public topic labels.

From the perspective of information foraging theory, users seek information by following cues that are expected to reduce the uncertainty associated with a broader information goal [6]. In social Q&A settings, a complex health concern may therefore be expressed through several related micro-demands, such as symptoms, diet, treatment, and healthcare access, within a single question. This characteristic makes flat, single-topic classification insufficient and motivates the identification of multiple, associated demand topics.



Academic Editor: Ting Chi

Received: 18 May 2026

Revised: 20 July 2026

Accepted: 28 July 2026

Published: 7 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Consequently, accurately identifying fine-grained, diverse information demand topics has emerged as a pressing practical challenge for optimizing topic-based information resource aggregation and enhancing information service efficiency [7].

However, existing topic identification methods have limitations in this setting. Traditional static topic models, including LDA [8] and other single clustering approaches, are mainly designed to extract global topics from large-scale corpora [9]. When applied to short questions on social Q&A platforms, they tend to generate coarse topics, miss secondary or long-tail demands within a single question, and provide limited support for multi-topic interpretation. In addition, social Q&A data often lack standardized labels, prior topic knowledge, or benchmark datasets. Under these conditions, models are more likely to encounter cold-start problems, unstable topic boundaries, and weak interpretability, which reduces consistency across topic recognition and classification tasks [10].

Recent advances in large language models (LLMs) offer a useful technical basis for this task [11]. Compared with traditional models, LLMs can make better use of contextual semantics and generate readable topic labels, which helps improve the relevance and interpretability of topic identification. This makes them suitable for recognizing multiple topics in short, open-ended user questions. However, LLM-based topic generation remains strongly shaped by the input corpus and prompt constraints. When clear external guidance is absent, the model may produce redundant labels, unstable granularity, or topics with weak semantic relevance. These outputs deviate from the intended research categories and are difficult to apply directly in practical contexts [12,13].

To address these limitations, we propose an integrated framework for dynamic seed topic construction and multi-demand topic identification. The framework automatically identifies users' multi-demand topics in specific domains of social Q&A platforms under conditions of insufficient prior knowledge. It first constructs representative and interpretable initial seed topics by combining BERTopic's clustering strengths with large language models' semantic naming capabilities. A TOP-K cross-filtering strategy then enables dynamic seed topic updates that balance quantitative frequency and semantic similarity. Guided by these dynamic seeds and carefully designed prompt templates, large language models perform integrated topic recognition, optimization, and allocation in a single workflow. The framework further constructs a fine-grained demand topic co-occurrence network to reveal key demand topics and their complex correlation characteristics. In doing so, it advances topic modeling in two key respects: (1) a dynamic seed topic construction strategy that effectively tackles the absence of standard datasets and prior knowledge while generating high-quality, research-aligned topics applicable to other modeling tasks; (2) an automatic, integrated multi-demand topic recognition method driven by large language models that accounts for short-text and diverse characteristics, improves task consistency, and reduces research costs. The identified fine-grained, interpretable, and strongly correlated topics ultimately provide actionable guidance for optimizing platform topic classification and information organization strategies.

## 2. Literature Review

### 2.1. Current Research Status of Topic Identification

Topic identification, also known as topic modeling, is a technique that mines latent topics from unstructured document collections based on semantic structures and content features [8]. It aims to automatically extract semantically consistent topic structures from large-scale text data, thereby providing important support for text organization, information retrieval, and knowledge discovery. Existing topic identification methods can generally be divided into manual identification, text-mining-based identification, and hybrid identification [14,15]. Manual methods include content analysis, manual annotation,

and conceptual mapping. The advantage of this approach lies in its ability to extract topics with fine granularity and high interpretability; however, it suffers from low efficiency, high costs, and a lack of generalizability when confronted with large-scale data. Technology-based identification grounded in text mining mainly includes methods such as keyword extraction, word-frequency analysis, and text clustering. Early research on technology-based identification relied primarily on statistical and probabilistic graphical models, such as Latent Dirichlet Allocation (LDA) and its variants, which inferred latent topic distributions through word co-occurrence statistics [8]. Although such methods exhibit strong interpretability in topic identification, they perform poorly when handling short texts, highly sparse data, and semantically complex tasks. Consequently, in the era of deep learning, especially after the rise of large-scale pre-trained models, topic recognition methods based on text representation and clustering algorithms have gradually gained popularity. Grootendorst proposed the BERTopic model, which employs Sentence-BERT for text representation and selects topic words within clusters according to c-TF-IDF weights to achieve topic representation; its document embeddings enhance the semantic representational capacity of topic clustering [16]. Compared with manual identification methods, text-mining approaches are weaker in terms of interpretability and semantic consistency. To compensate for the deficiencies of both manual identification and text-mining methods, some scholars have adopted hybrid strategies combining technological identification with manual intervention to improve the accuracy and reliability of topic identification.

In summary, although existing studies have conducted extensive work on topic identification, the identification of multiple topics in social question-and-answer platforms still faces three major challenges. First, topic identification granularity remains coarse. Traditional text-mining-based technological identification methods typically capture only the most prominent topics within documents, while secondary topics carrying semantic information are often overlooked, resulting in coarse topic granularity [17]. Second, topic interpretability is insufficient. Although manual identification methods can provide topics with strong interpretability, they suffer from high subjectivity and significant resource consumption when processing large-scale social Q&A data. Meanwhile, technological identification methods supported by machine learning and deep learning, despite their efficiency advantages, still exhibit notable deficiencies in topic interpretability [18]. Third, the consistency of multiple-topic identification tasks is inadequate. In scenarios such as social Q&A platforms, which lack standard data and prior knowledge, capturing the multidimensional semantic information implicit in texts and identifying the diverse topics underlying individual documents generally requires a sequential process of identification followed by classification. However, current methods often design the identification and classification stages independently, lacking semantic coherence between tasks; this leads to semantic deviations between the identified topics and subsequent classification results. Such disconnection between tasks not only weakens the semantic consistency of topics across stages but may also trigger topic drift or classification mismatches [19]. Therefore, identifying fine-grained, interpretable, and strongly associated multiple topics in the absence of predefined datasets and prior knowledge has become an urgent challenge.

## 2.2. LLM-Based Topic Identification Research

Traditional topic modeling methods rely on the bag-of-words assumption and probabilistic distributions. Although they perform relatively well on static corpora, their results often lack strong interpretability and struggle to adapt to complex scenarios such as short texts and highly noisy data. By comparison, large language models (LLMs), trained on large-scale corpora, show stronger capacities for contextual understanding and natural language generation [20]. LLMs can directly identify and generate topics from corpora

without depending on keywords or complex probabilistic inference, thereby rendering topic identification results more interpretable and semantically coherent and offering new opportunities to address the shortcomings of traditional topic identification methods.

Current LLM-based topic identification research is mainly organized around prompt engineering. A typical workflow uses LLMs for topic generation, refinement, representation, and evaluation. PromptTopic, for example, applies LLMs at several stages to extract, organize, and refine topics and combines PBM and WSM during optimization to alleviate sparsity and limited context in short-text topic modeling [21]. Nevertheless, the illusion problem of LLMs generating topics that are inconsistent with expectations or contain factual errors has become a key bottleneck limiting their reliability [22]. To address this issue and improve the quality of topic identification, existing research has primarily explored three strategies. The first is optimizing the granularity of prompts. By incorporating prompt elements such as seed topics or topic descriptions, the model is guided to eliminate hallucinations. Studies have indicated that, compared with detailed granularity descriptions, seed topics are more efficient in guiding LLMs to generate target topics; owing to their low cost, interpretability, and scalability, seed topics are particularly suitable for scenarios lacking labeled data and computational resources. The second is model fine-tuning. By adjusting model parameters on domain-specific data, LLMs acquire specialized, high-precision topic generation capabilities, thereby effectively reducing the number of generated topics [23]. Although this method can effectively mitigate topic deviation, it depends on sufficient labeled data and computational resources and suffers from inherent drawbacks such as high resource consumption and limited generalization capability. The third is retrieval-augmented generation (RAG). By introducing external knowledge bases, RAG supplies LLMs with real-time and accurate factual grounding, thereby enhancing the reliability, interpretability, and domain adaptability of topic identification [24]. RAG is well suited to context requiring dynamic knowledge augmentation and strong domain adaptation, with its advantages being especially pronounced when labeled data is absent but high-quality knowledge bases are available.

When applying the aforementioned strategies to specific contexts, trade-offs must be made in light of contextual characteristics and resource constraints. With respect to social question-and-answer platforms, platform content generally exhibits features such as short texts, multiple noise points, and diverse topics; user questions are characterized by non-professionalization, topic overlap, and dynamic evolution. These features collectively give rise to the dual challenges of insufficient prior knowledge and a lack of standard labeled datasets in the platform's topic identification tasks. In this context, the model fine-tuning strategy is difficult to apply owing to its strong dependence on data and resources. By comparison, the seed topic strategy can provide clear guidance at extremely low cost, assisting the model in identifying topics within complex and variable user questions; the semantic matching concept inspired by RAG, in turn, can effectively compensate for the platform's lack of prior knowledge by retrieving continuously updated external knowledge bases to guide the model, thereby significantly improving the accuracy and reliability of generated topics. Consequently, a combined strategy of seed topics and RAG-inspired strategy can achieve high-quality topic identification under resource-constrained conditions and may constitute a viable pathway for resolving topic identification on social question-and-answer platforms.

### 2.3. Research Related to Seed Topics

As a mechanism for supplying prior knowledge to models and guiding them to produce outcomes that meet expectations, the seed plays an important role in the field of topic identification. In traditional topic identification methods, seed words act as key points that

guide the formation of topics. By introducing a limited set of highly domain-specific lexical items, these methods help ensure that the inferred topics remain aligned with the pre-defined domain scope [25], thereby markedly improving topic clustering performance [26]. In LLM-based topic identification methods, although LLMs can directly generate coherent and interpretable topics from corpora, their powerful open-generation capability also introduces new challenges such as topics deviating from the domain, content redundancy, and hallucinations inconsistent with facts. To address this problem, the seed has been introduced as a prompt-constraint strategy to control the quality of LLM generation. This approach explicitly guides and constrains the scope and focus of generated topics by having users provide a set of initial seed topics, thereby helping the model more precisely capture identification intent and perform appropriate reasoning, ultimately producing more targeted topics.

However, existing research and applications of seed topics face two core challenges: seed topic initialization and dynamic updating. In initialization, the majority of studies rely on manually constructed sets or pre-defined labels associated with the dataset. This approach is difficult to implement in social question-and-answer platform scenarios characterized by insufficient prior knowledge and a lack of standard datasets. In terms of dynamic updating, once an initial seed topic set is established, it remains relatively fixed; with the emergence of new content, the initial seed topic set may fail to cover subsequently generated semantic spaces, resulting in incomplete topic identification coverage and semantic drift. In this context, endowing seed topics with dynamic adjustment capabilities is of critical importance [23]. How to enable initial seed topics to possess continuous expansion and self-evolution capabilities has become key to enhancing the comprehensiveness and accuracy of topic identification. To meet this challenge, the concept of dynamic seed topics has emerged. It refers to the automatic generation of domain-adaptive topic guidance words by dynamically capturing emerging concepts and semantic changes in the corpus during the topic identification process, followed by continuous updating of the initial seed topic set accordingly. Compared with traditional static seed topics, dynamic seed topics possess flexible matching and adaptive evolution capabilities, enabling the continuous optimization of the expression forms and coverage scope of seed topics throughout the topic identification process. On this basis, the present study focuses on the following core research question: in the context of social question-and-answer platforms lacking pre-defined datasets and prior knowledge, how to construct dynamic seed topics that can be automatically generated, continuously evolved, and effectively guide LLMs, thereby achieving precise identification of multiple topics within individual questions.

Recent prompt-based, seed-guided, and RAG-based topic modeling approaches provide important foundations but address different methodological objectives. TopicGPT focuses on prompt-based topic generation and refinement [12], while seed-guided topic discovery methods generally rely on user-provided seeds and multiple contextual signals to identify topic-indicative terms [27]. RAG-based topic modeling further retrieves external evidence to improve the reliability and interpretability of LLM-generated topics [24]. In contrast, the present study focuses on maintaining a domain-adaptive seed-topic set for unlabeled and evolving short-text corpora. Its novelty lies in treating the evolving seed-topic set itself as the retrieval target and retaining the union of frequency-based and semantic-similarity-based TOP-K candidates, thereby balancing globally prevalent topics with low-frequency topics that remain highly relevant to the current question.

### 3. Dataset

#### 3.1. Data Source

This paper used a Python 3.9 web crawler to collect targeted data on the topic of “health science popularization” from the Zhihu platform [28]. To discover core user needs from a large amount of text data, this paper collects discussion and key Q&A information under related topics on Zhihu. For each Q&A details page, the collected data include the question, question details, topic tags, number of followers, and number of views. After cleaning and deduplicating the data, we obtained 3529 questions related to trending “health science popularization” topics on Zhihu, covering January 2011 to December 2024.

In social Q&A platforms, answers are mostly short texts. The questions and custom hashtags directly reflect users’ health information needs. Therefore, this paper aggregates the self-created hashtags and question text from each Q&A, maximizing the retention of potential user needs while expanding the text length. Since data quality directly affects the outcome of topic modeling, we constructed appropriate user-defined dictionaries and custom stop-word dictionaries during data preprocessing. To effectively extract health knowledge units from user Q&A data, this study introduces the Chunyu Doctor Disease Topic Table and the Tsinghua University Medical Glossary when constructing the custom dictionary. Chunyu Doctor, as a professional medical and health platform, has a disease topic knowledge base reviewed by medical experts, covering common clinical disease symptoms and medical terms, providing a reference for effectively identifying professional health knowledge units in unstructured text. This paper uses web crawling to extract professional terms from web pages to construct the disease topic table. To more comprehensively capture the semantics of user questions and their core needs in social Q&A platforms, this paper also introduces the Tsinghua Medical Dictionary, which contains standardized medical terms, drug names, and commonly used health-related vocabulary. Combining the disease topic table with the Tsinghua Medical Dictionary can improve the accuracy of corpus cleaning.

#### 3.2. Data Preprocessing

The process of word segmentation for aggregated text data in this paper is as follows: (1) Clean the data, remove abnormal strings and emoticons from comments, remove meaningless letters and numbers. (2) Inspect high-frequency and low-frequency words in the segmented data and add noisy or invalid terms to the stop-word list based on the Harbin Institute of Technology stop words. For example, under the topic of health science popularization, “health”, “health science popularization”, and “health science popularization knowledge” are commonly used tags and words but lack practical meaning, so word classes are added to the stop word points. (3) Observe the text data and improve the user dictionary based on Spring Rain Doctor (<https://www.chunyuyisheng.com/>, accessed on 13 March 2026) and Tsinghua Medical Dictionary (Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, and Maosong Sun, THUOCL: Tsinghua University Open Chinese Lexicon, 2016, <http://thuocl.thunlp.org/>, accessed on 13 March 2026); (4) Use the Chinese natural language processing tool Jieba to segment the document and remove stop words, and input the cleaned text into the next stage.

#### 3.3. Analytical Framework

The overall analytical process is shown in Figure 1. First, the initial seed topics are generated and dynamically updated based on topic frequency and semantic similarity. The updated seed-topic set is then used to support topic identification, optimization, assignment, and performance evaluation. In addition, the identified topics are used to construct a co-occurrence network and examine the structural relationships among user demands.

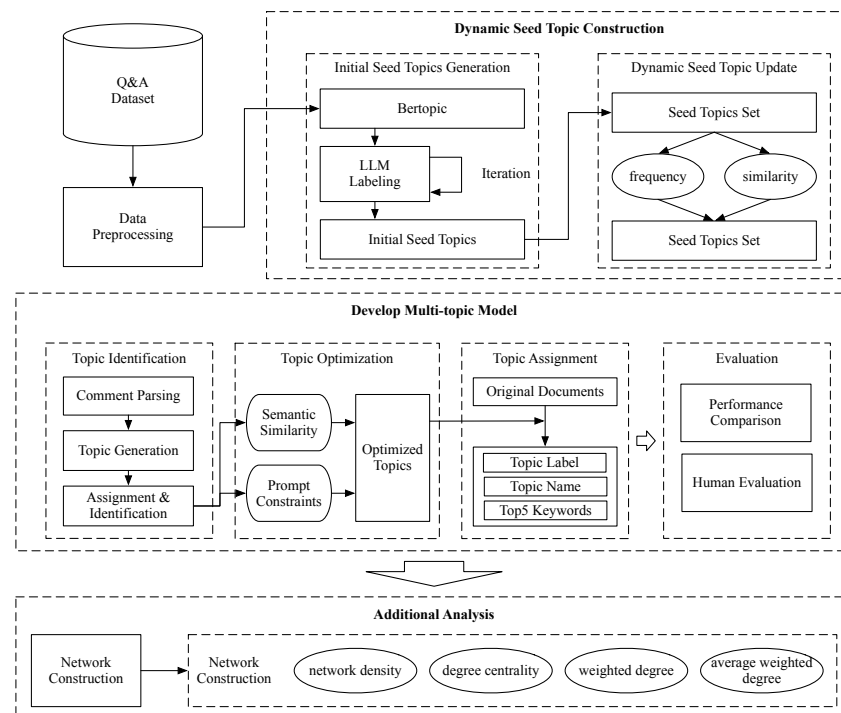


**Figure 1.** Research analytical framework.

## 4. Dynamic Seed Topic Construction

In social question-and-answer platforms, a deep analysis of user demand topics is of paramount importance, as it provides direct reference bases for the organization of information content and the adjustment of topic classification strategies. Traditional text mining approaches (such as the LDA topic model and single-text clustering) excel at identifying diverse global topics, but exhibit limitations of coarse topic granularity, difficulties in multi-classification of individual documents, and insufficient semantic representation capabilities when addressing the diverse characteristics of user information demands on social Q&A platforms [14]. Large language models (LLMs) possess powerful semantic understanding and summarization capabilities and have been widely applied to complex text tasks. However, in niche domains lacking standard datasets and prior knowledge, LLM-based topic modeling via prompt templates is prone to machine hallucinations due to excessive degrees of freedom or to generating fragmented topics that deviate from research expectations [22].

To address these challenges and better control topic granularity, this chapter proposes a Dynamic Seed Topic construction method. This method integrates the global topic clustering capabilities of BERTopic with the deep semantic understanding capabilities of large language models to automatically extract, name, and dynamically update seed topics from massive unlabeled social Q&A data, thereby laying the foundation for the subsequent precise identification of multiple demand topics within individual documents.

### 4.1. Conceptual Definition and Overall Framework of Dynamic Seed Topics

In natural language processing and topic modeling, “seed” constitutes a means of supplying prior knowledge to the model [25]. Prior studies have frequently employed expert-predefined seed words as anchors to guide the model toward outcomes consistent with human cognition [29]. However, topic hotspots on social Q&A platforms evolve rapidly over time. Reliance on manually constructed static topic label repositories not only incurs high costs but also struggles to adapt to the dynamic evolution of content.

Manually constructed static topic label repositories are therefore costly and fail to adapt to the dynamic evolution of the content.

To construct a flexible and representative external knowledge base under constraints of absent prior knowledge, this paper operationalizes the definition of Dynamic Seed Topic as: an initial topic template automatically generated according to the specific context and temporal distribution of the objective dataset and continuously subjected to semantic updates during the modeling process. In LLM topic modeling tasks, it functions as a supervisory or constraining condition that not only resolves cold-start and hallucination divergence problems confronting large models but also dynamically captures the evolutionary trajectories of topics over time.

From the perspective of the overall framework, Dynamic Seed Topic construction primarily comprises two progressive stages. The first stage is initial seed topic generation: clustering techniques are utilized to obtain a panoramic view of topics within the specific domain, followed by semantically clear and readable naming by the LLM; the second stage is Dynamic Seed Topic updating: To address the issue of fine-grained topic redundancy generated by large language models, this paper is inspired by retrieval-augmented generation (RAG) and references its semantic matching concept to achieve efficient cross-filtering of the seed set and dynamic retrieval updates [24].

#### 4.2. Generation Mechanism and Empirical Application of Initial Seed Topics

Confronted with large volumes of irregular, colloquial short texts in online communities, user questions constitute the direct expression of information needs. Extracting seed topics from objective user questions enables maximal alignment with authentic demand scenarios.

##### 4.2.1. Theory and Method

This paper first adopts BERTopic to perform global clustering on the question content. The complete technical workflow of BERTopic encompasses three core steps: first, a pre-trained language model is employed to extract embedding vector representations of the texts; second, the UMAP algorithm is applied for dimensionality reduction in the high-dimensional vectors, combined with the HDBSCAN algorithm for density-based clustering on the reduced representations; finally, class-based term frequency-inverse document frequency (c-TF-IDF) is used to extract the feature words of each topic cluster [16].

When evaluating clustering quality, topic diversity (Topic Diversity, TD) and topic coherence (Topic Coherence, TC) serve as two conventional metrics [21].

Topic diversity measures the differentiation and coverage among distinct topics [30]. A topic model with high diversity can capture the diverse topics present in the text data while avoiding the generation of repetitive or redundant topics. The calculation formula for topic diversity is shown in Equation (1). Here,  $N$  denotes the top  $N$  topic words selected for each topic, and  $K$  denotes the total number of topics.  $\bigcup_{i=1}^K$  represents the union operation performed on the sets of all topics, and  $\{w_1^i, w_2^i, \dots, w_N^i\}$  denotes the top  $N$  words of the  $i$ -th topic. The diversity score TD ranges from 0 to 1; values of TD closer to 1 indicate high differentiation among topics, whereas values approaching 0 indicate stronger redundancy between topics

$$TD = \frac{\left| \bigcup_{i=1}^K \{w_1^i, w_2^i, \dots, w_N^i\} \right|}{N \times K} \quad (1)$$

Topic coherence measures the semantic consistency and logical associations among keywords within a single topic.  $C_v$  (coefficient of variance) is one of the most commonly employed measures of topic coherence. This method performs topic-word segmentation via a sliding window and computes topic coherence using normalized pointwise

mutual information together with cosine similarity [31].  $C_v$  integrates normalized point-wise mutual information (NPMI) and cosine similarity to evaluate topic coherence more comprehensively and is well suited to short-text data. Its calculation formula is shown in Equation (2), where  $\text{cosine\_similarity}(v_i, v_j)$  represents the cosine similarity of vectors between different words:

$$C_v = \frac{1}{N(N-1)/2} \sum_{i=2}^N \sum_{j=1}^{i-1} \text{NPMI}(w_i, w_j) \cdot \text{cosine\_similarity}(v_i, v_j) \quad (2)$$

Because this paper seeks to identify multiple topics and their interrelationships, it is essential to detect topics of comparatively fine granularity that exhibit strong differentiation. Furthermore, given that keyword co-occurrence relationships are not pronounced in the short texts generated by social users, this paper, after comprehensive consideration, selects  $C_v$  as the evaluation metric for BERTopic clustering performance in order to effectively acquire preliminary topic clusters that lay the foundation for subsequent research.

However, the topic clusters produced by BERTopic are typically manifested solely as probability distributions of groups of feature words, with relatively ambiguous semantics; direct use for constraining large models would result in unstable connotations. Accordingly, this paper designs LLM-based prompt engineering for topic naming (see Table 1). The prompt template comprises four key elements: role (Role, defining the model as a topic annotator in a specific domain), objective (Objective, requiring the model to judge and generate topic labels), process (Process, creating a label if none exists or judging whether modification is required based on representative documents if a label already exists), and constraints (Constraints, preventing labels from being overly restrictive). Through this iteratively optimized naming calibration strategy, groups of feature words are transformed into high-quality, short-text topic labels that conform to human cognition.

**Table 1.** Seed topic label generation prompt.

| Prompt Engineering Template (for Generating Seed Topic Labels)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>[Role]</b><br>You are a topic annotation researcher in the field of {domain}                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>[Objective]</b><br>You will receive a set of {topic representative documents}, {topic feature words}, and {initial topic label: topic explanation}; your task is to determine whether the initial topic label requires updating.                                                                                                                                                                                                                                                                                                                                 |
| <b>Process</b><br>1. If the initial topic label is empty, please create a topic label for this topic.<br>2. If the initial topic label is not empty, please judge, on the basis of the {topic representative documents} and {topic feature words}, whether a new topic label must be created. <ul style="list-style-type: none"> <li>– If no new topic label needs to be created, return the initial topic label and topic explanation.</li> <li>– If a new topic label needs to be created, return the newly created topic label and topic explanation.</li> </ul> |
| <b>[Constraints]</b><br>1. The topic label must not be confined to a specific comment.<br>2. Any newly created topic label should constitute an optimization of the initial topic label.                                                                                                                                                                                                                                                                                                                                                                            |
| <b>[Input]</b><br>{Topic representative documents}<br>{Topic feature words}<br>{initial topic label}                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <b>[Output]</b><br>{Topic label: Topic explanation}                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

We selected the GLM-4-Plus large language model, which is the latest high-intelligence flagship large language model launched by Zhipu AI. It has made breakthroughs in language understanding and long text processing, and adopts innovative technologies such as

PPO to significantly improve reasoning and instruction following capabilities, making it better suited to complex natural language processing tasks.

#### 4.2.2. Empirical Implementation in the Zhihu Context

To validate the foregoing mechanism, this paper selects the “Health Science Popularization” topic on the Zhihu platform as the experimental data source. In the BERTopic text embedding stage, the paraphrase-multilingual-MiniLM-L12-v2 model, which possesses multilingual advantages, was chosen. We also account for the distribution of document counts when comparing BERTopic results. When document counts differ markedly and the number of topics is limited, some topics have very few documents (forming a long-tail distribution), while others have very many documents. Such an imbalance can adversely affect clustering quality (for example, small clusters may be absorbed by large ones). Consequently, this paper selected clustering results exhibiting a more balanced document-count distribution. After multiple rounds of parameter tuning, the below core clustering parameters were ultimately determined (see Table 2):

**Table 2.** BERTopic model training parameters.

| Parameter Name          | Parameter Interpretation                                 | Parameter Value |
|-------------------------|----------------------------------------------------------|-----------------|
| n_neighbors             | Number of neighboring points determined by UMAP          | 30              |
| n_components            | Dimensionality size after reduction                      | 5               |
| min_dist                | Minimum effective distance between embedding points      | 0               |
| min_cluster_size        | Minimum cluster size                                     | 30              |
| top_n_words             | Number of words extracted per Topic                      | 10              |
| calculate_probabilities | Calculate probabilities for all topics for each document | True            |

Under this configuration, the clustering model successfully identifies 26 topics and attains a high-level  $C_v$  coherence score of 0.764. Subsequently, the topic cluster feature words and representative documents output by BERTopic were fed into the large language model for naming. The initial seed topic set for Zhihu health science popularization ultimately constructed is shown in Table 3. The extraction results not only precisely accurately cover mainstream explicit demands such as “Traditional Chinese Medicine and Health” (269 documents), “Diet and Healthy Foods” (229 documents), and “Fitness and Exercise” (216 documents). In addition, the results retain a substantial volume of noise documents that were not clearly assigned by traditional clustering (marked as “-1 class”, totaling 1164 documents). These noise documents likewise carry highly divergent latent semantic demands from users and possess non-negligible mining value in the subsequent fine-grained identification performed by the large language model.

**Table 3.** Details of initial seed topics.

| LLM Topic Naming                        | Number of Documents | Topic Feature Words                                                             |
|-----------------------------------------|---------------------|---------------------------------------------------------------------------------|
| -1 Class                                | 1164                | -1_medicine_health_physical health_body                                         |
| Traditional Chinese Medicine and Health | 269                 | 0_traditional Chinese Medicine_staying up late_Chinese herbal medicine_medicine |
| Diet and Healthy Foods                  | 229                 | 1_diet_healthy diet_food_foodstuffs                                             |
| Fitness and Exercise                    | 216                 | 2_fitness_sports_weight loss_training                                           |
| Sleep and Health                        | 137                 | 3_sleep_insomnia_staying up late_sleep quality                                  |
| Cancer Prevention and Treatment         | 131                 | 4_cancer_liver cancer_tumor_lung cancer                                         |
| Lifestyle Habits and Physical Health    | 107                 | 5_physical health_life and health_life_lifestyle habits                         |
| Diabetes and Blood Sugar Management     | 105                 | 6_diabetes_blood sugar_hypertension_high blood sugar                            |
| Female Reproductive Health              | 104                 | 7_female health_female_pregnancy_premature ovarian failure                      |

**Table 3.** *Cont.*

| LLM Topic Naming                             | Number of Documents | Topic Feature Words                                                                  |
|----------------------------------------------|---------------------|--------------------------------------------------------------------------------------|
| Beverages and Health                         | 101                 | 8_beverages_drinking water_drinking alcohol_alcohol consumption                      |
| Mental Health                                | 94                  | 9_mental health_psychology_psychology_depression                                     |
| Subhealth Management                         | 85                  | 10_health management_life and health_health_subhealth                                |
| Medical Consultation and Healthcare Services | 85                  | 11_doctor_medicine_hospital_physical examination                                     |
| Skeletal Health                              | 71                  | 12_cervical vertebra_lumbar vertebra_pain_lumbar disc herniation                     |
| Respiratory Diseases                         | 62                  | 13_cough_lungs_rhinitis_common cold                                                  |
| Viral Infections and Immunity                | 60                  | 14_virus_influenza_epidemic_monkeypox                                                |
| Social Health Surveys                        | 60                  | 15_survey-type questions_education_life_female                                       |
| Medical Education and Common Knowledge       | 58                  | 16_medicine_hospital_medical care_doctor                                             |
| Vision Protection and Eye Health             | 58                  | 17_eyes_myopia_vision_ophthalmology                                                  |
| Intestinal Health                            | 53                  | 18_constipation_spleen and stomach deficiency_intestines and stomach_stomach disease |
| Kidney Health                                | 50                  | 19_uric acid_uremia_high uric acid_kidney failure                                    |
| Natural Sciences and Technological Progress  | 47                  | 20_natural science_life health_cutting-edge technology_pseudo-science popularization |
| Disease Symptoms and Diagnosis               | 46                  | 21_symptoms_manifestations_disease_accompanied                                       |
| Vitamins and Nutritional Supplements         | 41                  | 22_vitamins_deficiency_deficiency disease_multivitamins                              |
| Hearing and Ear Health                       | 32                  | 23_hearing aid_ears_tinnitus_hearing                                                 |
| Smoking Cessation and Healthy Living         | 32                  | 24_smoking cessation_smoking_cigarettes_secondhand smoke                             |
| Teeth and Oral Health                        | 32                  | 25_teeth_brushing teeth_oral cavity_electric toothbrush                              |

#### 4.3. Update Strategy and Evolution Analysis of Dynamic Seed Topic

After obtaining the initial seed topics, using them as a static set over time makes large language models prone to bias when processing newly arriving documents. As the corpus volume increases, large language models tend to handle new documents by generating excessively fragmented subtopics [32], resulting in rapid topic inflation and semantic redundancy within the topic system.

##### 4.3.1. RAG-Inspired Semantic Matching Mechanism

The proposed framework is inspired by the retrieval-augmented generation (RAG) paradigm, specifically its core idea of semantic similarity-based retrieval from an external knowledge base. To counteract topic redundancy, conventional practice introduces dynamic update mechanisms that eliminate low-frequency topics according to preset frequency thresholds [33]. Yet, frequency-driven update strategies harbor a critical defect: they readily delete core topics that are “semantically important yet low in frequency.” For example, should the “beverages” topic be eliminated for failing to meet the frequency threshold, the model, when subsequently handling similar questions, will diverge into generating more fragmented substitute labels such as “cold drinks” and “hot drinks,” thereby inducing disorder throughout the entire topic structure. To surmount the limitations of frequency thresholds, this paper introduces the similarity-matching concept from retrieval-augmented generation (RAG), upgrading the “replacement logic” of extant algorithms to “retrieval logic.” Concretely, the continuously evolving seed topic set is regarded as a dynamic external knowledge base; by comparing the semantic similarity between the current user question and each topic in the knowledge base, topics exhibiting high relevance to the question are preferentially retrieved for guidance, thereby effectively restraining the disorderly proliferation of redundant topics.

#### 4.3.2. Algorithm Implementation of Update Strategy

To jointly account for topic popularity and semantic relevance, we propose a TOP-K cross-filtering strategy based on topic frequency and semantic similarity. For each incoming question, the strategy retrieves two complementary subsets from the seed-topic repository: globally frequent topics and topics that are semantically relevant to the current question. Their union constitutes the dynamic seed-topic set used to guide the subsequent LLM-based multi-topic identification. The complete procedure is presented in Algorithm 1.

---

#### Algorithm 1 TOP-K cross-filtering for dynamic seed-topic retrieval

---

**Require:** A new question document  $w$ ; seed-topic set  $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ ; topic-frequency set  $\mathcal{F} = \{f_1, f_2, \dots, f_n\}$ ; seed-topic embedding set  $\mathcal{V}_{\mathcal{S}} = \{\mathbf{v}_{s_1}, \mathbf{v}_{s_2}, \dots, \mathbf{v}_{s_n}\}$ ; filtering parameter  $K$

**Ensure:** A document-specific dynamic seed-topic set  $\mathcal{S}_{\text{dynamic}}$

- 1: Encode the incoming question  $w$  into an embedding vector  $\mathbf{v}_w$
- 2: Rank all seed topics in  $\mathcal{S}$  in descending order according to their corresponding frequencies in  $\mathcal{F}$
- 3: Select the  $K$  seed topics with the highest frequencies:

$$\mathcal{S}_{\text{freq}} \leftarrow \text{TopK}_{\text{freq}}(\mathcal{S}, \mathcal{F}, K)$$

- 4: **for**  $i = 1$  to  $n$  **do**

- 5:     Compute the cosine similarity between the question vector  $\mathbf{v}_w$  and the seed-topic vector  $\mathbf{v}_{s_i}$ :

$$r_i \leftarrow \cos(\mathbf{v}_w, \mathbf{v}_{s_i}) = \frac{\mathbf{v}_w^\top \mathbf{v}_{s_i}}{\|\mathbf{v}_w\|_2 \|\mathbf{v}_{s_i}\|_2}$$

- 6: **end for**

- 7: Select the  $K$  seed topics with the highest semantic similarity scores:

$$\mathcal{S}_{\text{sim}} \leftarrow \text{TopK}_{\text{sim}}(\mathcal{S}, \{r_i\}_{i=1}^n, K)$$

- 8: Construct the dynamic seed-topic set by taking the union of the two candidate sets:

$$\mathcal{S}_{\text{dynamic}} \leftarrow \mathcal{S}_{\text{freq}} \cup \mathcal{S}_{\text{sim}}$$

- 9: Remove duplicate topics from  $\mathcal{S}_{\text{dynamic}}$

- 10: **return**  $\mathcal{S}_{\text{dynamic}}$
- 

The frequency-based subset preserves globally prevalent topics, whereas the similarity-based subset retains contextually relevant topics, including low-frequency topics with high semantic value. In the empirical implementation,  $K = 5$  was selected to balance topic coverage and retrieval precision. Larger values would introduce many loosely related topics and weaken the constraining effect of the dynamic seed-topic set, whereas smaller values might overlook semantically relevant but low-frequency topics. Therefore, it represents a practical balance between topic coverage and retrieval precision. Moreover, the proposed framework is not designed to depend on a single retrieval criterion. By jointly considering topic frequency and semantic similarity, the dynamic retrieval strategy reduces reliance on a specific  $K$  value and improves the stability of topic retrieval.

#### 4.3.3. Empirical Performance and Summary of the Update Strategy

Upon application of the aforementioned TOP-K cross-filtering update strategy to the Zhihu health science popularization data sequence, the model exhibited outstanding stability and dynamic adaptability. On the one hand, the strategy enabled global hotspot demands such as “physical health” and “medical services” to evolve steadily across the entire temporal cycle; on the other hand, the semantic recall mechanism effectively pre-

served highly specific demands of comparatively low frequency, such as “hearing and ear health” and “teeth and oral health,” preventing them from being forcibly truncated by the algorithm’s frequency threshold.

In summary, the Dynamic Seed Topic construction method effectively resolves the problems of deficient prior knowledge and uncontrolled granularity encountered by large models in complex socialized text mining. This update mechanism, which concurrently accommodates frequency evolution and semantic matching, not only restores the authentic dynamic trajectories of platform information but also furnishes precise benchmark references and boundary constraints for the subsequent chapter’s in-depth examination of individual question documents and the identification and mining of fine-grained multiple information demand topics. This update mechanism simultaneously considers frequency evolution and semantic matching. It effectively captures the dynamic evolution of platform information and provides clear benchmark references and boundary constraints for the subsequent in-depth analysis of individual questions and the identification of fine-grained multi-demand topics.

## 5. Develop Multi-Topic Model

### 5.1. Multi-Topic Based on Prompt

The primary goal of this stage is to develop a multi-topic model. The identification of multiple topics is a relatively complex natural language processing task, involving a compound process of semantic understanding, topic generation, topic calibration, and result validation. This complexity is particularly evident in social Q&A platforms, where user questions are typically expressed in short texts that combine colloquial language with non-standard wording, tag interference, and multiple coexisting intentions [34,35]. Against this background, this study constructs a multi-topic model based on large language models (LLMs) and prompt engineering, organized around three interrelated stages: topic identification, topic optimization, and topic assignment. The aim is to reveal, in a more comprehensive manner, one or more latent information-need topics embedded in user questions.

When topic identification relies only on a static topic set or one-round generation, it is often difficult to balance accuracy, completeness, and explainability. On the one hand, LLMs provide strong semantic understanding and contextual reasoning capabilities, enabling the recognition of core demand-related information in short texts [36]. On the other hand, without explicit constraints, such capabilities may also result in imbalanced topic granularity, blurred semantic boundaries, or unstable assignment criteria [33]. Accordingly, the central objective at this stage is to improve the stability and credibility of multi-topic identification. This framework offers valuable insights into user demand in social Q&A and provides a scalable solution that can be adapted to analyze other similar platforms or context.

#### 5.1.1. Topic Identification

First, we identify the topic corresponding to the question through prompts. To improve the alignment between topic identification and users’ actual intentions, a prompt template composed of role, objective, constraints, instructions, and examples is designed. Within the prompt, the boundaries of topics in a specific domain are predefined in order to reduce ambiguity in semantic judgment. The core purpose of this design is not to increase prompt complexity, but to guide the model toward stable topic identification under a shared semantic standard through a clearly specified task framework. Accordingly, the topic identification stage is divided into three sequential tasks: text analysis, topic generation, and topic assignment with keyword identification.

Among them, text analysis constitutes the prerequisite for topic identification. User texts on social Q&A platforms are generally short, and a considerable proportion of content contains surface-level health keywords while being unrelated to actual health information demands [34,35]. Compared with traditional methods that rely on stop-word lists and domain dictionaries, LLMs can identify and filter noisy text at the semantic level, thereby improving the corpus quality for subsequent modeling. By combining contextual semantics with the preliminary definition of health information demands, 170 invalid statements were removed, accounting for 4.8% of the original data. On the basis of the valid text, the model further performs the topic generation task. Because users' information demands are inherently open-ended, a fixed set of seed topics is insufficient to fully cover real demands [37]. Therefore, at this stage, the model is allowed to identify potential new topics beyond the existing topic set, while unconstrained expansion is suppressed through mechanisms such as granularity checking and cross-validation. The value of this design lies not in increasing the number of topics, but in strengthening the capacity of the topic system to accommodate complex needs [37,38].

Ultimately, the model identified 51 topic instances from 3359 valid questions, representing 38 distinct topics, as shown in Table 4. Topics marked with an asterisk (\*) denote unique topics. All initial seed topics fall within this set of unique topics. This finding indicates both the necessity of the initial seed topics and the effectiveness of the dynamic expansion mechanism in complementing demand content that cannot be adequately covered by the original topic set.

**Table 4.** Initial topic identification.

| Original Seed Topics                        | Counts | New Topics                          | Counts |
|---------------------------------------------|--------|-------------------------------------|--------|
| *Disease Symptoms and Diagnosis             | 1029   | *Physical Health                    | 409    |
| *Lifestyle Habits and Physical Health       | 836    | Disease Prevention and Treatment    | 310    |
| *Medical Education and General Knowledge    | 572    | *Dietary Habits and Physical Health | 104    |
| *Diet and Health Foods                      | 427    | Adolescent Health                   | 91     |
| *Mental Health                              | 256    | *Workplace Health                   | 61     |
| *Fitness and Exercise                       | 247    | *Elderly Health                     | 57     |
| *Sleep and Health                           | 167    | *Personal Growth                    | 43     |
| *Social Health Surveys                      | 155    | Male Health                         | 38     |
| *Viral Infections and Immunity              | 147    | *Beauty and Skincare                | 36     |
| *Sub-health Management                      | 141    | Female Health                       | 36     |
| *Cancer Prevention and Treatment            | 131    | *Public Health                      | 34     |
| *Traditional Chinese Medicine               | 124    | *First Aid and Emergency Response   | 33     |
| *Beverages and Health                       | 117    | Pharmaceutical Manufacturing        | 30     |
| *Medical Visits and Healthcare Services     | 111    | Liver Health                        | 29     |
| *Reproductive Health for Women              | 104    | *Scientific Weight Loss             | 28     |
| *Bone Health                                | 103    | Nutritional Science                 | 22     |
| *Vitamins and Nutritional Supplements       | 94     | *Daily Life and Health              | 19     |
| *Diabetes and Blood Sugar Management        | 87     | *Child Health                       | 18     |
| *Vision Protection and Eye Health           | 68     | Cardiovascular Health               | 12     |
| *Kidney Health                              | 68     | *Law and Responsibility             | 7      |
| *Dental and Oral Health                     | 68     | Psychological Science               | 7      |
| *Gut Health                                 | 59     | Gender Relationships and Emotions   | 4      |
| *Respiratory Diseases                       | 55     | Medical Devices                     | 3      |
| *Hearing and Ear Health                     | 43     | Sports and Health                   | 3      |
| *Natural Science and Technological Progress | 39     | Gastroenterology                    | 1      |
| *Smoking Cessation and Healthy Living       | 38     |                                     |        |

### 5.1.2. Topic Optimization

It is still necessary to optimize the topic set. An increase in the number of topics does not imply an improvement in topic quality. When different topics overlap substantially at the semantic level, a more fine-grained topic set may instead reduce the overall explainability of the analysis. Without predefining the number of topics, this study introduces an optimization strategy based on semantic similarity and prompt constraints to address topic duplication, ambiguous naming, and inconsistent granularity [27,39]. Low-frequency

topics are more likely to lie at the semantic periphery or to exhibit blurred boundaries [40]. Prioritizing such topics is therefore conducive to improving the overall structure without disrupting high-frequency core topics. Specifically, topic frequencies are first tallied across all topics from the initial identification round, with the lowest-frequency topic designated as the starting point. Candidate topics are identified by computing the average semantic similarity between feature words, retaining up to three topics that exceed a similarity threshold of 0.5. If no topic meets this criterion, the procedure terminates. The initial topic, candidate topics, and their feature words are then submitted alongside the prompt template (in Appendix A) to the large language model. If a merge occurs, the feature word set is updated and c-TF-IDF scores are recalculated. This process iterates until all topics have been traversed.

Through this process, the 51 identified topics are ultimately optimized into 36 topics. Compared with the initial seed topics, the optimized topic set retains distinctions in user needs. The results of each stage are presented in Table 5. Compared with the preliminary identification results, the optimized topic set exhibits clearer topic boundaries. In other words, the evolution from 26 topics to 51 and then the convergence to 36 reflects not a simple numerical change but a process in which the topic system moves from coarse-grained identification to fine-grained expansion and then returns to a structure with higher discriminative power.

**Table 5.** Evolution of topic sets across identification and optimization stages.

| Stage                                                     | Topic Num. | Topic Set                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------------------------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Initial seed topics based on BERTopic and prompts         | 26         | Disease Symptoms and Diagnosis; Lifestyle Habits and Physical Health; Medical Education and General Knowledge; Diet and Health Foods; Mental Health; Fitness and Exercise; Sleep and Health; Social Health Surveys; Viral Infections and Immunity; Sub-health Management; Cancer Prevention and Treatment; Traditional Chinese Medicine; Beverages and Health; Medical Visits and Healthcare Services; Reproductive Health for Women; Bone Health; Vitamins and Nutritional Supplements; Diabetes and Blood Sugar Management; Vision Protection and Eye Health; Kidney Health; Dental and Oral Health; Gut Health; Respiratory Diseases; Hearing and Ear Health; Natural Science and Technological Progress; Smoking Cessation and Healthy Living                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Topics based on dynamic seed topics and RAG-inspired      | 51         | Physical Health; Disease Prevention and Treatment; Dietary Habits and Physical Health; Adolescent Health; Workplace Health; Elderly Health; Personal Growth; Male Health; Beauty and Skincare; Female Health; Public Health; First Aid and Emergency Response; Pharmaceutical Manufacturing; Liver Health; Scientific Weight Loss; Nutritional Science; Daily Life and Health; Child Health; Cardiovascular Health; Law and Responsibility; Psychological Science; Gender Relationships and Emotions; Medical Devices; Sports and Health; Gastroenterology; Disease Symptoms and Diagnosis; Lifestyle Habits and Physical Health; Medical Education and General Knowledge; Diet and Health Foods; Mental Health; Fitness and Exercise; Sleep and Health; Social Health Surveys; Viral Infections and Immunity; Sub-health Management; Cancer Prevention and Treatment; Traditional Chinese Medicine; Beverages and Health; Medical Visits and Healthcare Services; Reproductive Health for Women; Bone Health; Vitamins and Nutritional Supplements; Diabetes and Blood Sugar Management; Vision Protection and Eye Health; Kidney Health; Dental and Oral Health; Gut Health; Respiratory Diseases; Hearing and Ear Health; Natural Science and Technological Progress; Smoking Cessation and Healthy Living |
| Optimized topics based on semantic similarity and prompts | 36         | Sports and Health; Diet and Health; Lifestyle Habits and Health; Mental Health; Sub-health Management; Disease Management and Health; Sleep and Health; Smoking Cessation and Healthy Living; Cancer Prevention and Treatment; Social Health Surveys; Medical Education and General Knowledge; Respiratory Diseases; Viral Infections and Immunity; Bone Health; Vision Protection and Eye Health; Medical Visits and Healthcare Services; Natural Science and Technological Progress; Diabetes and Blood Sugar Management; Hearing and Ear Health; Female Health; Traditional Chinese Medicine; Dental and Oral Health; Male Health; Pharmaceutical Management; Workplace Health; Beauty and Skincare; Adolescent Health; First Aid and Emergency Response; Elderly Health; Liver Health; Law and Responsibility; Child Health; Cardiovascular Health; Scientific Weight Loss; Digestive System Diseases; Kidney Health                                                                                                                                                                                                                                                                                                                                                                                      |

### 5.1.3. Topic Assignment

Following topic identification and topic optimization, the focus shifts to the correspondence between topics and individual documents. At this point, a set of interpretable topics with clear semantic distinctions is retained. To ensure semantic completeness and avoid

confusion, the goal of the topic assignment stage is to establish and refine a more effective and interpretable linkage between the topic list and the documents in the dataset [41]. At this stage, the prompted LLM assigns one or more topics to each document. The final output includes not only the assigned topic labels and topic keywords but also detailed rationales for topic assignment, which substantially enhances the explainability and transparency of the results. Building on the prompt used for topic identification, the input prompt is further optimized by removing the judgment and validation instructions related to new-topic generation, thereby enabling the LLM to focus more directly and accurately on the current task of topic assignment. We report the prompt template used in Appendix A.

Among the 36 topics, the top three are “Disease Management and Health,” “Lifestyle Habits and Health,” and “Diet and Health,” with frequencies of 777, 640, and 637, accounting for 12%, 10%, and 10% of the total, respectively. The results show that disease management and health ranks first, reflecting users’ concern with disease prevention, treatment, and health management. Diet and health and lifestyle habits and health are each mentioned over 10%, highlighting users’ information demands regarding the relationship of diet and lifestyle habits to health. In addition, topics such as exercise and health and mental health also receive substantial attention. Overall, the data exhibit diverse and comprehensive characteristics. On the basis of topic keywords, c-TF-IDF scores are calculated to present each topic together with its top five representative feature words, as reported in Table 6. According to the commonalities across topics, we find that users’ health science popularization demands on social Q&A platforms can be grouped into four broad categories: lifestyle-oriented demands, disease prevention and management demands, demands related to specific populations, and demands related to medical and health education.

**Table 6.** Optimized topics and representative keywords.

| No. | Topic                                      | Top 5 Topic Keywords                                                                          |
|-----|--------------------------------------------|-----------------------------------------------------------------------------------------------|
| 1   | Disease Management and Health              | disease, prevention, symptom, treatment, ataxia                                               |
| 2   | Lifestyle Habits and Health                | life, health, wellness, lifestyle, staying up late                                            |
| 3   | Diet and Health                            | diet, nutrition, food product, food, vitamin                                                  |
| 4   | Medical Education and General Knowledge    | medicine, common knowledge, popular science, doctor, clinical medicine                        |
| 5   | Mental Health                              | mental health, psychology, psychology, emotion, anxiety                                       |
| 6   | Sports and Health                          | exercise, fitness, workout, running, weight loss                                              |
| 7   | Sub-health Management                      | management, sub-health, physical health, body, medical examination                            |
| 8   | Social Health Surveys                      | society, survey, issue, China, population                                                     |
| 9   | Medical Visits and Healthcare Services     | hospital, healthcare, doctor, surgery, medical examination                                    |
| 10  | Viral Infections and Immunity              | COVID-19, virus, epidemic, pneumonia, AIDS                                                    |
| 11  | Scientific Weight Loss                     | weight loss, method, obesity, fat loss, weight                                                |
| 12  | Female Health                              | woman, pregnancy, menstruation, girl, infertility                                             |
| 13  | Sleep and Health                           | sleep, insomnia, staying up late, quality, bed                                                |
| 14  | Traditional Chinese Medicine               | health preservation, traditional Chinese medicine, Chinese herbal medicine, blood, deficiency |
| 15  | Bone Health                                | protrusion, cervical spine, lumbar disc, lumbar spine, cervical spondylosis                   |
| 16  | Digestive System Diseases                  | constipation, gastrointestinal tract, digestion, stomach, diarrhea                            |
| 17  | Cancer Prevention and Treatment            | cancer, tumor, lung cancer, treatment, liver cancer                                           |
| 18  | Natural Science and Technological Progress | technology, natural science, frontier, life, medicine                                         |
| 19  | Respiratory Diseases                       | cough, rhinitis, pneumonia, chronic, nodule                                                   |
| 20  | Diabetes and Blood Sugar Management        | diabetes, blood sugar, diet, hyperglycemia, complication                                      |
| 21  | Beauty and Skincare                        | skincare, skin, beauty, care, basic                                                           |
| 22  | Child Health                               | child, baby, kid, infant, parenting                                                           |
| 23  | Pharmaceutical Management                  | drug, medicine, pharmaceutical, side effect, taking medicine                                  |
| 24  | Vision Protection and Eye Health           | eye, vision, myopia, health care, nearsightedness                                             |
| 25  | Cardiovascular Health                      | heart, heart disease, cardiovascular system, blood pressure, blood vessel                     |
| 26  | Elderly Health                             | older adult, elderly person, old age, aging, elderly care                                     |
| 27  | Dental and Oral Health                     | tooth, oral cavity, toothbrushing, oral medicine, oral ulcer                                  |
| 28  | First Aid and Emergency Response           | first aid, bite, doctor, treatment, resuscitation                                             |
| 29  | Liver Health                               | liver, hepatitis, liver disease, liver cancer, fatty liver                                    |
| 30  | Adolescent Health                          | young person, adolescent, puberty, development, college student                               |
| 31  | Smoking Cessation and Healthy Living       | smoking cessation, smoking, cigarette, smoker, electronic cigarette                           |

Table 6. Cont.

| No. | Topic                  | Top 5 Topic Keywords                                                 |
|-----|------------------------|----------------------------------------------------------------------|
| 32  | Hearing and Ear Health | ear, hearing aid, hearing, tinnitus, deafness                        |
| 33  | Workplace Health       | work, workplace, overtime, going to work, life                       |
| 34  | Kidney Health          | kidney disease, kidney, nephritis, kidney failure, kidney deficiency |
| 35  | Law and Responsibility | law, doctor, responsibility, family member, staff member             |
| 36  | Male Health            | man, boy, care, male, sperm                                          |

### 5.2. Performance of the Multi-Topic Model

Given that user demands on social Q&A platforms are complex, open-ended, and lack standardized labels, it is difficult to fully demonstrate the effectiveness of multi-topic identification through a single evaluation approach. Consequently, this study conducts a comprehensive assessment of the identification results of both quantitative and human evaluation [42]. The quantitative evaluation is used to assess the structural quality of the topic system, whereas the human evaluation further examines whether the assignment results can authentically and comprehensively reflect user demands.

#### 5.2.1. Comparison with Other Topic Model

For objective quantitative evaluation, topic coherence (TC), topic diversity (TD), and topic similarity are adopted as the evaluation metrics [43]. We select the top 20 words with the highest c-TF-IDF scores under each topic as topic words, and calculate N-topic similarity from the top 20 high-frequency topics. Since the method in this paper is unsupervised, the number of topics is specified as  $K = 36$  as the final output of the large language model; to ensure comparability, LDA is fixed at  $K = 36$ ; one of the advantages of the BERTopic topic clustering method is that it can autonomously find the optimal number of topics, and after parameter tuning, it can generate a maximum of 34 topics, so  $K = 34$  is used for comparison. Specifically, the calculation of topic similarity is shown in Equation (3), where  $Topic_i$  denotes the sum of the word embeddings in topic  $i$ . These three types of metrics are selected because they correspond respectively to semantic consistency within topics, distinctiveness across topics, and the degree of redundancy among high-frequency topics.

$$Similar_N = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N \cos(\text{Topic}_i, \text{Topic}_j)}{N(N-1)/2} \quad (3)$$

To verify the effectiveness of the proposed method combining BERTopic and Dynamic Seed Topics (DSTs) in topic recognition, this paper selects LDA [8], BERTopic, LLM-DST, LLM-DST-refine, LLM-BERTopic, BERTopic-LLM-DST, and BERTopic-LLM-DST-refine as baseline models for a comprehensive comparative analysis of topic recognition performance. Among these, Latent Dirichlet Allocation (LDA) is a traditional topic model where each topic represents a probability distribution over words; LDA plays a crucial role in demand-based topic recognition. DST-refine refers to the dynamic seed topic strategy proposed in this paper, while DST refers to the frequency-based seed replacement strategy proposed in previous research. As shown in Table 7, the results indicate that BERTopic-LLM-DST achieves the best performance in topic coherence, indicating that its internal topic semantics are more concentrated. BERTopic obtains the highest score in topic diversity, reaching 0.826, which suggests a clear advantage in distinguishing different topics. The proposed method ranks second on this metric with a score of 0.792, indicating that the model still maintains good topic coverage and discriminability after the introduction of dynamic seed topics. At the same time, BERTopic-LLM-DST obtains the lowest score in 20-topic similarity, suggesting that it is more effective in avoiding topic repetition. Overall, the proposed method achieves a desirable balance among coherence, diversity, and non-redundancy.

**Table 7.** Performance comparison of topic models.

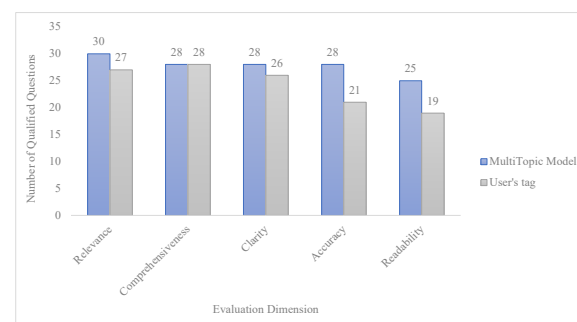
| Model                   | K (Topic Num.) | TC    | TD    | Topic Similarity |
|-------------------------|----------------|-------|-------|------------------|
| BERTopic                | 34             | 0.703 | 0.826 | 0.743            |
| LDA                     | 36             | 0.546 | 0.576 | 0.923            |
| LLM-DST                 | 112            | 0.742 | 0.492 | 0.951            |
| LLM-DST-refine          | 96             | 0.714 | 0.534 | 0.892            |
| BERTopic-LLM-DST        | 46             | 0.903 | 0.674 | 0.759            |
| BERTopic-LLM-DST-refine | 36             | 0.987 | 0.792 | 0.698            |

### 5.2.2. Human Evaluation

Although quantitative metrics can reveal the statistical characteristics of topic structure, information demands topics themselves be highly interpretive and subjective. It is therefore still necessary to introduce human evaluation to examine whether the model output is genuinely close to users' actual expressions. This is particularly important on social Q&A platforms, where there is an absence of standardized ground-truth labels. It means that human evaluation serves not only as a complementary approach but also as a necessary basis for judging whether topic assignment is acceptable in practice. Thus, human evaluation serves as both a complementary approach and a necessary basis for assessing whether the topic assignment is acceptable in practice. Original tags are treated as preliminary descriptions of users' own demands and are compared with the information demands topics generated by the model in order to assess whether the latter can more fully capture the intention underlying users questions.

In the evaluation process, 30 user questions were provided to each volunteer, together with the original topic tags and the information demand topics generated in this study. The evaluation criteria included five dimensions: relevance, comprehensiveness, clarity, accuracy, and readability. One point was assigned for each satisfied criterion, so the score for each topic label ranged from 0 to 5. To ensure sufficient platform experience, 17 postgraduate volunteers majoring in information management or information systems were recruited. All evaluators had at least three years of active experience using social Q&A platforms such as Zhihu, as well as prior experience with text analysis and topic interpretation. Before the evaluation, all participants independently completed the assessment following a unified evaluation protocol and identical scoring criteria, thereby reducing subjective variation.

The evaluation results are shown in Figure 2. Across the five specific dimensions, the multi-topic model performs at least as well as or better than user-defined tags in relevance, comprehensiveness, clarity, accuracy, and readability, with more evident advantages in relevance and accuracy. These findings suggest that the multi-topic demand identified in this study not only exhibits good semantic quality at the structural level, but also more closely approximates the genuine expression of users' latent demands. The human evaluation serves as complementary evidence to verify the practical usability of the identified topics and is intended to supplement, rather than replace, the quantitative experimental evaluation.

**Figure 2.** Results of human evaluation.

## 6. Additional Analysis

### 6.1. Methodology

#### 6.1.1. Network Construction

Demand topics are generally regarded as basic knowledge units within online knowledge environments. An individual topic can only reflect a relatively independent category of demand content, whereas the associations among topics further reveal how the structure of user demands is organized. In other words, there are multiple topics involved in a user's question. It does not represent a simple combination of parallel elements but rather tends to indicate latent connections among these topics. Building upon the topic modeling framework, we further examine the logical structure among topics from the perspective of a topic association network. The analysis aims to identify topics that serve as hubs within the user demand system and to characterize the connectivity patterns that exist between them.

This study constructs a topic co-occurrence network based on the co-occurrence relationships. On social Q&A platforms, when multiple demand topics are simultaneously identified within a question, these topics can form an associative edge between them. As a large number of questions are incorporated into the analysis, repeated co-occurrence gradually gives rise to a topic network with an overall structure. In this network, nodes represent demand topics, edges represent co-occurrence relationships, and edge weights encode the frequency of co-occurrence. Through this construction, the otherwise dispersed information needs of individual users are transformed into an interconnected system with identifiable structural characteristics. It provides a methodological foundation for identifying key topics and understanding their functional roles.

#### 6.1.2. Network Metrics and Analysis

To comprehensively characterize the structural properties of this network, the analysis is carried out through both network structure analysis and key node analysis. The former focuses on whether the overall network is tightly connected and whether linkages among topics are broadly present, while the latter focuses on identifying topics that occupy structurally prominent positions and thus exert greater influence over demand organization and information flow. The former examines the overall connectivity of the network and the breadth of linkages among topics. The latter identifies topics that occupy central positions and exert greater influence over demand organization and information flow.

##### (1) Network Structural Analysis

Network structural analysis employs the metric of network density to quantify the overall degree of connectivity in the topic co-occurrence network. This metric indicates whether the demand form a highly interwoven global structure. Network density is defined as the ratio of the actual number of edges present in the network to the maximum number of edges, as formulated in Equation (4), where  $E$  denotes the number of edges actually present and  $N$  denotes the number of nodes. Higher network density indicates that different topics tend to co-occur more readily within user queries and that the associations among topics are correspondingly closer. By contrast, a lower network density suggests that relationships among topics are relatively limited, and that the network may contain more weak ties or even mutually disconnected topics.

$$Density = \frac{E}{N(N-1)/2} \quad (4)$$

##### (2) Key Node Analysis

Beyond the overall structure, it is necessary to identify those topic nodes that play pivotal roles within the network. Key node analysis reveals which topics are of greater

importance in connecting other topics, sustaining the network's structural integrity, or representing central demand categories. Relying solely on topic frequency is insufficient to determine whether a given topic possesses the capacity to bridge different demand categories within the network. Accordingly, three metrics are employed to characterize the importance of topic nodes: degree centrality, weighted degree, and average weighted degree.

First, degree centrality measures the breadth of direct connections that a topic establishes with other topics, defined as the number of nodes to which it is directly linked. The corresponding calculation is given in Equation (5). In the demand topic co-occurrence network, higher degree centrality indicates that a topic co-occurs more frequently with a greater variety of other topics. It suggests that it is more likely to occupy a central position within the user demand structure. Such topics usually do not merely represent a single independent category of demand but may instead serve an important role in linking multiple categories of content.

$$C_D(v) = \deg(v) \quad (5)$$

Second, weighted degree further takes connection strength into account rather than considering only the number of connections. In the present study, edge weights are defined by the co-occurrence frequency between topic pairs. Accordingly, weighted degree reflects the cumulative relational intensity between a given topic and all its connected counterparts, as formulated in Equation (6). Compared with degree centrality, weighted degree places greater emphasis on how strong these connections are. It can identify nodes that maintain frequent and stable co-occurrence relationships with several important topics. Such topics often represent content with stronger stickiness and aggregation effects within user demands.

$$C_w(v) = \sum_{u \in N(v)} w(v, u) \quad (6)$$

Finally, the average weighted degree of neighboring nodes quantifies the overall connectivity strength of a node's immediate neighborhood. It is calculated by the mean weighted degree of its adjacent nodes, as formulated in Equation (7). This metric is motivated by the recognition that a node's structural importance is determined not only by its own connectivity profile but also by the local structural context in which it is embedded. When a topic's neighbors themselves exhibit high weighted degrees, the topic is more likely situated within a high-intensity associative region of the network.

$$C_{Aw}(v) = \frac{1}{\deg(v)} \sum_{u \in N(v)} w(v, u) \quad (7)$$

In summary, the construction of a demand topic co-occurrence network and integrating indicators enable a comprehensive analysis of both the global connectivity and key node characteristics. From a network perspective, this approach can reveal the core topics, strongly associated topics, and latent linkage patterns within the structure of user questions. It also provides support for subsequent efforts to understand the organizational logic of user demands and the knowledge supply of the platform.

## 6.2. Result

The health science popularization demand topic co-occurrence network illustrates the co-occurrence relationships among different health science popularization demand topics. The network density reaches 0.69, a value that clearly indicates relatively tight connections among the various nodes within the network and demonstrates strong associations among the health science popularization demand topics embedded in user questions.

This study measures the role of topics in the topic co-occurrence network through three topological indicators: degree centrality, weighted degree, and average weighted degree. The top-10 topics under each indicator along with their respective scores are presented in Table 8. Degree centrality measures the number of connections a node maintains with other nodes in the network, thereby reflecting the breadth of thematic associations. The topic “Disease Management and Health” ranks first in degree centrality, indicating that it has the closest associations with other topics in the topic co-occurrence network. In addition, the topics “Diet and Health” and “Medical Education and Common Knowledge” both achieve a degree centrality of 1.00, likewise demonstrating high levels of association. The weighted degree indicator incorporates connection strength or weight on the basis of connection quantity, thereby providing a more comprehensive reflection of the degree of association among topics. The topic “Lifestyle Habits and Health” performs most prominently under the weighted degree indicator, indicating the existence of high-frequency and high-strength connections between it and other nodes. At the same time, the topics “Disease Management and Health” and “Diet and Health” also exhibit relatively high weighted degrees. Average weighted degree is obtained by dividing the weighted degree by the degree, thereby supplying the average strength of each connection. The topic “Lifestyle Habits and Health” again stands out most prominently in average weighted degree, reaching 26.82 and markedly exceeding other topics. The average weighted degrees of the topics “Disease Management and Health” and “Diet and Health” are 21.51 and 17.89, respectively, ranking second and third.

**Table 8.** Topic node centrality (top 10).

| Topic                                        | Degree Centrality | Topic                                        | Weighted Degree | Topic                                  | Average Weighted Degree |
|----------------------------------------------|-------------------|----------------------------------------------|-----------------|----------------------------------------|-------------------------|
| Disease Management and Health                | 1.06              | Lifestyle Habits and Health                  | 885             | Lifestyle Habits and Health            | 26.82                   |
| Diet and Health                              | 1.00              | Disease Management and Health                | 796             | Disease Management and Health          | 21.51                   |
| Medical Education and Common Knowledge       | 1.00              | Diet and Health                              | 626             | Diet and Health                        | 17.89                   |
| Medical Consultation and Healthcare Services | 0.97              | Medical Education and Common Knowledge       | 565             | Medical Education and Common Knowledge | 16.14                   |
| Lifestyle Habits and Health                  | 0.94              | Mental Health                                | 407             | Mental Health                          | 12.72                   |
| Mental Health                                | 0.91              | Exercise and Health                          | 360             | Exercise and Health                    | 12.00                   |
| Exercise and Health                          | 0.86              | Subhealth Management                         | 343             | Subhealth Management                   | 11.83                   |
| Subhealth Management                         | 0.83              | Social Health Surveys                        | 324             | Social Health Surveys                  | 11.17                   |
| Social Health Surveys                        | 0.83              | Scientific Weight Loss                       | 242             | Scientific Weight Loss                 | 9.31                    |
| Cancer Prevention and Treatment              | 0.80              | Medical Consultation and Healthcare Services | 240             | Viral Infections and Immunity          | 8.74                    |

This study further examined the strength of topic co-occurrence pairs, defined as the frequency with which different topics appear simultaneously in the same context, in order to quantitatively assess the closeness of the association between topics. The top-10 topic co-occurrence pairs are shown in Table 9. As can be seen from the table, the co-occurrence pair “Diet and Health—Lifestyle Habits and Health” registers the highest co-occurrence frequency, indicating that users typically mention “Diet and Health” and “Lifestyle Habits and Health” simultaneously in their questions; the two exhibit extremely high association and interdependence. The next highest is “Diet and Health—Disease Management and Health”, revealing that users’ “Diet and Health” demands are also closely linked with “Disease Management and Health” demands within the health science popularization

domain on social question-and-answer platforms. The strength of topic co-occurrence pairs reflects the degree of association among different topics and further demonstrates the correlation and semantic relatedness among user demand topics, thereby furnishing certain reference value for information resource aggregation strategies such as topic-label-based recommendations.

**Table 9.** Topic co-occurrence pair strength (Top 10).

| Co-Occurrence Topic Pair                                             | Co-Occurrence Frequency |
|----------------------------------------------------------------------|-------------------------|
| Diet and Health—Lifestyle Habits and Health                          | 109                     |
| Diet and Health—Disease Management and Health                        | 106                     |
| Lifestyle Habits and Health—Subhealth Management                     | 100                     |
| Lifestyle Habits and Health—Mental Health                            | 97                      |
| Disease Management and Health—Medical Education and Common Knowledge | 79                      |
| Exercise and Health—Scientific Weight Loss                           | 72                      |
| Lifestyle Habits and Health—Medical Education and Common Knowledge   | 72                      |
| Disease Management and Health—Viral Infections and Immunity          | 64                      |
| Disease Management and Health—Lifestyle Habits and Health            | 61                      |

## 7. Conclusions and Discussion

### 7.1. Summary

In contemporary society, social question-and-answer platforms have become important platforms for satisfying users' information needs. The increasing diversification of users' information needs has led to information accumulation within the platforms; therefore, there is an urgent need to construct targeted information resource aggregation strategies. However, existing aggregation strategies based on topic classification and popular topic labels suffer from problems such as the semantic ambiguity of topics, disconnection from actual demands, and logical disorder, making it difficult to meet users' complex information needs and retrieval requirements. To address this, this paper proposes a dynamic seed topic construction and multi-demand topic identification method, which effectively achieves fine-grained, interpretable, and semantically consistent identification of multiple information-demand topics. It also reveals the associative features among topics and thus provides strong support for optimizing topic-based information aggregation strategies on social Q&A platforms.

### 7.2. Discussion

The results indicate that the proposed framework improves the organization of short and multi-demand questions by combining global topic discovery with document-level semantic guidance. The evolution from 26 initial seed topics to 51 identified topic instances and subsequently to 36 optimized topics should not be interpreted as simple topic inflation and reduction. Rather, it reflects a two-stage process in which the model first expands the semantic coverage of the initial topic repository and then consolidates overlapping or inconsistently granular topics.

The quantitative results provide complementary evidence regarding the quality of the resulting topic system. Compared with the baseline models, the proposed method achieved higher topic coherence and lower inter-topic similarity while retaining relatively high topic diversity. This pattern suggests that the dynamic seed mechanism and subsequent optimization do not merely increase the number of labels but help generate topics with more internally consistent and distinguishable semantic boundaries. The human evaluation further shows that the identified topics can capture information demands that are not fully represented by users' original tags, particularly in terms of relevance and accuracy.

The co-occurrence analysis extends these findings from individual topic identification to the relational structure of user demands. The prominence of disease management,

lifestyle, diet, and medical education topics indicates that health-related questions often combine disease-specific concerns with behavioral and informational needs. This implies that topic organization on social Q&A platforms should not treat user demands as isolated categories but should support multi-label and association-aware information aggregation.

### 7.3. Contributions

Compared with existing studies, the contributions of this paper are manifested in two aspects:

First, an innovative dynamic seed topic construction strategy is proposed. To guide the large language model in generating demand topics consistent with research expectations, this paper integrates the topic clustering advantages of the BERTopic model on short-text datasets with the topic naming capability of the large language model, thereby providing a readable initial seed topic set for social platforms that lack standard datasets and prior knowledge. Second, in terms of seed topic updating, this paper designs a TOP-K dynamic seed topic update strategy that comprehensively considers semantic similarity and frequency. This strategy significantly reduces redundancy and repetition in topic identification and guides the model toward generating important topics with greater semantic differentiation. The dynamic seed topic construction method proposed in this paper furnishes an effective solution for topic identification tasks and offers useful references for topic identification research in other fields.

Second, based on dynamic seed topic construction, this paper proposes an automated and integrated multiple demand topic identification method for social question-and-answer platforms driven by large language models. This method effectively addresses the characteristics of short texts and diverse user demands on social question-and-answer platforms, enhances the semantic consistency of topics across different topic identification and classification tasks, and significantly reduces the resource consumption and subjectivity issues associated with manual annotation. The fine-grained and interpretable multiple information demand topics identified by this method provide strong guidance for optimizing information resource aggregation strategies based on topic classification on social question-and-answer platforms.

### 7.4. Limitations and Future Prospects

The dynamic seed topic construction and multiple demand topic identification method proposed in this paper accurately and comprehensively identifies and assigns user demand topics on social question-and-answer platforms and further reveals the association features among demand topics. Nevertheless, this paper still has certain limitations: First, this study validated the proposed framework using questions from the Health Science Popularization topic on the Zhihu platform, which provides diverse user-generated content suitable for evaluating topic identification methods. However, the empirical validation was limited to a single platform and domain. Although the framework is designed to be domain-independent, its applicability to other platforms and domains has not yet been empirically verified. Future research should therefore conduct cross-platform and cross-domain validation to further examine its generalizability.

Second, the human evaluation in this study adopts a comparative scoring design rather than an inter-annotator agreement scheme. While this approach effectively demonstrates the relative advantage of the proposed method over user-assigned tags, the sample size and the absence of cross-validated inter-annotator agreement measures constitute limitations. Future work will expand the evaluation scale and incorporate standardized inter-annotator reliability metrics to further validate the method.

In addition, although dynamic seed updating may provide a technical basis for responding to evolving information needs, the present study does not directly model or evaluate temporal semantic drift. Drawing on the Sense-Making perspective, future research could combine time-aware embeddings with longitudinal validation to examine whether, and under what conditions, dynamic seeds can track changes in the meanings and boundaries of information-demand topics.

**Author Contributions:** Conceptualization: Y.Z. and X.Y.; Data curation: X.Y., T.Q. and L.L.; Investigation: T.Q. and L.L.; Methodology: Y.Z., X.Y. and T.Q.; Writing—original draft: Y.Z., X.Y. and T.Q.; Writing—review & editing: Y.Z., X.Y. and L.L.; Supervision: Y.Z.; Funding acquisition: Y.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Entrusted Project of Sichuan Provincial Social Science Fund (Grant No. SCJJ25RKX131).

**Data Availability Statement:** The data generated and analyzed during the current study are subject to privacy and ethical restrictions and therefore cannot be shared.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A

**Table A1.** Topic identification prompt template.

| Prompt Engineering Template (for Topic Identification)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>[Role]</b><br/>You are a topic identification expert in the field of {domain}</p> <p><b>[Goal]</b><br/>Based on user comments, comprehensively identify user demand topics within {topic/domain}</p> <p><b>[Constraints]</b><br/>           1. A demand topic refers to the information or services that users are concerned with and need under {topic/domain}.<br/>           2. Topic independence: Topics represent mutually exclusive dimensions of intervention.<br/>           3. Topic generality: each topic must be applicable across multiple comments.</p> <p><b>[Instructions]</b><br/>           Step 1 – Comment Parsing:<br/>           Identify the intent behind the user’s question.<br/>           – If the primary focus is unrelated to information demands within {topic/domain}, terminate immediately and return <i>only</i> None(reason).<br/>           – If it is related, proceed to Step 2.<br/>           Step 2 – Topic Generation:<br/>           First, identify any topics already present in the document and prioritise recognition of <i>existing topics</i>.<br/>           Second, if a newly identified topic is absent from the existing topic set, run a topic review:<br/>           ① Granularity check: verify that the new topic is a high-level topic capable of accommodating sub-topics; if not, the review fails.<br/>           ② Cross-validation: check whether the new topic overlaps with or duplicates any existing topic; if so, the review fails.<br/>           ③ Hierarchy check: verify that the new topic and the existing topics share the same hierarchical level; if a parent–child relationship exists, the review fails.<br/>           ④ If the review passes, retain the new topic and append it to the existing topic set.<br/>           Step 3 – Topic Assignment &amp; Feature Word Identification:<br/>           Based on the updated existing topic set, assign one or more directly relevant topics to the comment. Return each topic, its topic feature words, and the assignment rationale. (Different contents pointing to the same topic should be merged.)</p> <p><b>[Example]</b><br/>           {seed topics} {user comment} {output}</p> <p><b>[Input]</b><br/>           Seed topics: {seedtopics}<br/>           User comment: {text}</p> <p><b>[Output]</b><br/>           Return <i>only</i> the topic assignment result in the format: ('Topic', 'Topic feature words', 'Topic assignment rationale'). When multiple topics apply, return a composite tuple.</p> |

**Table A2.** Topic optimisation prompt template.

| Prompt Engineering Template (for Topic Optimisation)                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>[Role]</b><br>You are a topic optimisation research expert in the field of {domain}                                                                                                                                                                                                                                                                       |
| <b>[Goal]</b><br>You will receive {n} groups of topics. Each topic entry includes a topic label and topic feature words. Among the feature words, those appearing earlier carry greater interpretability for the topic.<br>Please merge topics that are nearly identical or highly similar in order to reduce redundancy and improve classification clarity. |
| <b>[Constraints]</b><br>{Topic description}                                                                                                                                                                                                                                                                                                                  |
| <b>[Input]</b><br>{candidate topics}                                                                                                                                                                                                                                                                                                                         |
| <b>[Example]</b><br>{candidate topics} {output}                                                                                                                                                                                                                                                                                                              |
| <b>[Output]</b><br>Return <i>only</i> the following content; do not append any explanation.<br>– If no merging is needed, return None.<br>– If merging is needed, return the relevant topics as a list where the last item is the merged (resulting) topic and all preceding items are the topics that participated in the merge.                            |

## References

- Feng, X.; Wang, X.; Xue, Y.; Yu, H. Analysis of the Characteristics and Evolution of Knowledge Label Networks in the Q&A Community: Taking the Zhihu Platform as an Example. *Electron. Libr.* **2023**, *41*, 242–263. [\[CrossRef\]](#)
- Li, M.; Li, Y.; Lou, W.; Chen, L. A Hybrid Recommendation System for Q&A Documents. *Expert Syst. Appl.* **2020**, *144*, 113088. [\[CrossRef\]](#)
- Pan, X.; Tang, Z.; Liu, Y.; Ren, J. Analysis on Childhood Fever Health Information Seeking Behaviors in Online Health Community Using a Mixed-Methods Approach. *Digit. Health* **2024**, *10*, 20552076241282622. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yang, Y.; Ma, Y.; Wu, G.; Guo, Q.; Xu, H. The Insights, “Comfort” Effect and Bottleneck Breakthrough of “E-Commerce Temperature” during the COVID-19 Pandemic. *J. Theor. Appl. Electron. Commer. Res.* **2022**, *17*, 1493–1511. [\[CrossRef\]](#)
- Meng, F.; Jia, Y. Data-Driven Prioritization of User Requirements in Health E-Commerce: An Explainable Machine Learning Study. *J. Theor. Appl. Electron. Commer. Res.* **2026**, *21*, 104. [\[CrossRef\]](#)
- Pirolli, P.; Card, S. Information Foraging. *Psychol. Rev.* **1999**, *106*, 643–675. [\[CrossRef\]](#)
- Li, Y.; Li, M.; Zhang, X.; Ding, J. Dynamic Q&A Multi-Label Classification Based on Adaptive Multi-Scale Feature Extraction. *Appl. Soft Comput.* **2025**, *170*, 112740. [\[CrossRef\]](#)
- Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
- Wang, X.; Lu, X.; Guo, Y.; Zhu, L.; Wang, F. Multimodal Topic Models for Brand Competitive Analysis. *Electron. Commer. Res.* **2026**. [\[CrossRef\]](#)
- Yang, Y.; He, W.; Gao, C.; Xu, Z.; Xia, X.; Liu, C. TopicAns: Topic-informed Architecture for Answer Recommendation on Technical Q&A Site. *ACM Trans. Softw. Eng. Methodol.* **2023**, *33*, 20. [\[CrossRef\]](#)
- Gatto, J.; Basak, M.; Srivastava, Y.; Bohlman, P.; Preum, S.M. Scope of Large Language Models for Mining Emerging Opinions in Online Health Discourse. *arXiv* **2024**, arXiv:2403.03336.
- Pham, C.M.; Hoyle, A.; Sun, S.; Resnik, P.; Iyyer, M. TopicGPT: A Prompt-based Topic Modeling Framework. *arXiv* **2023**, arXiv:2311.01449. [\[CrossRef\]](#)
- Reuter, A.; Khadka, B.; Thielmann, A.; Weisser, C.; Fischer, S.; Säfken, B. GPTopic: Dynamic and Interactive Topic Representations. *arXiv* **2024**, arXiv:2403.03628. [\[CrossRef\]](#)
- Qian, Y.; Gui, W. Identifying Health Information Needs of Senior Online Communities Users: A Text Mining Approach. *Aslib J. Inf. Manag.* **2020**, *73*, 5–24. [\[CrossRef\]](#)
- Oh, S.; Zhang, Y.; Park, M.S. Cancer Information Seeking in Social Question and Answer Services: Identifying Health-Related Topics in Cancer Questions on Yahoo! *Inf. Res.* **2016**, *21*, 718. Available online: <https://informationr.net/ir/21-3/paper718.html> (accessed on 27 July 2026). [\[CrossRef\]](#) [\[PubMed\]](#)
- Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794. [\[CrossRef\]](#)

17. Zhao, T.; Zhang, Y.; Miao, D.; Pedrycz, W. Fine-Grained Local Label Correlation for Multi-Label Classification. *Knowl.-Based Syst.* **2025**, *314*, 113210. [CrossRef]
18. Jeong, D.; Hwang, J.; Choi, Y.; Kim, Y.Y. Automatic Seed Word Selection for Topic Modeling. *IEEE Access* **2025**, *13*, 31269–31285. [CrossRef]
19. Wang, R.; Li, Z.; Zheng, Y.; Luo, G.; Zhou, Y.; Zhou, Y. Topic Modeling and Alignment with Large Language Models for Multi-Labeled Text Corpora. *Expert Syst. Appl.* **2026**, *307*, 130958. [CrossRef]
20. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:cs/2005.14165. [CrossRef]
21. Wang, H.; Prakash, N.; Hoang, N.K.; Hee, M.S.; Naseem, U.; Lee, R.K.W. Prompting Large Language Models for Topic Modeling. *arXiv* **2023**, arXiv:cs/2312.09693. [CrossRef]
22. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42. [CrossRef]
23. Mu, Y.; Dong, C.; Bontcheva, K.; Song, X. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N., Eds.; Association for Computational Linguistics: Torino, Italy, 2024; pp. 10160–10171.
24. Spielberger, G.; Artinger, F.M.; Reb, J.; Kerschreiter, R. Retrieval Augmented Generation for Topic Modeling in Organizational Research: An Introduction with Empirical Demonstration. *arXiv* **2025**, arXiv:2502.20963 [CrossRef]
25. Li, C.; Chen, S.; Xing, J.; Sun, A.; Ma, Z. Seed-Guided Topic Model for Document Filtering and Classification. *ACM Trans. Inf. Syst.* **2019**, *37*, 9. [CrossRef]
26. Jagarlamudi, J.; Daumé, H., III; Udupa, R. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*; Daelemans, W., Ed.; Association for Computational Linguistics: Avignon, France, 2012; pp. 204–213.
27. Zhang, Y.; Zhang, Y.; Michalski, M.; Jiang, Y.; Meng, Y.; Han, J. Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. In *WSDM '23: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 429–437. [CrossRef]
28. Zhihu. Available online: <https://www.zhihu.com/> (accessed on 27 July 2026).
29. Ding, L.; Xiao, J.; Zhou, S.; Yang, C.; Han, J. Topic-Oriented Open Relation Extraction with A Priori Seed Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 13834–13845. [CrossRef]
30. Dieng, A.B.; Ruiz, F.J.R.; Blei, D.M. Topic Modeling in Embedding Spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [CrossRef]
31. Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*; Barzilay, R., Johnson, M., Eds.; Association for Computational Linguistics: Edinburgh, Scotland, UK, 2011; pp. 262–272.
32. Kosar, A.; De Pauw, G.; Daelemans, W. Comparative Evaluation of Topic Detection: Humans vs. LLMs. *Comput. Linguist. Neth. J.* **2024**, *13*, 91–120.
33. Mu, Y.; Bai, P.; Bontcheva, K.; Song, X. Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling. *arXiv* **2024**, arXiv:2405.00611. [CrossRef]
34. Kumar, S.; Le, T. FoTo: Targeted Visual Topic Modeling for Focused Analysis of Short Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N., Eds.; Association for Computational Linguistics: Torino, Italy, 2024; pp. 7406–7416.
35. Zhao, Z.; Chen, Y.; Bangash, A.A.; Adams, B.; Hassan, A.E. An Empirical Study of Challenges in Machine Learning Asset Management. *Empir. Softw. Eng.* **2024**, *29*, 98. [CrossRef]
36. Rijcken, E.; Scheepers, F.; Zervanou, K.; Spruit, M.; Mosteiro, P.; Kaymak, U. Towards Interpreting Topic Models with ChatGPT. In *Proceedings of the 20th World Congress of the International Fuzzy Systems Association*, Daegu, Republic of Korea, 20–24 August 2023.
37. Wang, J.; Qu, Z.; Wu, S.; Skitmore, M.; Ma, L. Automatic Classification of Construction Accident Reports Using BERTopic-GLDA Approach. *Eng. Constr. Archit. Manag.* **2025**, *EarlyCite*. [CrossRef]
38. Zheng, M.; Jiang, K.; Xu, R.; Qi, L. An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification. *IEEE Access* **2023**, *11*, 92273–92284. [CrossRef]
39. Huang, J.; Xie, Y.; Meng, Y.; Shen, J.; Zhang, Y.; Han, J. Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion. In *Proceedings of WWW '20: The Web Conference 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 2188–2198. [CrossRef]

40. Henning, S.; Beluch, W.; Fraser, A.; Friedrich, A. A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*; Vlachos, A., Augenstein, I., Eds.; Association for Computational Linguistics: Dubrovnik, Croatia, 2023; pp. 523–540. [[CrossRef](#)]
41. Liu, X.; Liu, Y.; Qian, Y.; Jiang, Y.; Ling, H. Learning Consumer Preferences through Textual and Visual Data: A Multi-Modal Approach. *Electron. Commer. Res.* **2025**, *25*, 2955–2984. [[CrossRef](#)]
42. Laureate, C.D.P.; Buntine, W.; Linger, H. A Systematic Review of the Use of Topic Models for Short Text Social Media Analysis. *Artif. Intell. Rev.* **2023**, *56*, 14223–14255. [[CrossRef](#)] [[PubMed](#)]
43. Parlak, I.B.; Şahin, M.Ş.; Acarman, T.; Adel, M.; Bourennane, S. Quantification of Overlapping and Network Complexity in News: Assessment of Top2Vec and Fuzzy Topic Models. *Appl. Sci.* **2025**, *15*, 9627. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.