IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 June 2024, accepted 18 July 2024, date of publication 29 July 2024, date of current version 26 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434648

METHODS

Building Lightweight Domain-Specific Consultation Systems via Inter-External Knowledge Fusion Contrastive Learning

JIABIN ZHENG[1], HANLIN WANG[2], AND JIAHUI YAO[2]
[1]School of Computer Science, Peking University, Beijing 100871, China
[2]Institute of Social Science Survey, Peking University, Beijing 100871, China
Corresponding author: Jiahui Yao (isssyaojh@pku.edu.cn)

ABSTRACT Large language models (LLMs) have demonstrated their vast potential and value in natural language processing tasks and beyond. However, when these models are applied to develop consultation systems for industrial domains, such as e-government, intelligent diagnosis, and legal consultancy, they encounter many unresolved technical issues. These include the vast scale of the models, lack of tight fusion with existing industry knowledge, along with occurrences of model hallucinations and inadequate explainability. Unlike general-purpose dialogue systems, building a consultation system for a specific industrial domain requires not only the integration of extensive external knowledge from the Internet but also the incorporation of precise, specialized knowledge from specific industries, making the challenges even more complex. In response to these challenges, we propose the Inter-External Knowledge Fusion Contrastive Learning Technique. This technique facilitates the integration of internal industry knowledge with widely accessible external knowledge from the Internet and provides a universal framework for building lightweight, domain-specific consultation systems. Utilizing this technique enables the straightforward creation of precise, professional, and constantly updated domain-specific consultation systems applicable across various industries. Overcoming the inherent limitations associated with LLMs, this technique achieves a performance level comparable to LLMs. To validate the effectiveness of our proposed technique, we conducted extensive experiments in the development of real-world industry consulting systems. By testing on seven real datasets covering diverse tasks, we demonstrate the system's exceptional performance: our lightweight consultation system utilizes only 4% of the parameters of an LLM but achieves over 90% of its performance level.

INDEX TERMS Knowledge fusion, domain-specific consultation system, contrastive learning, large language models, lightweight system.

I. INTRODUCTION

With the rapid advancement of artificial intelligence technology, Large Language Models (LLMs) have demonstrated significant application potential across a variety of domains and tasks, with their pre-training on massive datasets endowing them with exceptional language understanding and generation capabilities [1]. However, in specialized industries such as e-government (as shown in Figure 1), intelligent diagnosis, and legal consultancy, these models often exhibit limited performance due to insufficient pre-training on specific domain knowledge. These models tend to prioritize general knowledge, and may overlook crucial domain-specific information. Moreover, training LLMs from scratch incurs prohibitively high costs for many organizations. By contrast, fine-tuning existing models on domain-specific data emerges as a viable alternative. Researchers have provided new insights and methodologies by fine-tuning LLMs with specific datasets within consultation systems in domains such as e-government, intelligent diagnosis, and legal consultancy, significantly enhancing the system's efficiency and accuracy in answering professional questions.

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

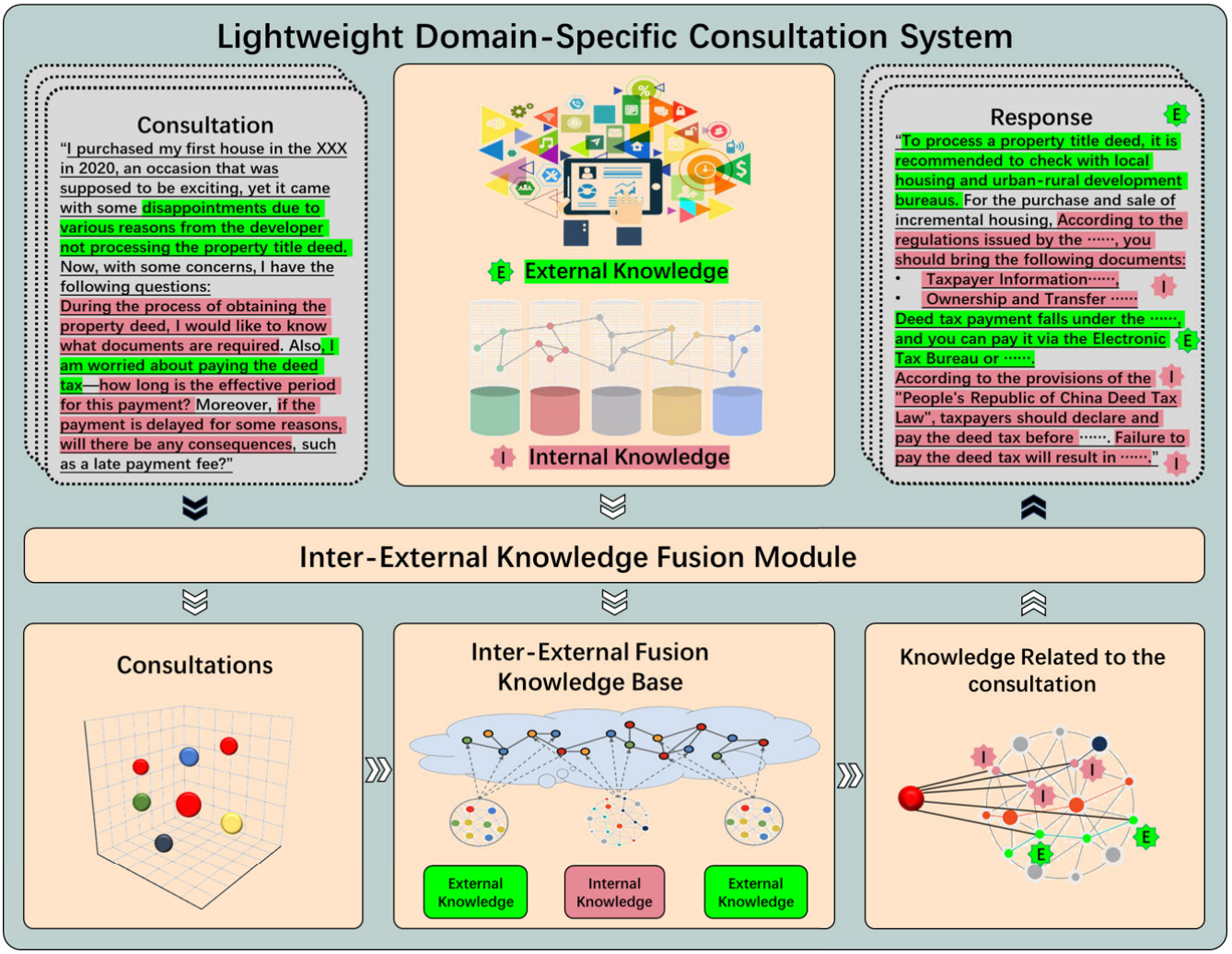

FIGURE 1. A real-world example of a domain-specific consultation system. In order to address intricate user consultations, the system needs to combine specialized internal knowledge and external information from the Internet to deliver accurate responses.

For instance, in the domain of e-government, Han et al. [2] significantly improved the efficiency and accuracy of handling citizen inquiries by fine-tuning LLMs with specialized datasets. In healthcare, Li et al. [3]. demonstrated how analyzing extensive patient-physician conversations could refine the accuracy of medical consultations. Within the legal sector, Cui et al. [4] developed an LLM specifically designed for interpreting legal terminology, incorporating case law and other contextual enhancements to improve explanation quality and reduce misinformation. These efforts not only validate the applicability of LLMs in professional domains but also provide valuable insights and inspiration for our research.

Despite the promising insights offered by LLMs in professional domains, they also come with significant challenges. The main issues include the high costs and complexity of using and updating these models due to their large number of parameters. This is particularly concerning in areas that require frequent updates and the processing of complex data. Additionally, LLMs sometimes produce "hallucinations," generating information that is inconsistent with facts or entirely incorrect. In industrial domains, this risk of erroneous information can lead to severe consequences, making this issue especially serious.

However, when focusing on domain-specific consultation systems, beyond the challenges associated with the use of LLMs, the inherent characteristics of these systems themselves introduce a series of additional challenges. These include the disparity between colloquial consultations and the formal presentation of professional knowledge, as well as the core challenge of achieving fine-grained matching of complex information. On one hand, users' consultations often lean towards being colloquial, utilizing informal expressions, whereas knowledge in professional domains is typically presented in written form or filled with specialized terminology. Bridging this gap between colloquial speech and professional text is essential for the systems to ensure accurate understanding and responses.

On the other hand, both user consultations and professional knowledge usually involve multiple points of information, requiring the consultation systems to have the capability for fine-grained matching. This not only involves filtering out a vast amount of irrelevant information but also accurately aligning with related professional knowledge points, thereby providing precise and comprehensive answers.

In this paper, we propose a novel technique called "IEK-Fusion: Inter-External Knowledge Fusion Contrastive Learning" to address the technical issues encountered by LLMs when deployed in industrial domain consultation systems. This technique facilitates the integration of internal industry knowledge with widely accessible external knowledge from the Internet and provides a universal framework for building lightweight, domain-specific consultation systems. Specifically, through a multi-view contrastive learning framework, it constructs a fusion network for diverse types of knowledge, systematically optimizing the representations of both internal industry knowledge and external internet knowledge, while establishing dynamic comparative relationships between different knowledge representations. Additionally, the Gated Hierarchical Knowledge Encoder proposed in this paper enables precise fine-grained knowledge extraction and noise elimination, ensuring the meticulous alignment and fusion of internal and external knowledge in complex scenarios. Our objective is to build lightweight, precise, and professional consultation systems that can efficiently process domain-specific consultations. To validate the effectiveness of our proposed technique, we conducted extensive experiments in the development of real-world industry consulting systems. Utilizing seven real datasets that span various industry tasks, we demonstrate the system's exceptional performance: despite employing merely 4% of the parameter of an LLM, it manages to deliver performance levels exceeding 90% of those achieved by the LLM.

The main contributions of this research are summarized as follows:

- We proposed a novel technique called "**IEK-Fusion: Inter-External Knowledge Fusion Contrastive Learning**," which effectively facilitates the integration of internal industry knowledge and external knowledge from the Internet, ensuring that domain-specific consultation systems respond to users in a more comprehensive and professional manner.
- Based on this technique, we proposed **a universal framework for building lightweight domain-specific consultation systems applicable across various industries.**
- We conducted extensive experiments in the development of real-world industry consulting system. By testing on seven real datasets covering diverse tasks, we demonstrate the system's exceptional performance: **our lightweight consultation system utilizes only 4% of the parameters of an LLM but achieves over 90% of its performance level.**

This article is organized as follows. Section II provides an overview of the technologies involved in this paper. Section III provides a detailed introduction to our IEK-Fusion technique. Section IV describes the experimental setup used to validate the effectiveness of the proposed technique, along with the presentation of results and their analysis. Finally, the paper concludes with a summary of key findings and explores potential avenues for future research.

II. RELATED WORK

A. LLMS IN SPECIALIZED DOMAINS

Large Language Models (LLMs) have become increasingly popular in various natural language processing tasks. These models, built upon the Transformer architecture [5], are typically trained on general public information, enabling them to handle a wide range of tasks with versatility. However, when it comes to specialized domains, such as Public Service Consultation on Welfare Issues, the need for domain-specific LLMs becomes apparent. One approach to address this issue is to repurpose LLMs for specialized domains. Tag-LLM, a model-agnostic method introduced in recent research, aims to repurpose general LLMs into effective task solvers for specialized domains [6]. These models are designed to capture the unique jargon and context of a specific industry or use case, making them more effective in generating relevant and accurate outputs. While generic LLMs like GPT-3 [7] have their place, the deployment of specialized LLMs in specific domains can lead to more tailored and efficient results. By fine-tuning or grounding these models in domain-specific datasets, they can achieve higher performance levels in specialized tasks. In conclusion, the adaptation of LLMs to specialized domains is a crucial area of research that can significantly enhance the capabilities of these models in various applications. By repurposing LLMs, developing domain-specific models, and utilizing a combination of specialized LLMs, researchers and developers can unlock the full potential of these powerful language models [8]. Despite the significant improvements offered by specialized domain-specific Large Language Models (LLMs) over their general-purpose counterparts, these models still present substantial challenges, such as high costs associated with the use, maintenance, and updating due to their immense model parameters.

B. DOMAIN-SPECIFIC CONSULTATION SYSTEMS

The development of domain-specific consultation systems has been a topic of interest in various domains. Istiadi et al. [9] developed an infectious disease expert system using the Dempster Shafer method, which demonstrated high accuracy and usability in health services recommendations. Huang et al. [10] introduced a framework for an AI-based medical consultation system using knowledge graph embedding and reinforcement learning components to support diagnosis based on patient evidence. Moreover, Liu and Xu [11] addressed challenges in building knowledge

graphs from ''dirty'' clinical electronic medical records for intelligent consultation applications by proposing a data cleaning framework. Finally, Chen et al. [12] highlighted the importance of machine learning in improving the efficiency of automatic medical consultation systems and proposed frameworks for doctor-patient dialogue understanding and task-oriented interaction. Research on domain-specific consultation systems predominantly targets hot domains such as e-government, intelligent diagnosis, and legal consultancy. However, there is currently a lack of a universal solution that can be applied across all domains.

C. MULTI-VIEW LEARNING

Multi-view learning has gained significant attention in recent years as it provides a powerful approach to analyze data from multiple views. Tang et al. [13] explored multi-view learning technique with the LINEX loss for pattern classification, emphasizing the importance of loss functions in multi-view learning tasks. Zhang et al. [14] introduced a joint representation learning approach for multi-view subspace clustering, aiming to learn view-specific representations and low-rank tensor representations in a unified framework. Phan et al. [15] focused on multi-view learning for audio and music classification tasks, demonstrating the superiority of their proposed multi-view network over single-view and multi-view baselines. Lastly, Chen et al. [16] introduced a distinguishable feature fusion approach for rumor detection using multi-view learning techniques. Overall, the literature on multi-view learning showcases a variety of approaches and frameworks that aim to leverage multiple views of data to enhance representation learning, clustering, and classification tasks. Researchers continue to explore novel methods to address the challenges associated with multi-view data analysis and to improve the performance of multi-view learning algorithms. Previous multi-view learning has primarily focused on multimodal data, while exploration of different forms of multi-view learning within a single modality remains insufficiently researched.

III. METHODOLOGY

This section introduces our core technique, ***IEK-Fusion***: *Inter-External Knowledge Fusion Contrastive Learning*. Based on this technique, we propose a Universal Framework for Building lightweight Domain-Specific Consultation Systems. An overview of the technique and the framework is presented in Figure 2. This section is structured as follows:

- **Part A:** Introduces the process of constructing diverse views of domain-specific knowledge, aimed at providing a solid knowledge foundation for building a domain-specific consultation system.
- **Part B:** To accurately encode various types of knowledge, We propose two types of encoders for accurately encoding various types of knowledge. The first is the Hierarchical Knowledge Encoder (HKE), a hierarchical BERT structure. The second is the Gated Hierarchical Knowledge Encoder (GHKE), which enhances the HKE with gated linear units. Differences in model architecture and training approaches between these two encoders are discussed in detail.
- **Part C:** Following our discussion on encoding internal and external knowledge, this part addresses the fusion of internal and external knowledge. We propose the *IEK-Fusion*, a technique that integrates multi-view contrastive learning with knowledge fusion, to achieve precise amalgamation of internal and external knowledge.

Through this section, we aim to demonstrate how IEK-Fusion can effectively enhance the performance of the domain-specific consultation system, improving its accuracy and efficiency by precisely integrating internal and external knowledge.

A. CONSTRUCTION OF DIVERSE VIEWS OF KNOWLEDGE

The domain-specific consultation system solution presented in this paper is centered on both internal and external knowledge. This section provides a detailed explanation of these two types of knowledge.

1) VIEWS OF INTERNAL KNOWLEDGE

The domain-specific consultation system distinguishes itself from traditional dialogue systems in two fundamental ways. Firstly, user inquiries are confined to domain-specific consultations rather than general information or assistance requests. Secondly, the responses provided by a domain-specific consultation system depend heavily on a broad spectrum of specialized knowledge within that domain. This section introduces the concept of internal knowledge, which includes a range of professional knowledge directly relevant to the domain of the system, necessary for supporting precise responses in the consultation system. Examples of this include government functions and policies within the realm of e-government, symptoms and treatment plans within intelligent medical diagnosis, and legal statutes within the legal consultancy domain.

Building on the definition of internal knowledge as discussed, internal knowledge consists of specialized information derived from authoritative sources within specific professional domains. It is cultivated through extensive research and practical experience by experts, lending it a high level of credibility and authority. Such knowledge is pivotal not only for enhancing understanding of specialized domains but also for providing crucial guidance on challenges within those domains.

The development of an internal knowledge base is fundamental for domain-specific consultation systems. It forms the backbone that enables the system to deliver precise and dependable responses to user consultations. This paper

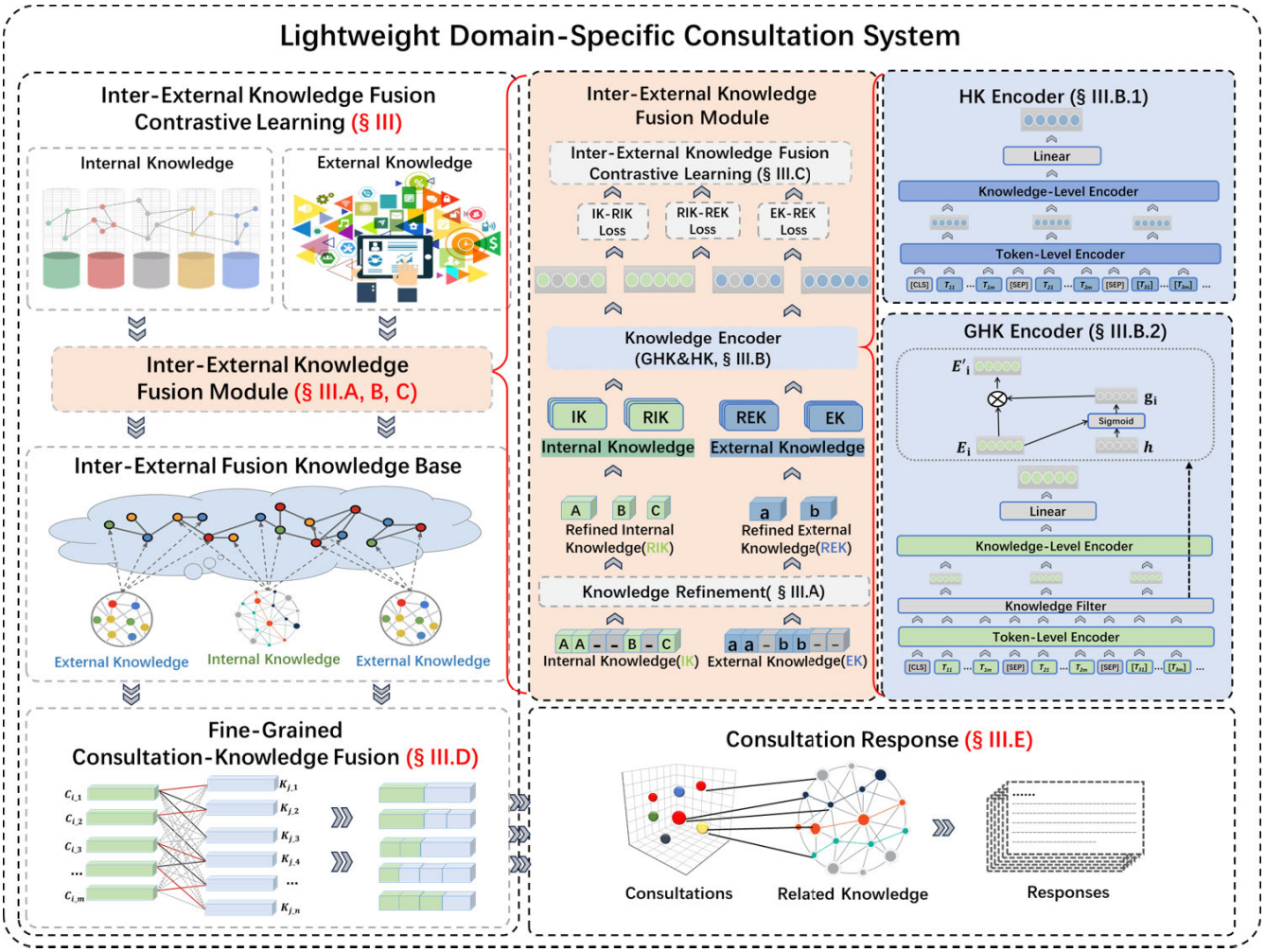

FIGURE 2. Overview of the IEK-fusion technique and the universal framework for building lightweight domain-specific consultation systems.

categorizes internal knowledge into two distinct types, detailed as follows:

- Internal Knowledge (IK): IK refers to the foundational data and information that have not undergone any additional processing or refinement. IK includes raw facts, unedited data, and primary documents collected directly from reliable and authoritative sources within the domain. This form of knowledge is essential as it provides an unaltered base from which further analysis and synthesis can be performed, ensuring that the consultation system has access to the most authentic and accurate information available.
- Refined Internal Knowledge (RIK): RIK is the result of further refinement and annotation based on IK. The process of creating RIK involves a comprehensive analysis and detailed annotation of the information contained in IK. This methodical breakdown transforms broad data into finer-grained, actionable information units. Such enhanced granularity facilitates the provision of more precise and targeted responses, optimizing the consultation system's effectiveness in addressing specific user inquiries.

2) VIEWS OF EXTERNAL KNOWLEDGE

In domain-specific consultation systems, in addition to relying on internal knowledge for accurate professional information support, it is also necessary to provide reasonable responses to user consultations. We propose a solution similar to retrieval-based dialogue systems to obtain responses, specifically by analyzing historical interaction data from the consultation system published in the domain, which consists of a vast number of consultation-response pairs. We refer to these extensive consultation-response pairs as external knowledge. Adequate external knowledge ensures that the system can handle a variety of flexible consultations. Similar to internal knowledge, we have also introduced two types of external knowledge, which are described in detail below:

- External Knowledge (EK): Refers to the raw, unprocessed consultation-response pairs collected from historical interactions within the domain-specific consultation systems. These pairs capture the real-world exchanges between users and the system, providing a foundational dataset that reflects genuine user inquiries and the corresponding system responses. EK serves as a vital resource for understanding the types of questions users ask and how effectively the system can address these queries without any modifications or enhancements.
- Refined External Knowledge (REK): involves the enhancement of original external knowledge by filtering out colloquial expressions and noise, preserving only the core, valuable information. This process focuses on refining the clarity and usability of the data, stripping away irrelevant details that do not contribute to the system's effectiveness. Additionally, the consultation content is broken down into fine-grained semantic units. This decomposition allows for more precise and targeted responses by enabling the system to address specific aspects of a user's inquiry with greater accuracy and relevance.

B. HIERARCHICAL KNOWLEDGE ENCODERS

To precisely capture the semantic information of both internal and external knowledge, we utilize a hierarchical BERT encoder [17] as our foundational architecture. This architecture is specifically designed to deeply understand and effectively harness the complex details present in both types of knowledge through hierarchical processing.

More specifically, we detail two specialized hierarchical encoders: the Hierarchical Knowledge Encoder (HKE) and the Gated Hierarchical Knowledge Encoder (GHKE). The HKE is dedicated to encoding and processing refined knowledge, while the GHKE focuses on encoding original knowledge. These encoders integrate and synthesize information through their distinct mechanisms, significantly boosting the model's expressiveness and precision in domain-specific contexts.

1) HIERARCHICAL KNOWLEDGE ENCODER

The Hierarchical Knowledge Encoder (HKE), a hierarchical encoder built upon the BERT architecture, is specifically designed to handle domain-specific knowledge with intricate structures. Unlike traditional BERT encoders, the HKE integrates a hierarchical processing mechanism by encoding different sub-clauses separately at the first layer, and then synthesizing these encodings at a second layer, enabling the model to capture diverse semantic information from different sub-clauses within complete sentences. This innovative approach allows the HKE to comprehend not only the fundamental semantics of the text but also finer-grained details. As a result, it provides more comprehensive and precise knowledge representations, making it exceptionally suited for addressing complex consultations. In this paper, we employ the Hierarchical Knowledge Encoder (HKE) to encode refined knowledge.

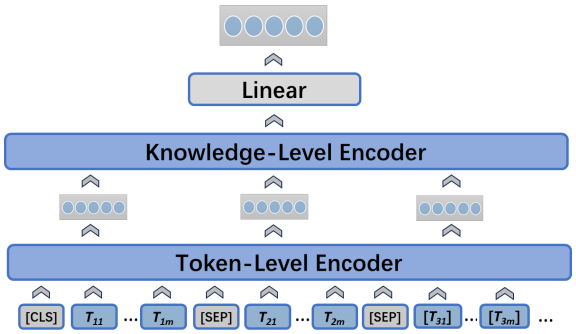

FIGURE 3. The architecture of the hierarchical knowledge encoder.

We define the Hierarchical Knowledge Encoder (HKE) based on the BERT architecture, specifically tailored for handling complex, domain-specific text as follows:

The encoder processes input texts by first tokenizing them into smaller semantic units such as words or sub-tokens. Each sentence or sub-clause is then encoded independently at the token level using a special module called the **Token-Level Encoder**. Sub-clauses are demarcated by the special token `[SEP]`.

Mathematically, the encoding of the i-th sub-clause, which comprises tokens $T_{i1}, T_{i2}, \ldots, T_{in}$, can be represented as follows:

$$E_i = \text{BERT}(\texttt{[SEP]}, T_{i1}, T_{i2}, \ldots, T_{in}, \ldots) \tag{1}$$

In this formula, E_i denotes the vector representation of the i-th sub-clause. Importantly, the output corresponding to the `[SEP]` token at the end of each sub-clause is used as the representative vector for that sub-clause. The vectors E_i for each sub-clause are further processed to form the final comprehensive representation of the entire input text. This layered and detailed approach ensures that each sub-clause is treated as an independent unit while retaining the contextual continuity provided by BERT's architecture.

The second layer of the hierarchy, known as the **Knowledge-Level Encoder**, synthesizes these individual sub-clause representations into a comprehensive output, This output is then passed through an additional linear layer to enhance and refine the final representation:

$$O = \text{FNN}(\text{BERT}(E_1, E_2, \ldots, E_n)) \tag{2}$$

where O is the final output of the HKE. The linear layer serves to linearly transform the synthesized output, allowing for further customization and refinement based on specific domain requirements. This final representation provides a unified and highly contextual output that integrates both detailed and abstract semantic information from the encoded sub-clauses.

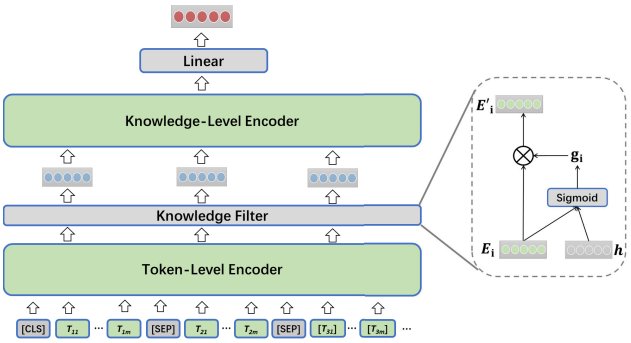

FIGURE 4. The architecture of the gated hierarchical knowledge encoder.

The architectural diagram of the Hierarchical Knowledge Encoder (HKE) is shown in Figure 3. This diagram illustrates the multi-layered processing structure of the HKE, including the token-level encoding at the first layer, the sentence-level synthesis at the second layer, and the additional linear transformation layer that refines the final output. The architectural diagram of the Hierarchical Knowledge Encoder (HKE) is shown in Figure 3. This diagram illustrates the multi-layered processing structure of the HKE, including the token-level encoding at the first layer, the sentence-level synthesis at the second layer, and the additional linear transformation layer that refines the final output.

2) GATED HIERARCHICAL KNOWLEDGE ENCODER

As introduced in Section III, our encoding subjects include four types: IK, RIK, EK, and REK. Both RIK and REK have undergone fine-grained sub-clausal annotation and removal of irrelevant information, which clearly delineates which parts constitute an independent semantic unit. However, such annotations are absent in the processing of IK and EK. To address this, we have adapted the structure of the Hierarchical Knowledge Encoder (HKE) by incorporating knowledge filter module, a Gated Linear Unit (GLU) following the Token-Level Encoder. This gating mechanism enables the model to autonomously learn which sequences of sub-clauses form coherent semantic modules and to identify dispensable knowledge. This encoder is referred to as the Gated Hierarchical Knowledge Encoder (GHKE). The integration of the GLU provides GHKE with a versatile mechanism for information filtering and fusion, enhancing not only the model's ability to filter out irrelevant information but also its capability to recognize and capture fine-grained knowledge.

The architectural diagram of the Gated Hierarchical Knowledge Encoder (GHKE) is illustrated in Figure 4. This figure provides a visual representation of the complex structure of GHKE, incorporating the integration of the Gated Linear Unit (GLU) for advanced information modulation and filtering.

The gating mechanism is crucial for controlling the flow of information based on the context and the relevance of the input data. We consider E_i as the output vector from the original sub-clause or token sequence, and s as an additional state information. The gate g_t is computed as follows:

$$g_t = \text{sigmoid}(W_g[E_i, s] + b) \tag{3}$$

Here, g_t represents the gating coefficient, computed using a sigmoid activation function, ensuring the output is between 0 and 1. The weight matrix W_g and the bias b are parameters learned during training, which transform the concatenated vector of E_i and s. The purpose of this gate is to modulate the influence of the input vector E_i based on the context provided by s.

The gated output $\tilde{E}_i$ is then computed by applying the gate g_t to the vector E_i:

$$\tilde{E}_i = g_t \cdot E_i \tag{4}$$

This operation results in $\tilde{E}_i$, which is a contextually modulated version of E_i, allowing the model to selectively emphasize or deemphasize certain features based on the gate's output. This mechanism enhances the model's ability to focus on relevant features and ignore irrelevant or less important information.

C. INTER-EXTERNAL KNOWLEDGE FUSION CONTRASTIVE LEARNING

In this paper, the construction of domain-specific consultation systems necessitates precise fusion of internal and external knowledge based on user consultations. Notably, since external knowledge encompasses historical consultations which are analogous to real-time consultations, their integration poses fewer challenges. However, the specialized domain-specific expressions inherent in internal knowledge present significant disparities from user consultations, making their integration particularly challenging.

Given that consultations and external knowledge originate from the same source, this paper achieves precise alignment between consultations and internal knowledge through the effective fusion of both knowledge types. To develop a lightweight and accurate consultation system, the system not only fuses internal and external knowledge but also aligns the original and refined content within each category, namely IK, RIK, EK, and REK. As a result, the system bypasses the need for complex fine-grained segmentation and intensive content distillation of the original knowledge. By directly utilizing the original consultations along with internal and external knowledge, the system can deliver accurate responses efficiently. Considering the need to simultaneously align four distinct types of knowledge, we propose a novel technique that incorporates both internal knowledge and external knowledge correlations to create a robust and unified knowledge representation. As illustrated in Figure 5, our technique exploiting multi-view contrastive

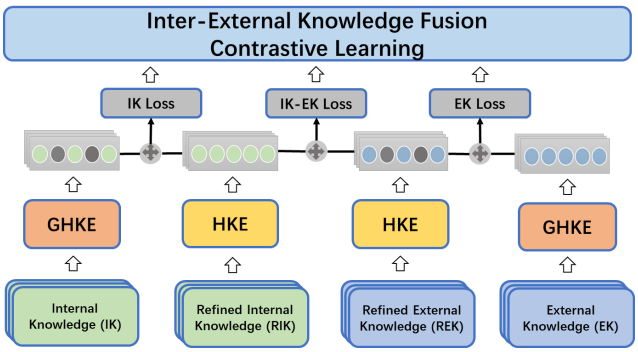

FIGURE 5. The overview of the IEK-Fusion technique, utilizing the HKE and GHKE encoders to encode four knowledge views. Through multi-view contrastive learning technique, alignment among the four knowledge views is achieved.

learning strategies [18] that strengthen the representations not just between but also within the different knowledge categories.

This section delineates the contrastive loss functions utilized for aligning the various forms of knowledge in our multi-view learning framework.

Following InfoNCE [19], The contrastive loss for internal knowledge, represented by L_{IK}, is formulated to minimize the distance between the original internal knowledge representations, h_{IK}, and their refined counterparts, h_{RIK}. This is captured by the following equation:

$$L_{IK} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp\left(h_{IK}^{i\top}h_{RIK}^{i}/\tau\right)}{\sum_{j=1}^{N}\exp\left(h_{IK}^{i\top}h_{RIK}^{j}/\tau\right)} \tag{5}$$

Similarly, the contrastive loss for external knowledge, L_{EK}, aims to align the original external knowledge, h_{EK}, with the refined external knowledge, h_{REK}, as expressed by:

$$L_{EK} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp\left(h_{EK}^{i\top}h_{REK}^{i}/\tau\right)}{\sum_{j=1}^{N}\exp\left(h_{EK}^{i\top}h_{REK}^{j}/\tau\right)} \tag{6}$$

The alignment of internal and external knowledge is quantified by L_{IK-EK}, which measures the congruence between refined internal knowledge and refined external knowledge, ensuring semantic consistency across the two knowledge domains:

$$L_{IK-EK} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp\left(h_{RIK}^{i\top}h_{REK}^{i}/\tau\right)}{\sum_{j=1}^{N}\exp\left(h_{RIK}^{i\top}h_{REK}^{j}/\tau\right)} \tag{7}$$

The temperature parameter τ plays a critical role in regulating the sharpness of the distribution and N denotes the number of internal/external knowledge in the batch.

The combined fusion loss, L_{fusion}, is a weighted sum of the three contrastive losses. The weights α, β, and γ are hyperparameters that are empirically determined to balance

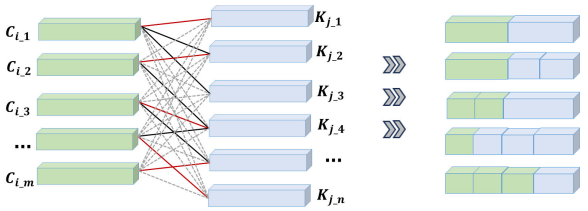

FIGURE 6. The illustration of the fine-grained consultation-knowledge Fusion method.

the contributions of each individual loss function to the overall objective:

$$L_{\text{fusion}} = \alpha L_{IK} + \beta L_{EK} + \gamma L_{IK-EK} \tag{8}$$

D. FINE-GRAINED CONSULTATION-KNOWLEDGE FUSION

To enhance the precision of the consultation system, we introduce the Fine-Grained Consultation-Knowledge Fusion algorithm. This algorithm not only considers the global semantic vectors when determining the match between consultation and knowledge but also takes into account the semantic vectors of each sub-clause. Specifically, we construct a weighted bipartite graph based on consultation and knowledge each containing multiple sub-clauses, where each sub-clause of the consultation or knowledge is treated as a node in the graph, and the vector similarity between sub-clauses is represented as the edges of the bipartite graph. Subsequently, we employ a maximum weight matching algorithm to identify the optimal alignment within the bipartite graph. Through this method, we can not only accurately determine whether consultation and knowledge match but also provide a more detailed list of the sub-information matches between the two, thus yielding a more precise matching score. An example of the algorithm is illustrated in Figure 6. The diagram displays the fine-grained matching between consultation$_i$ and knowledge$_j$, where all solid lines represent fine-grained information matches. The red lines indicate the optimal matching results. In this example, the five red lines signify that there are five corresponding fine-grained information matches between the consultation and the knowledge, resulting in a matching score of 5.

E. THE UNIVERSAL FRAMEWORK FOR BUILDING LIGHTWEIGHT DOMAIN-SPECIFIC CONSULTATION SYSTEMS

Based on the IEK-Fusion technique described in this paper, we propose a universal framework for building a lightweight domain-specific consultation system. This framework aims to integrate internal and external knowledge through IEK-Fusion technique and align it with user consultations to provide immediate and accurate responses. Specifically, the system generates a universal response based on the alignment between the consultation and external knowledge, while obtaining more precise knowledge support based on the alignment between the consultation and internal

industry knowledge. These two sources together form a comprehensive system response, ensuring that users receive integrated answers that combine the practicality of external consultation data with the depth of internal knowledge. This dual-response mechanism makes the system highly efficient and effective for real-time applications. The overall of the framework is shown in Figure 2.

IV. EXPERIMENTS

In this section, we conduct real experiments using a complex consultation system within the industrial domain of e-government. We validate the effectiveness of the system on seven different real-world datasets involving multiple tasks.

A. CONSTRUCTION OF A MULTI-VIEW KNOWLEDGE BASE

1) ORIGINAL INTERNAL AND EXTERNAL KNOWLEDGE BASE

- Original Internal Knowledge Base: In the domain-specific consultation system of the e-government, internal knowledge should encompass government-related policies and regulations, administrative management systems, public service processes, government service guidelines, responsibilities and functions of government departments, as well as specific content and operational regulations for various public service projects. These knowledge points serve as fundamental materials required by government consultation systems, enabling the provision of accurate and authoritative consultation answers to users.

  The main sources of internal knowledge include government official websites and portal websites, including those of national, provincial, municipal, and county-level government departments. They primarily cover government policies, regulations, documents, announcements, etc.

  Internal knowledge plays a crucial role in the system. First, it provides an accurate and comprehensive data foundation for the e-government consultation system, enabling it to respond based on the latest information on government policies and responsibilities. Moreover, internal knowledge also enables the intelligent consultation system to better adapt to the constantly changing policy environment by continuously updating the knowledge base, ensuring the timeliness and accuracy of the information provided.
- Original External Knowledge base: In the domain-specific consultation system of the e-government, the main sources of external knowledge include regional public-government interaction modules, and government advisory service platforms. Consultations posed on these platforms are responded by government personnel based on their personal experience and relevant internal knowledge. We define these consultation-response pairs as external knowledge. Compared to internal knowledge, this type of data is much larger in scale and is crucial to ensuring that the consultation system can cover a wide range of flexible consultations.

TABLE 1. Overview of the internal and external knowledge base.

Knowledge Base	Description	Data Source	Data Volume
Internal Knowledge	Foundational government information for accurate and authoritative responses.	Government official websites and portals.	700k
External Knowledge	Information obtained from official online consulting platforms.	Government official websites and public service platforms.	3M

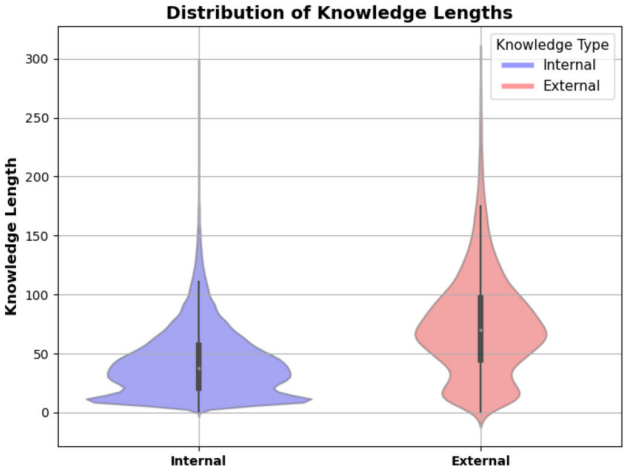

FIGURE 7. The distributions of knowledge lengths.

The specific database construction details are presented in Table 1.

In the construction process of the consultation system in this paper, it is necessary to integrate and fuse internal and external knowledge. To illustrate the necessity and difficulty of this task, we compared the distribution differences between the two in terms of text length and sentiment. As shown in Figure 7, the length distribution of internal knowledge is mainly concentrated in a shorter range, with most sentences not exceeding 50 characters, reflecting the typically concise and direct mode of expression adopted in internal knowledge. In contrast, the length distribution of external knowledge shows a broader range, with many sentences exceeding 100 characters, indicating the detailed and complex nature of expression in external knowledge. This comparison reveals a tendency towards generalization in internal knowledge expression, whereas external knowledge displays a higher degree of informational richness.

On the emotional dimension,[1] as revealed by Figure 8, the sentiment tendency of IK is primarily neutral, likely reflecting the objective requirements of professional or official texts. Conversely, EK exhibits a pronounced distribution of both positive and negative sentiments, which may be

[1]The model used here can be find at https://huggingface.co/techthiyanes/chinese_sentiment

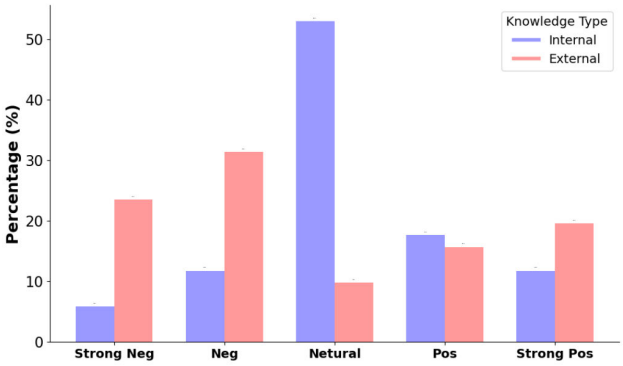

FIGURE 8. The distributions of knowledge sentiment.

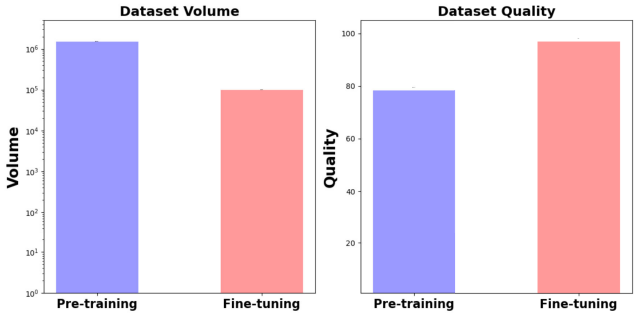

FIGURE 9. Results of data annotation.

attributed to the subjectivity and emotional expressiveness of users within the context of the Internet. This distinction in sentiment distribution lies in the fact that IK content leans towards statements of fact and data, whereas EK more frequently mirrors personal opinions and emotional responses.

The disparities in length and sentiment distribution not only reveal the challenges inherent in the process of knowledge base integration, but also highlight the necessary adaptations and adjustments for achieving effective amalgamation of knowledge. By meticulously analyzing and understanding these differences, we can better guide the design of knowledge integration strategies, thereby enhancing the quality and applicability of the combined knowledge base.

2) DATA ANNOTATION

To support the training of the model proposed in this paper, we carried out a series of data annotation tasks. To minimize the cost of annotation, we initially divided the data into two parts: pre-training data annotated solely by Large Language Models (LLMs) and fine-tune data that combines LLM[2] and human assistance. This section will specifically introduce our annotation methods and tasks.

We adopted the CoAnnotating method proposed by Kan et al. [20] This method introduces a new paradigm, namely, human-LLMs co-annotation of large-scale unstructured texts. CoAnnotating leverages the zero-shot capability of LLMs in text annotation tasks, using them as an effective complement to human annotation. This approach not only reduces annotation costs but also improves annotation efficiency. It has been demonstrated in practice to achieve significant performance improvements across various annotation tasks. [20]

- **Data annotation of knowledge alignment:** We first undertake the task of knowledge alignment, specifically aiming to align the consultation portions of internal and external knowledge. This task is fundamentally designed to enable the retrieval of internal knowledge in later consultations. We build indexes on two dimensions using ElasticSearch: textual (BM25 [21]) and semantic (sentence embedding). Through these indexes, we construct initial data pairs of internal knowledge and external knowledge (consultation part). After this initial setup, we proceed with detailed annotation to generate pre-training and fine-tune data. The specific process is outlined in Algorithm 2, and the annotation results are shown in Figure 9.
- **Data annotation of knowledge refinement:** During our experiments, we identified two characteristics of internal and external knowledge that affected the final effects of knowledge fusion. One issue common to both internal and external knowledge is that the collected data often contain multiple semantic angles within a complete sentence structure. For example, a single sentence about government responsibilities in internal knowledge may contain multiple clauses, each representing a different responsibility. Similarly, a complaint in external knowledge may consist of different clauses representing various demands. Additionally, there is a unique problem with external knowledge: due to the consultation side, it includes a large amount of colloquial information irrelevant to the knowledge.
  In response to these issues, we refined the annotation of both internal and external knowledge by breaking down the sentence components and removing irrelevant content from both types of knowledge. Specific results are presented in Figure 10.

B. DOMAIN-SPECIFIC CONSULTATION EXPERIMENTS

1) KNOWLEDGE FUSION

Our core goal is to build a lightweight, flexible domain-specific consultation system. We base our consultation responses on the constructed internal and external knowledge bases. First, we achieve the fusion of internal knowledge with external knowledge, and then we implement the system based on the inter-external fusion knowledge base. This section presents the experimental results of the task of knowledge fusion.

[2]The LLM used is Llama2-Chinese-7b-Chat, available at https://github.com/LlamaFamily/Llama-Chinese

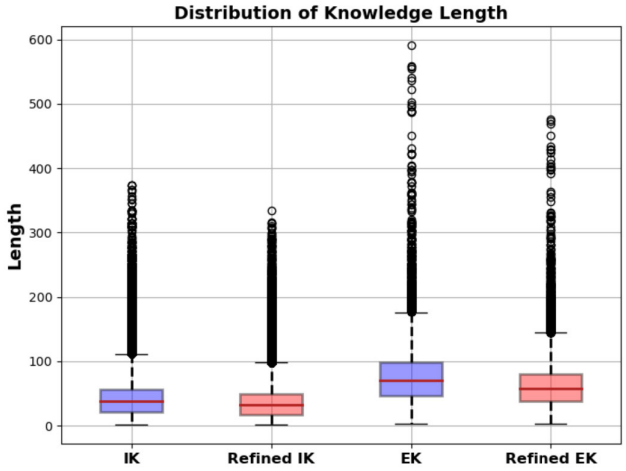

FIGURE 10. Result of knowledge refinement, After knowledge refinement, the lengths of both the internal and external knowledge have become shorter.

- **Baseline:** We compare several sentence representation methods on the Knowledge fusion task, which include GloVe embeddings [22], BERT [23], BERT-flow [24], Bert-Whitening [25], DiffCSE [26], SimCSE [27], DCLR [28], DenoSent [29].
- **Dataset:** Our core training data consists of original internal knowledge (IK), refined internal knowledge (RIK), original external knowledge (REK) and refined external knowledge (REK), with 1.5 million instances for pre-training and 10,000 for fine-tuning. The training objectives are twofold: to achieve precise matching between consultation and internal knowledge, as well as between consultation and external knowledge (consultation side).
- **Evaluation:** During the evaluation, we annotate 2,000 instances each of consultation and the four types of knowledge matches, similar to the STS dataset [30]. Each instance is scored from weak to strong match, divided into six levels (0, 1, 2, 3, 4, 5). We use the Spearman's rank correlation coefficient to assess the final performance of the model.
- **Implement details:** We use the IEK-Fusion technique proposed in Chapter III to achieve the fusion of four types of knowledge: IK, RIK, EK, REK. During the testing phase, we assess the match between the consultation and these four types of knowledge, corresponding to C2I, C2I-R, C2E, C2E-R respectively. As the baseline models we selected are semantic representation models, we first evaluate the performance of the native baseline models, here we also list the performance of the baseline models on the traditional semantic textual similarity dataset (STS) to highlight the differences between our proposed task and the standard semantic similarity task), the performance of the baseline models further trained on our dataset, and the performance of the IEK-Fusion model. During the evaluation of the IEK-Fusion model, we conduct ablation experiments to assess the impact of our pre-training data, finetune data, and the gate mechanism in the knowledge encoder.
- **Main Results:** As shown in Table 2, we first compared the performance of various baseline sentence representation models on four different knowledge fusion tasks, covering four types of knowledge: original internal knowledge (C2I), refined internal knowledge (C2RI), original external knowledge (C2E), and refined external knowledge (C2RE). We also listed the average metrics for these models on the general semantic similarity STS dataset (STS Avg).
  From Table 2, we can draw the following conclusions: the refinement operation can enhance model performance, with an overall average improvement of 1.72%. The difficulty of matching consultation with internal knowledge is noticeably higher, with final metrics being 31% lower than those achieved with external knowledge.

Algorithm 1 Data Annotation of Knowledge Alignment

Input: Internal Knowledge (IK), External Knowledge (EK)
Output: Annotated Data for Model Training
Data Preparation and Index Construction:
Collect Internal Knowledge (IK) and External Knowledge (EK)
Use Elasticsearch to build indexes:
- Configure index using BM25
- Configure index using semantic vector similarity
Data Recall:
for each IK in Internal Knowledge **do**
Retrieve the top 10 EK matches from the index based on BM25 and semantic embedding similarity
end for
Data Streamlining:
Categorize knowledge pairs into:
- Massive data for pre-training
- Small-scale data for fine-tuning
Data Annotation and Model Application:
for each pair in constructed IK-EK (Internal-External knowledge) pairs **do**
Retrieve data based on two BM25 indexes and semantic embedding index.
end for
for each annotated pair in pre-training data **do**
Apply Uncertainty-Guided Data Annotation Method for fine-tuning
end for
Output and Further Processing:
Integrate annotated data
Train models using integrated data
Evaluate model performance and optimize

TABLE 2. Performance of baseline models, Without training with domain-specific dataset. We report Spearman's correlation on all tasks. Results of STS taken from [29], Here, C2I represents Consultation to Internal knowledge, C2RI represents Consultation to Refined Internal Knowledge, C2E represents Consultation to External Knowledge, and C2RE represents Consultation to Refined External Knowledge.

Models	C2I	C2RI	C2E	C2RE	C2I Avg.	C2E Avg.	STS Avg.
GloVe embeddings	30.52	32.76	56.35	58.57	31.64	57.46	61.32
BERT	20.63	21.49	49.49	51.72	21.06	50.605	56.7
BERT-flow	35.32	36.82	60.25	64.83	36.07	62.54	66.55
BERT-whitening	34.68	35.06	59.32	62.78	34.87	61.05	66.28
DiffCSE	36.94	37.6	70.18	71.5	37.27	70.84	78.49
SimCSE	35.72	36.8	71.45	72.06	36.26	71.75	76.25
DCLR	36.76	37.03	70.54	73.87	36.89	72.20	77.22
DenoSent	36.65	37.94	72.78	74.34	37.29	73.56	79.33

TABLE 3. Performance of all models, with continue training with domain-specific dataset. We report Spearman's correlation on all tasks, In this context, "pt" represents pre-training, and "ft" represents fine-tuning.

Models	C2I	C2I-R	C2E	C2E-R	Avg.
BERT	72.35	73.21	69.42	70.14	71.28
BERT-flow	76.37	80.23	72.85	74.68	76.03
BERT-whitening	77.25	81.75	73.81	74.62	76.85
DiffCSE	79.5	84	72.56	73.06	77.28
SimCSE	81.75	86.25	74.39	75.83	79.55
DCLR	83	87.5	76.8	78.02	81.33
DenoSent	84.31	88.48	77.25	78.37	82.10
IEK-Fusion w/o pt	52.36	53.62	72.39	73.05	62.855
IEK-Fusion w/o ft	73.58	74.31	72.83	73.15	73.46
IEK-Fusion w/o gate	80.48	81.39	76.8	78.02	79.17
IEK-Fusion	**85.75**	**86.43**	**88.25**	**88.91**	**87.33**

However, both are still below the performance of the model on the general semantic similarity dataset STS, where the STS metrics exceed the average C2E metrics by 4.74%. This highlights the challenges of specialized domain-specific text matching.

Subsequently, in Table 3, we compared the performance of the baseline models trained on specialized datasets with the model proposed in this paper, once again comparing across four datasets.

The experimental results indicate that both pre-training and fine-tuning, as well as the configuration of gating mechanisms, have a significant positive impact on the outcomes.

2) CONSULTATION RESPONSE

For the Consultation Response task, we evaluate the accuracy of the system's internal and external knowledge in responding to user consultations across various real-world datasets.

- **Baseline:** For this task, we take the semantic encoding model used in the knowledge fusion task as the baseline model. After sentence embedding is performed by the semantic encoding model (including IEK-Fusion), indexes are built using Faiss [31].
- **Evaluation:** We prepare the Consultation Response task across six datasets, each dataset is directly related to a government department and contains 200 examples. We employ a detailed scoring system, rating responses from 0 to 5, where:
  - **Score 0:** The response is completely irrelevant or incorrect.
  - **Score 1:** The response contains significant errors or irrelevant information but includes minimal relevant details.
  - **Score 2:** The response is partially correct but lacks details or contains minor inaccuracies.
  - **Score 3:** The response is mostly correct but includes some incorrect or irrelevant information.
  - **Score 4:** The response is largely accurate and relevant with only minor errors.
  - **Score 5:** The response is completely accurate, fully relevant, and provides comprehensive information.
- **Implementation details:** For the Consultation Response task, we conducted evaluation experiments on datasets from six different real-world scenarios. We compared our model with baseline models and also assessed the performance of a LLM on this task.
- **Main Results:** The experimental results for the Consultation Response task are shown in Figure 11. From the results depicted in the figure, we can see that compared to other mainstream sentence representation models, IEK-Fusion has a clear advantage in the consultation response task across multiple scenarios, and its performance closely approaches that of the LLM in this task.

  Specifically, our results exceed those of the well-performing sentence representation model DenoSent by 7 percentage points, are more than 10 percentage points higher than the scores of general models, and are only 3.5 percentage points lower than the performance of the LLM.

3) ANALYSIS OF HANDLING CAPABILITY FOR COMPLEX CONSULTATIONS

The experiment focuses on testing the IEK-Fusion model's capability to handle complex queries, which involve multiple knowledge points or require deeper understanding. Specifically, it aims to assess the model's performance

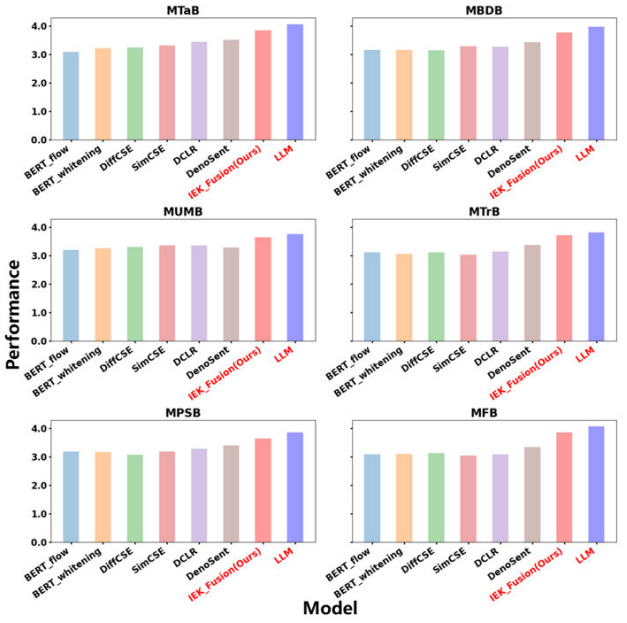

FIGURE 11. Result of the domain-specific consultation system, Among them, different datasets represent consultation scenarios directly related to different government departments. MTaB represents the Municipal Tax Bureau, MBDB represents the Municipal Big Data Bureau, MUMB represents the Municipal Urban Management Bureau, MTrB represents the Municipal Transportation Bureau, MPSB represents the Municipal Public Security Bureau, and MFB represents the Municipal Finance Bureau.

in comprehending and responding to intricate queries that encompass various aspects of knowledge. For this purpose, we curated a specialized dataset comprising a series of complex queries. Each query necessitates the model to utilize both internal knowledge and external Q&A data to furnish accurate responses. These queries are devised to real-world scenarios of user inquiries regarding government vertical livelihood services. They encompass diverse areas such as policy interpretation, and public service guidelines. Details can be found in Figure 12. It can be observed that the more complex the consultation, the more pronounced the advantage of our IEK-Fusion model.

4) ANALYSIS OF THE SIZE OF THE PRE-TRAINING DATASET

This section analyzes the impact of pre-training data at different scales on model performance. The model was pre-trained on large-scale roughly annotated data and fine-tuned on small-scale finely labeled data. Details can be found in Figure 13.

From the results in the figure, it can be seen that utilizing a large amount of slightly lower-quality pre-training data can effectively improve the performance of the final model.

5) COMPARISON OF IEK-FUSION WITH LLM

In this section, we compared the performance of our models against LLM based on our proposed IEK-Fusion technique. Our models, trained with 250M parameters building upon

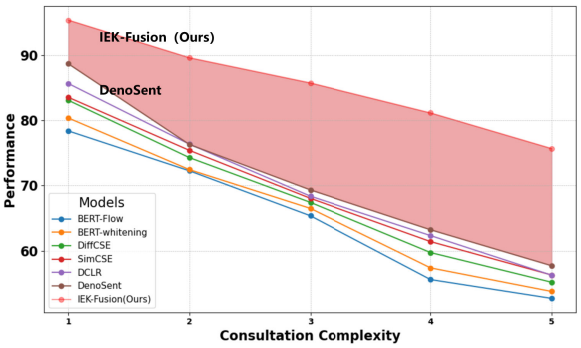

FIGURE 12. A comparison of different models' performance in handling complex consultations, with the horizontal axis representing the complexity of the consultation and the vertical axis representing model performance. The shadow in the legend highlights the advantage of our proposed model over the optimal baseline model.

FIGURE 13. Comparison of the Effects of Pre-training Data at Different Scales.

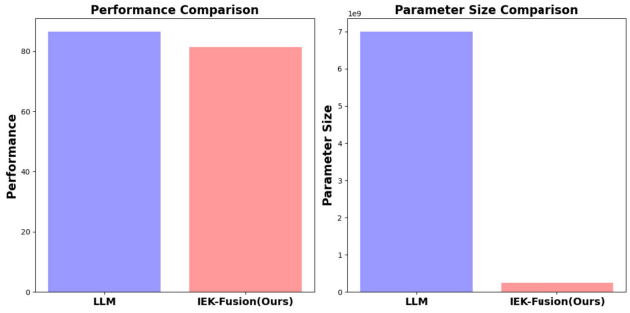

FIGURE 14. Comparison of IEK-Fusion with LLM.

the Roberta [32] model,[3] were contrasted with an LLM named llama [33],[4] which has 7B parameters, as illustrated in Figure 14. In tasks involving knowledge fusion and consultation response, our models achieved over 90% effectiveness compared to LLM, utilizing only 4% of its parameter size in both tasks.

[3]The model is trained on RoBERTa_zh_L12 models, available at https://github.com/brightmart/roberta_zh

[4]The LLM used is Llama2-Chinese-7b-Chat, available at https://github.com/LlamaFamily/Llama-Chinese

V. CONCLUSION

In conclusion, we propose a novel technique to constructing domain-specific consultation systems that effectively leverage Internal and External Knowledge Fusion Contrastive Learning. This methodological innovation addresses the significant challenges posed by Large Language Models (LLMs), including their substantial demand for computational resources, their propensity for generating hallucinations. This technique is designed for industrial specific domains such as e-government, intelligent diagnosis, and legal consultancy, our technique not only simplifies the integration of vast external knowledge from the Internet but also ensures the incorporation of precise, specialized knowledge within these industrial domains.

Our lightweight model utilizes merely 4% of the parameters of the LLM while maintaining over 90% of their performance, marks a significant advancement. This model's reduced computational resource requirements facilitate easier deployment and use, particularly in environments constrained by resources.

Our research contributes to industrial domains by offering a technically innovative solution that significantly enhances the performance of domain-specific consultation systems through the integration of internal and external knowledge sources. Furthermore, the development of a cost-effective, lightweight consultation framework opens new avenues for deploying advanced consulting systems in resource-limited environments.

In this study, although our model employs far fewer parameters than traditional Large Language Models (LLMs), we have successfully maintained high performance through efficient knowledge fusion and optimization techniques. This achievement demonstrates that, even with fewer parameters, our carefully designed model structure and algorithms can still deliver excellent system performance. However, we did encounter several challenges during the experiments. These challenges primarily stem from the significant differences between internal and external knowledge, which make accurate knowledge fusion difficult. These differences include variations in expression, length, and sentiment. We performed targeted refinements on both internal and external knowledge and achieved knowledge fusion through multi-view contrastive learning techniques. Another challenge is the fine-grained alignment of internal and external knowledge in complex scenarios. To address this issue, we designed a special gated hierarchical encoder and proposed the Fine-Grained Consultation-Knowledge Fusion algorithm to solve this problem.

The experiments conducted have demonstrated that our proposed universal framework is capable of effectively facilitating the integration of internal and external knowledge to build lightweight, domain-specific consultation systems. In the future, considering specialized industries where there is a severe lack of either internal or external knowledge, we plan to integrate few-shot learning methods into our universal framework to address the unique requirements of these specialized scenarios. Additionally, some industries have stringent requirements for internal data security management. In response, we plan to explore privacy-preserving data mining techniques to investigate knowledge fusion and model building while ensuring privacy.

REFERENCES

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[2] J. Han, J. Lu, Y. Xu, J. You, and B. Wu, "Intelligent practices of large language models in digital government services," *IEEE Access*, vol. 12, pp. 8633–8640, 2024.

[3] Y. Li et al., "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," 2023, *arXiv:2303.14070*.

[4] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan, "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2023, *arXiv:2306.16092*.

[5] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[6] J. Shen, N. Tenenholtz, J. Brian Hall, D. Alvarez-Melis, and N. Fusi, "Tag-LLM: Repurposing general-purpose LLMs for specialized domains," 2024, *arXiv:2402.05140*.

[7] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, vol. 33, 2020, pp. 1877–1901.

[8] J. Shen, N. Tenenholtz, J. B. Hall, D. Alvarez-Melis, and N. Fusi, "Mark my words: Repurposing LLMs for specialized domains via ability tokens," 2024. [Online]. Available: https://openreview.net/forum?id=GsNp4ob8BY

[9] I. Istiadi, E. B. Sulistiarini, R. Joegijantoro, and D. U. Effendy, "Infectious disease expert system using Dempster Shafers with recommendations for health services," *Jurnal Rekayasa Sistem dan Teknologi Informasi*, vol. 4, no. 1, pp. 17–27, Feb. 2020.

[10] Y. Huang, M. Chen, and K. Tang, "Training like playing: A reinforcement learning and knowledge graph-based framework for building automatic consultation system in medical field," 2021, *arXiv:2106.07502*.

[11] X. Liu and L.-Q. Xu, "Knowledge graph building from real-world multisource 'dirty' clinical electronic medical records for intelligent consultation applications," in *Proc. IEEE Int. Conf. Digit. Health (ICDH)*, Sep. 2021, pp. 260–265.

[12] W. Chen, Z. Li, H. Fang, Q. Yao, C. Zhong, J. Hao, Q. Zhang, X. Huang, J. Peng, and Z. Wei, "A benchmark for automatic medical consultation system: Frameworks, tasks and datasets," *Bioinformatics*, vol. 39, no. 1, Jan. 2023, Art. no. btac817.

[13] J. Tang, W. Xu, J. Li, Y. Tian, and S. Xu, "Multi-view learning methods with the LINEX loss for pattern classification," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107285.

[14] G.-Y. Zhang, Y.-R. Zhou, C.-D. Wang, D. Huang, and X.-Y. He, "Joint representation learning for multi-view subspace clustering," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 113913.

[15] H. Phan, H. Le Nguyen, O. Y. Chén, L. Pham, P. Koch, I. McLoughlin, and A. Mertins, "Multi-view audio and music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 611–615.

[16] X. Chen, F. Zhou, G. Trajcevski, and M. Bonsangue, "Multi-view learning with distinguishable feature fusion for rumor detection," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108085.

[17] J. Lu, M. Henchion, I. Bacher, and B. M. Namee, "A sentence-level hierarchical BERT model for document classification with limited labelled data," in *Proc. 24th Int. Conf. Discovery Sci.*, Halifax, NS, Canada. Cham, Switzerland: Springer, 2021, pp. 231–241.

[18] B. Shan, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-ViL 2.0: Multi-view contrastive learning for image-text pre-training," 2022, *arXiv:2209.15270*.

[19] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[20] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, and D. Yang, "CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation," 2023, *arXiv:2310.15638*.
[21] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.
[22] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[24] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," 2020, *arXiv:2011.05864*.
[25] J. Su, J. Cao, W. Liu, and Y. Ou, "Whitening sentence representations for better semantics and faster retrieval," 2021, *arXiv:2103.15316*.
[26] Y.-S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljačić, S.-W. Li, W.-T. Yih, Y. Kim, and J. Glass, "DiffCSE: Difference-based contrastive learning for sentence embeddings," 2022, *arXiv:2204.10298*.
[27] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.
[28] K. Zhou, B. Zhang, W. X. Zhao, and J.-R. Wen, "Debiased contrastive learning of unsupervised sentence representations," 2022, *arXiv:2205.00656*.
[29] X. Wang, J. He, P. Wang, Y. Zhou, T. Sun, and X. Qiu, "DenoSent: A denoising objective for self-supervised sentence representation learning," 2024, *arXiv:2401.13621*.
[30] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity–multilingual and cross-lingual focused evaluation," 2017, *arXiv:1708.00055*.
[31] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024, *arXiv:2401.08281*.
[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[33] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

JIABIN ZHENG is currently an Artificial Intelligence Algorithm Engineer with the School of Computer Science, Peking University. His research interests include information retrieval and natural language processing.

HANLIN WANG is currently an Intern with the Institute of Social Science Survey, Peking University. His research interests include data processing and analysis.

JIAHUI YAO is currently an Engineer with the Institute of Social Science Survey, Peking University. Her research interests include big data technology and machine learning.