Semantically-Guided Two-Stage Classification for Scientific Texts: Integrating Structural Awareness and Expert Routing

Meng Wang[1], Jing Xie[1], Yang Li[1,2], Zhixiong Zhang[1,2] and Hanyu Li[1,2‡]

[1]*National Science Library, Chinese Academy of Sciences, Beijing 100190, China*

[2] *School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China*

{wangmeng2022, xiej, liyang2022, zhangzx, lihy}@mail.las.ac.cn

Keywords: Text Classification; Two Stage Classfication; Semantic Feature Encoding; Mixture of Experts; Imbalanced Data Learning

Abstract

As the pace of scientific knowledge production accelerates and the volume of domain-specific literature continues to grow, the development of classification models capable of capturing semantic subtlety and adapting to imbalanced label distributions has become increasingly central to intelligent text understanding. However, existing multi-label classification approaches for scientific texts often fall short in fully leveraging semantic features and struggle with accurately identifying low-resource or semantically overlapping categories. To address these challenges, this study proposes a two-stage classification framework that integrates domain-aware semantic feature encoding into large language models (LLMs). In the first stage, we construct a dynamic-window-based topic semantic extractor and a hierarchical structure-aware encoder, embedding these features into the [CLS] and [SEP] positions of the LLM to enhance semantic representation. In the second stage, a fine-grained semantic routing strategy is introduced within a mixture-of-experts (MoE) architecture, enabling adaptive expert allocation informed by semantic cues. Experimental results on both the DBpedia benchmark and a domain-specific dataset of scientific value sentences demonstrate consistent performance improvements: our approach achieves a 4.37%–5.82% gain in F1 score over existing semantic encoding methods, reaches a peak F1 of 94.19% on value sentence recognition, improves macro-average F1 to 93.35% on balanced data (a 9.04% increase), and boosts F1 for minority classes by over 25% on imbalanced datasets. The results suggest that semantically guided representation enhancement may offer a promising pathway toward more accurate and resilient classification in scientific text mining.

‡Corresponding author: Hanyu Li (Email: lihy@mail.las.ac.cn; ORCID: 0000-0003-1426-3242).

1. Introduction

Text multi-classification is a fundamental task in the field of Natural Language Processing (NLP), aiming to accurately assign textual data to predefined categories based on semantic content and linguistic features [1]. This task plays a crucial role in parsing the semantic structures of scientific texts, extracting key information, and organizing knowledge [2]. As a result, optimizing and improving multi-classification models has remained a central focus in NLP research [3]. Semantic features, which encapsulate the core content and deep meaning of texts, are pivotal to enhancing the accuracy and efficiency of multi-classification models [4]. However, with the increasing scale of scientific textual data and the growing complexity of classification taxonomies, the semantic boundaries between different categories have become increasingly blurred. Moreover, class distribution is often significantly imbalanced, making it difficult for models to capture subtle semantic differences between classes. This is particularly problematic for minority classes, where recognition accuracy tends to be much lower. Therefore, embedding semantic features of scientific texts into multi-classification models is of critical importance.

Existing semantic-aware multi-classification models are primarily based on Pre-trained Language Models (PLMs), which leverage embedded representations of key textual tokens. PLMs are trained on large-scale corpora to learn distributed representations of salient words in context, and utilize attention mechanisms to compute their associations with surrounding tokens to capture semantic features of the text [5,6]. While this token-level representation learning approach improves model performance on multi-class classification tasks, it often treats feature tokens as isolated semantic units, neglecting the semantic dependencies among them [7,8]. To improve the model's semantic comprehension, researchers have explored various semantic fusion mechanisms—such as attention modules over feature tokens and graph-based semantic dependency modeling—to integrate richer semantic information [9,10]. However, these existing strategies often adopt uniform processing techniques and fail to account for the varying contributions of different textual features across classification tasks, which may hinder the model's ability to identify fine-grained semantic distinctions between classes.

With the rapid advancement of Large Language Models (LLMs) in the field of NLP, the strong capabilities of LLMs in semantic understanding and knowledge representation offer new avenues for addressing the challenge of semantic feature utilization in text multi-classification tasks [11,12]. Through pretraining on massive textual corpora, LLMs can effectively capture both topical semantic features and structural semantic information, thereby laying the foundation for more comprehensive semantic representations. Additionally, LLMs possess fine-grained semantic discrimination capabilities, enabling them to recognize the unique semantic characteristics of different text categories—a promising direction for mitigating class imbalance in multi-class scenarios [13]. Nevertheless, a critical challenge remains: how to effectively integrate the semantic strengths of LLMs while achieving accurate identification of underrepresented or minority classes.

To address these challenges, this paper proposes a two-stage classification method with embedded semantic feature encoding. In the first stage, we leverage the semantic comprehension capabilities of LLMs to construct a dual semantic encoding mechanism that embeds both topical and structural semantic features into the model, thereby enhancing its holistic understanding of the text. In the second stage, we incorporate a fine-grained semantic routing strategy within a Mixture-of-Experts (MoE) framework. This mechanism enables the model to route inputs based on class-specific semantic cues, allowing expert networks to focus on subtle distinctions between categories. As a result, the proposed approach significantly improves the identification accuracy of underrepresented classes in imbalanced datasets. For instance, when distinguishing between semantically similar sentence categories, such as background and motivation statements, the proposed method leverages dual semantic encoding (topical and structural) to enhance contextual understanding in the first stage. This helps disambiguate semantically overlapping texts. In the second stage, the semantic-aware routing mechanism further amplifies category-specific signals by directing sentences to specialized expert networks. This hierarchical setup not only alleviates the ambiguity of semantic boundaries but also improves the model's ability to recognize minority classes with limited samples.

In response to the limitations observed in scientific text multi-classification tasks, and informed by the related literature, we identify three key challenges currently facing the field: (1) Insufficient

utilization of semantic features: Existing LLM-based classification methods primarily focus on output-level representations while overlooking the rich internal semantic information within the model and its connection to text understanding. There is a need to develop encoding mechanisms that can fully exploit these internal semantics to enhance the model's comprehension of textual meaning. (2) Imbalanced category feature representation: Most current approaches to class imbalance emphasize data-level augmentation, lacking explicit modeling of category-specific semantic characteristics. It is necessary to design routing mechanisms capable of accurately perceiving the unique semantic features of each class and guiding expert models to perform discriminative recognition accordingly. (3) Overly simplistic classification architectures: Existing multi-classification frameworks often rely on a single-stage architecture, which struggles to achieve both general understanding and fine-grained categorization. A hierarchical architecture that integrates broad semantic comprehension with nuanced feature differentiation is needed to improve overall classification performance.

Compared with existing scientific text multi-classification methods, the proposed two-stage classification approach with enhanced semantic feature encoding introduces three key innovations:(1) A dual-semantic feature encoding mechanism is designed to extract both topical and structural semantic features and directly embed them into the internal representations of LLMs. This enhances the model's global semantic awareness and improves the efficiency of semantic feature utilization. (2) A fine-grained semantic routing framework is proposed within a Mixture-of-Experts (MoE) architecture. By embedding category-specific semantic features into the MoE routing layer, the model can perform semantic-aware expert assignment, thereby improving classification accuracy—particularly under imbalanced category distributions. (3) A "general-to-specific" two-stage classification architecture is constructed. In the first stage, semantically enhanced LLMs perform coarse-grained filtering and preliminary classification. In the second stage, fine-grained semantic routing directs expert model selection, forming a progressive classification pipeline from general recognition to precise differentiation.

2. Related Work

2.1 Text Classification

Research on text classification has primarily focused on two paradigms: classification based on Pre-trained Language Models (PLMs) and classification based on Large Language Models (LLMs). Within the PLM-based framework, efforts have concentrated on fine-tuning models such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) [14] to leverage their semantic representation capabilities learned from large-scale corpora. Li et al. [15] proposed a BERT-CNN-based model for identifying research question sentences, where BERT was employed as a word embedding generator and CNN was used to extract hierarchical linguistic features from sentence vectors, thereby enabling effective sentence classification. The experimental results demonstrated that BERT's representational power significantly improved classification accuracy. In 2019, Liu et al. [14] introduced RoBERTa, an enhanced version of BERT. Compared to its predecessor, RoBERTa was pretrained on a larger corpus with extended training time to better capture linguistic subtleties and complexity. Additionally, RoBERTa removed the Next Sentence Prediction (NSP) objective, thereby focusing learning on more accurate contextual representations, and adopted a dynamic masking strategy, which helped the model generalize across broader language patterns. Liu and Cao [16] applied RoBERTa, BERT, and BiLSTM-CRF (Bidirectional Long Short-Term Memory with Conditional Random Fields) to perform named entity recognition in Winter Olympics news articles. Their results indicated that RoBERTa outperformed BERT and BiLSTM-CRF by 0.77% and 1.81% in F1-score, respectively. Further, Liu et al. [17] proposed a causal intervention-based approach using RoBERTa to mitigate the influence of confounding factors in few-shot relation classification tasks. Their results showed that RoBERTa significantly enhanced semantic representation and feature extraction capabilities, achieving a classification accuracy of 93.38% on the FewRel dataset. This highlights RoBERTa's superior performance in scientific text classification tasks. Nevertheless, PLMs fundamentally rely on masked language modeling and sentence-level pretraining tasks, which limits their ability to capture fine-grained semantic features and complex contextual dependencies in nuanced classification scenarios.

Compared with PLMs, Large Language Models (LLMs) possess stronger semantic understanding and knowledge transfer capabilities, offering new directions for optimizing text classification tasks. Research on LLM-based text classification can be broadly divided into two categories. (1) Instruction-tuned classification approaches enhance model comprehension of classification tasks by designing tailored prompt templates. For example, Han [18] proposed a fine-tuning strategy based on the Qwen-7B model for financial text classification. Experimental results showed that the fine-tuned LLM achieved an accuracy of 82.27%, surpassing both the DeBERTa-V3-base and DeBERTa-V3-large models. This indicates that fine-tuning LLMs can effectively improve classification performance. Zhang et al. [19] proposed an adaptive enhancement framework for text classification by adjusting the distribution of training samples and iteratively fine-tuning LLMs, ultimately forming a specialized classification model. Benchmark evaluations revealed that this adaptive LLM model outperformed PLMs by 1.36% in accuracy. Chae and Davidson [20] explored the use of LLMs in supervised text classification, comparing prompt-based zero-shot and few-shot learning with fine-tuning using larger annotated datasets, as well as instruction tuning that combines prompts with task-specific training data. Their findings demonstrated that instruction tuning is particularly effective in improving LLM performance on complex classification tasks. Fatemi et al. [21] applied a model merging technique by integrating single-task, domain-specific fine-tuned models with a base LLM to perform financial domain classification. Experimental results showed that the merged LLM approach significantly improved zero-shot classification performance, highlighting its effectiveness in low-resource scenarios.

However, instruction-tuned LLMs for text classification tend to learn dominant class features due to data imbalance, resulting in limited ability to recognize minority classes. This limitation is exacerbated when different categories exhibit subtle differences in semantic expression, which simple prompt templates often fail to capture effectively. Consequently, the classification performance of LLMs is often constrained when applied to imbalanced datasets. To address this issue, (2) data augmentation-based approaches have been proposed to enhance the model's ability to learn category-specific features. Peng and Shao [22] developed a data augmentation framework targeting class imbalance in text classification. Their method adjusts the mapping from classification labels to instruction prompts and uses GPT-4 to generate synthetic data. Experimental results demonstrate that instruction tuning on the augmented dataset improves classification accuracy. Meguellati et al. [23] employed LLMs to clean noisy text and provide context-rich explanations, thereby enriching the training set without significantly increasing the overall data volume. Results showed that while zero-shot enhanced LLMs underperformed compared to supervised models, the integration of LLM-driven semantic augmentation enabled performance comparable to that of human-annotated datasets. Guo et al. [24] proposed a method that treats LLMs as data annotators to expand limited training data. The augmented data was then used to fine-tune both PLMs and LLMs. Results revealed that RoBERTa trained on GPT-4-generated data achieved performance equal to or better than models trained solely on human-labeled data.

2.2 Mixture-of-Experts Mechanism

The Mixture-of-Experts (MoE) model was first introduced by Jacobs et al. [25] in 1991 as a "divide-and-conquer" strategy for solving complex classification problems. As noted by Peralta et al. [26], MoE improves model performance by partitioning it into smaller, specialized sub-models, making it well-suited for handling high-complexity classification tasks. In recent years, MoE has demonstrated considerable advantages in the field of text classification. Le et al. [27] proposed an improved MoE Transformer model tailored for small-scale clinical text classification. While maintaining classification accuracy, the model significantly reduced computational resource consumption. Validated on a French clinical text dataset, it achieved 87% accuracy and an F1 score of 86%. Although slightly inferior to a biomedical pre-trained BERT model in terms of accuracy, the training speed was improved by a factor of 190—offering a highly efficient and feasible solution in resource-constrained clinical environments. Chen et al. [28] addressed data conflicts during mixed-instruction fine-tuning of multimodal LLMs by proposing a sparse LoRA-based expert mixture model, LLaVA-MoLE. This model introduces a set of LoRA expert modules within the Transformer layers and uses a routing function to assign tokens from different domains to the most suitable expert. This design enables adaptive learning across heterogeneous domains. Experimental results indicate that, compared with conventional LoRA

approaches, LLaVA-MoLE mitigates performance degradation caused by mixed-instruction datasets while maintaining similar computational cost—and even surpasses baseline models trained on double the data volume. To further enhance MoE performance, Chowdhury et al. [29] proposed an expert pruning method during fine-tuning. Wu et al. [30] tackled the uncertainty in expert routing by introducing GW-MoE, a fine-tuning approach based on the Global Workspace Theory. This method broadcasts uncertain tokens to multiple experts during fine-tuning, allowing them to benefit from the knowledge of various experts and reducing sensitivity to routing choices. Experimental results show that GW-MoE achieves consistent performance improvements on tasks such as text classification and question answering without increasing inference cost. Lin et al. [31] introduced a novel modality-aware MoE architecture, MoMa, which separates expert modules into modality-specific groups for handling visual and textual sequences. Experimental results indicate that, compared with a dense baseline of equivalent computational cost, MoMa achieves a 270% increase in overall computational efficiency, including a 160% improvement in text processing efficiency. These gains outperform the approximately 200% efficiency improvement of the standard MoE architecture.

3. Two-Stage Text Multi-Classification Method with Embedded Semantic Feature Encoding

Conventional approaches to scientific text multi-classification often adopt an end-to-end framework, where PLMs are directly fine-tuned for multi-class tasks based on general-purpose semantic understanding. However, this strategy may overlook semantic overlap between categories, thereby reducing classification accuracy—particularly in scenarios where class distributions are highly imbalanced. In such cases, models tend to favor majority classes with more abundant samples, leading to performance degradation.

To address this issue, we propose a two-stage classification method that integrates "general semantic comprehension" with "fine-grained semantic differentiation." In the first stage, we extract topical and structural semantic features from scientific texts and embed them into an LLM to perform a binary classification task. This step not only filters high-quality data for downstream fine-grained classification but also benefits from the relative simplicity of binary classification, which allows for easier acquisition of sufficient training samples. In the second stage, we focus on enhancing category discrimination by designing a fine-grained semantic feature extraction mechanism that emphasizes subtle inter-class distinctions—thereby improving recognition accuracy for minority classes.

3.1 Overall Framework

As illustrated in Figure 1, the overall framework of the proposed two-stage text multi-classification method with embedded semantic feature encoding consists of two core modules: a binary classification stage using LLMs with general semantic feature embedding, and a multi-class classification stage leveraging a Mixture-of-Experts (MoE) mechanism with fine-grained semantic routing. In the first stage, we construct a semantic feature extraction module that captures both the topical and structural semantic features of scientific texts. These features are embedded into the encoder layers of the LLM, allowing them to be integrated with the contextual information. This enhances the semantic representation of the input and enables the LLM to perform binary classification at the sentence level. In the second stage, based on the sentence set filtered from the first stage (e.g., value-bearing sentences in scientific texts), we combine fine-grained semantic features of various sentence types with an MoE routing mechanism. A set of expert feed-forward networks (FFNs) is used to carry out precise classification of different types of value sentences.

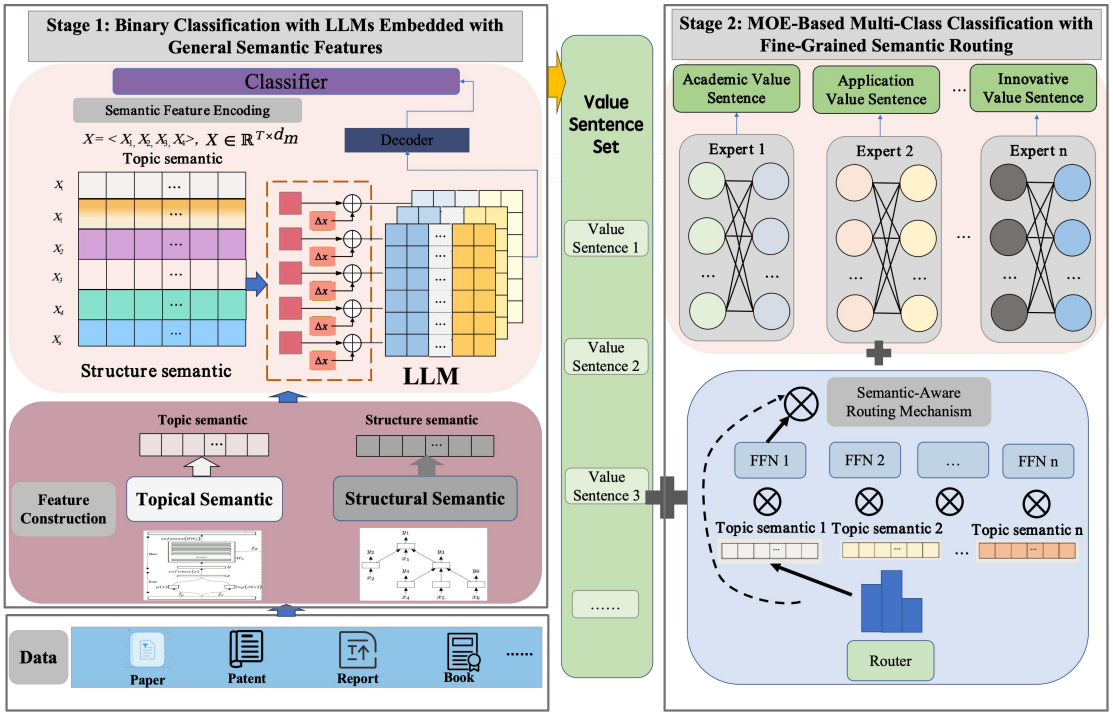

Fig.1 The Framework of Two-Stage Classification Method with Embedded Semantic Feature Encoding

3.2 Stage 1: Binary Classification with LLMs Embedded with General Semantic Features

Taking the multi-classification of value-bearing sentences in scientific texts as an example, this work first constructs a unified semantic representation that integrates topical and structural semantic features to enable efficient identification of such sentences. Specifically, a dynamic window-based local-global extraction method is used to obtain topical semantic features, while a hierarchical structural feature extraction mechanism is employed to capture inter-sentence logical relations. These two types of semantic features are then embedded into a large language model via a semantic encoder, forming enhanced semantic representations. The representations are subsequently processed by the decoder of the LLM, and the output is fed into a classifier to perform binary classification of value sentences. The resulting classification yields a high-quality subset of value-bearing sentences, which serves as the foundation for the fine-grained classification in the second stage.

(1) Topical Semantic Feature Construction

Topical semantic features are core components of text classification, as they effectively capture the main subject and semantic content of the text, thereby improving classification performance. Traditional topic extraction methods, such as TF-IDF and Latent Dirichlet Allocation (LDA), typically require global computations over the entire document and iterative optimization, which are computationally expensive and struggle to capture fine-grained semantic associations. Moreover, these methods rely solely on global statistical information, neglecting local contextual dependencies, which limits their representational power. To address these limitations, we propose a local-global topic feature extraction approach based on dynamic windows. By fusing local and global features, this method balances efficiency in semantic feature construction with improved expressive capability in downstream models. The detailed process of topical semantic feature extraction is illustrated in Figure 2.

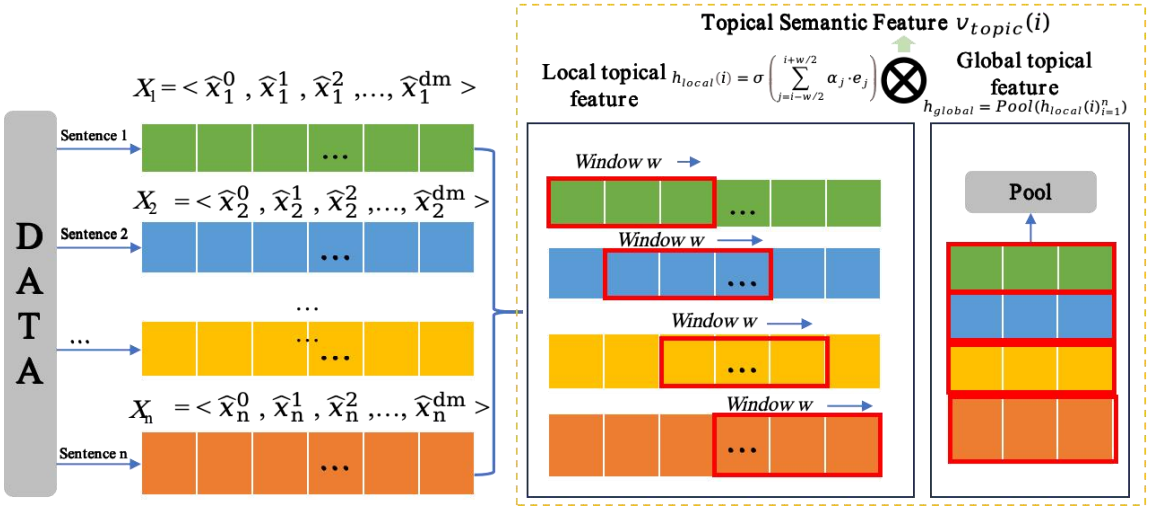

Fig.2 Topical Semantic Feature Extraction

Given an input text sequence, the construction of topical semantic features proceeds as follows:

1) Local Topic Feature Extraction

To effectively capture fine-grained semantic associations, a sliding window ω is applied over the input text to compute an aggregated topical representation for each word within the window:

$$h_{local}(i) = \sigma\left(\sum_{j=i-w/2}^{i+w/2} \alpha_j \cdot e_j\right) \tag{1}$$

where σ(·) denotes the activation function, w is the window size, e_j represents the vector representation of the j-th token, and α_j is the corresponding attention weight. The attention weights are computed as:

$$\alpha_j = softmax\left(e_j \cdot W_q \cdot e_i\right) \tag{2}$$

In this formulation, W_q is a learnable query parameter, and e_i denotes the vector of the target (center) token. The softmax function normalizes the attention scores across the windowed context.

2) Global Topical Relevance Computation

To compensate for the limitations of using only local context, a global topical representation is constructed and combined with local features for interactive enhancement:

$$h_{global} = Pool\left(h_{local}(i)_{i=1}^{n}\right) \tag{3}$$

Here, Pool(·) denotes a pooling operation that aggregates all local features to obtain a global representation. Accordingly, the final semantic feature $v_{topic}(i)$ representation in this work is defined as follows:

$$v_{topic}(i) = tanh\left(W_t \cdot \left[h_{local}(i); h_{global}\right] + b_t\right) \tag{4}$$

In this formulation, W_t and b_t represent the transformation matrix and bias term, respectively; [;] denotes the feature concatenation operation; and tanh is the hyperbolic tangent activation function.

(2) Structural Semantic Feature Construction

Given that different types of sentences exhibit diverse organizational structures, structural semantic features play a crucial role in understanding the logical relationships between sentences in scientific texts. Traditional methods, which often rely on bag-of-words models or simple sequential architectures, struggle to accurately capture hierarchical inter-sentence structural relations. To address this issue, we design a lightweight structural feature extraction approach that leverages hierarchical feature aggregation to derive structural semantic information from the text.

1) Basic Feature Representation: For a given input text, the initial representation of each token is obtained as follows:

$$h_{init}(i) = FFN([e_i; pos_i]) \tag{5}$$

Here, e_i denotes the token embedding, pos_i represents the positional encoding, *FFN* refers to a feed-forward neural network, and [;] indicates the concatenation operation between the two feature vectors.

2) Hierarchical Structure Aggregation: Based on dependency relations between tokens, hierarchical feature aggregation is performed as follows:

$$h_l(i) = \sigma\left(\sum_{j \in N(i)} w_j \cdot h_{l-1}(j)\right) \tag{6}$$

Here, $\sigma(\cdot)$ denotes the activation function, $N(i)$ represents the set of neighboring tokens of token *i*, w_j is the aggregation weight, *l* denotes the layer index, and $h_{l-1}(j)$ is the feature representation of token j at the *l* - 1-th layer.

3) Inter-Sentence Relation Representation:

To model sentence-level structural relations, we construct a structural representation at the sentence level:

$$v_{struct} = \text{Pool}\left(\{h_L(i)\}_{i=1}^{n}\right) \tag{7}$$

Here, $\text{Pool}(\cdot)$ denotes the pooling operation, $h_L(i)$ represents the token representation at the final layer, and v_{struct} is the final structural feature representation.

It should be noted that dependency parsing may introduce errors when processing complex academic sentence structures. While our ablation studies (see Section 5.1, Table 4) demonstrate the effectiveness of structural features in the overall framework, further investigation into the robustness and error patterns of structural encoding remains an important area for future research.

(3) Semantic Feature Encoding and Embedding Mechanism

To effectively embed the extracted semantic features into the large language model, we propose a feature replacement encoding mechanism, as illustrated in Figure 3. Traditional approaches typically use the [CLS] and [SEP] tokens of PLMs for text representation, where [CLS] is used to capture global semantic representations and [SEP] is used for segmenting and marking sequence boundaries. Considering that the topical semantic feature v_{topic} also encodes global semantic information, and the structural semantic feature v_{struct} captures boundary and organizational structure information, these two types of features can functionally substitute the [CLS] and [SEP] tokens, respectively.

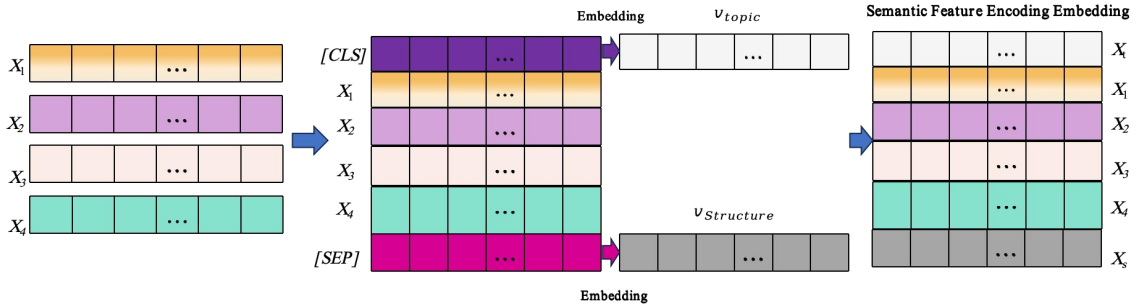

Fig.3 Feature replacement encoding mechanism

The proposed semantic feature encoding and embedding mechanism proceeds as follows:

First, the original input text sequence is passed through the topical and structural feature extraction modules to obtain the corresponding semantic representations. Then, the extracted topical semantic feature v_{topic} is used to replace the [CLS] token representation at the beginning of the sequence (represented by the purple block in the figure), and the structural semantic feature v_{struct} is used to replace the [SEP] token at the end of the sequence (represented by the pink block). The original token features in the middle of the sequence remain unchanged. The final embedded feature sequence X_{embed} is illustrated as follows:

$$X_{embed} = \left[v_{topic}, X_1, X_2, X_3, X_4, v_{struct}\right] \tag{8}$$

Through the proposed feature injection mechanism, the model is able to retain the original textual representation while effectively integrating domain-specific semantic information. This enhances the expressiveness of the features and provides more targeted representations for the binary classification of sentences.

3.3 Stage 2: MOE-Based Multi-Class Classification with Fine-Grained Semantic Routing

Taking the multi-class classification of valuable sentences in scientific texts as an example, we aim to achieve efficient and precise identification of different types of value-bearing sentences. Based on the filtered sentence set obtained in the first stage, we propose a semantic-aware routing framework built upon a Mixture of Experts (MoE) architecture. Specifically, fine-grained semantic feature extraction methods are designed for each subcategory, focusing on capturing the distinctive semantic representations specific to each type of value sentence, on top of the general topical semantics. These subcategory-specific semantic features are then embedded into the routing function of the MoE to construct a semantic-aware routing mechanism, which enables expert assignment based on semantic similarity. Finally, an expert selection strategy is devised to guide each expert model to focus on its designated category features, thereby improving recognition performance for underrepresented classes. The proposed semantic encoding–based MoE framework aims to alleviate the impact of class imbalance and achieve fine-grained classification of value sentences.

(1) Subcategory Semantic Feature Extraction

To capture the distinctive semantic expressions of various types of value sentences, we design a fine-grained subcategory-specific semantic feature extraction method based on the topical feature extraction described in Section 3.2.1. Unlike the first stage, which focuses on the overall semantics of the text, this stage emphasizes features that distinguish among different types of value sentences. Given an input sequence of value sentences $X = <X_1, X_2, ..., Xn>$ and their corresponding class labels Y, the subcategory semantic feature extraction process is as follows:

1) Class-Specific Feature Template Construction

To highlight the unique semantic patterns of each class, we first construct a class-specific feature template:

$$T_c = \text{Pool}\left(\{v_{\text{topic}}(x_i) \mid y_i = c\}\right) \tag{9}$$

Here, c denotes the value sentence class, and $v_{\text{topic}}(x_i)$ is the topical semantic feature obtained in the first stage.

2) Subcategory Semantic Feature Generation

Based on the class prototype features, the subcategory semantic feature of a sentence is computed as:

$$v_{sub}(i) = \sigma\left(W_c \cdot \left[v_{topic}(x_i); P_c\right]\right) \tag{10}$$

where W_c is the class-specific transformation matrix, σ(·) is the activation function, and $v_{sub}(i)$ is the semantic representation of the i-th sentence in subcategory c.

Through this approach, the subcategory semantic feature not only preserves the original topical information but also incorporates class-specific semantic patterns. This enables the extraction of more discriminative features tailored to each type of value sentence, thereby facilitating downstream semantic routing and expert selection.

(2) Semantic-Aware Routing Mechanism

Based on the extracted subcategory semantic features, we design a semantic-aware routing mechanism, as illustrated in Figure 4. In this mechanism, the topical semantic features are embedded into the routing function of the Mixture of Experts (MoE). A router dynamically allocates different Feed-Forward Network (FFN) experts, thereby enabling semantic-driven expert selection.

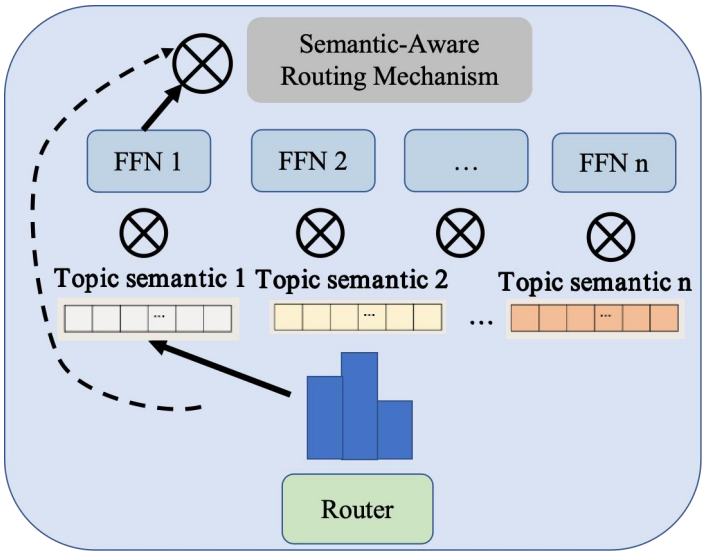

Fig.4 Semantic-aware routing mechanism

Specifically, for an input sample x, the semantic-aware routing process is as follows:

1) Routing Feature Construction

The topical semantic feature is first interacted with the routing parameters:

$$r(x) = W_r \bullet \text{Topic semantic} \tag{11}$$

where W_r is the routing weight matrix, and $r(x)$ denotes the routing feature.

2) Expert Selection Probability Calculation

Based on the routing feature, the matching probability between the sample and each FFN expert is computed as:

$$p(j/x) = softmax\left(r(x) \cdot h_j\right) \tag{12}$$

where h_j represents the semantic representation vector of the *j*-th FFN expert.

3) Dynamic Expert Assignment

According to the expert selection probabilities, the top-k most relevant FFN experts are selected for processing:

$$y = \sum \left(p(j/x) \cdot FFNj(x)\right) \tag{13}$$

where *k* is the pre-defined number of experts, and *y* is the final output feature.

(3) Expert Model Design

Based on the output of the semantic-aware routing mechanism, this section focuses on designing the expert selection strategy and the coordination mechanism. To accommodate the characteristics of different types of value sentences, we adopt heterogeneous expert architectures along with a coordinated training strategy.

1) Expert Architecture Design

Each expert model is implemented as a two-layer feed-forward network:

$$E_j(x) = FFN2_j\left(FFN1_j(x)\right) \tag{14}$$

where $FFN1_j$ and $FFN2_j$ denote the two-layer feed-forward networks of the *j*-th expert, with different parameter scales to accommodate tasks of varying complexity.

2) Expert Selection Strategy

Based on the routing probability $p(j|x)$, a dynamic thresholding method is adopted for expert selection:

$$S = \{j|p(j|x) > \tau_j\}$$
$$\tau_j = \beta \cdot avg\left(p(j|X)\right) \tag{15}$$

here, *S* denotes the selected expert set, τ_j is the dynamic threshold, β is a tunable coefficient, and $avg\left(p(j|X)\right)$ is the average selection probability of the expert.

3) Expert Collaboration Mechanism

The selected experts collaborate through weighted aggregation as follows:

$$y = \sum \left(w_j \cdot E_j(x)\right)/|S| \tag{16}$$

where w_j denotes the expert weight, which is computed as follows:

$$w_j = softmax\left(q_j \cdot k_x\right) \tag{17}$$

Here, q_j denotes the expert-specific query vector, and k_x represents the key vector of the input sample.

4. Experimental design

4.1 Experimental method

To validate the effectiveness of our proposed method, we designed a systematic experimental scheme. The experiments are divided into two stages: the first stage validates the improvement of embedded universal semantic features on binary classification performance, and the second stage validates the performance of the semantic routing-based MOE framework on multi-classification tasks.

The first stage includes three experiments: (1) To validate the effectiveness of the feature encoding mechanism, we conduct experiments on the public dataset DBPedia [32], comparing traditional [CLS] and [SEP] encoding approaches with our proposed embedding of topic semantic features and structural semantic features, thereby validating the advantages of the feature fusion strategy. (2) We perform feature ablation experiments on this dataset, separately validating the contributions of topic semantic features and structural semantic features to demonstrate the necessity and complementarity of these two feature types. (3) To validate the effectiveness of large language models with embedded universal semantic features for binary classification, we conduct model performance comparison experiments on our constructed scientific literature value sentence dataset to evaluate the performance advantages of our method.

In the second stage, to validate the multi-classification capability of our model, we design 2 experiments using our constructed 3-type fine-grained value sentence datasets (academic value sentences, application value sentences, and innovation value sentences): (1) To validate the model's classification capability under ideal data distribution, we use a balanced category value sentence dataset (with equal numbers of each type of value sentence) and compare pre-trained model fine-tuning methods with our proposed semantic routing MOE method to validate the multi-classification advantages of our approach. (2) To validate the model's capability in handling class imbalance problems, we use an imbalanced value sentence dataset (with different numbers of each type of value sentence) to demonstrate the effectiveness of the semantic routing mechanism for minority class sample recognition.

Regarding model selection, to comprehensively validate the universality and effectiveness of our method, we select BERT, SciBERT [33], and RoBERTa as baseline PLMs; for LLMs, we select Qwen3-14B [34], LLaMa4-17B [35], and GLM4-9B [36] as baseline models.

4.2 Datasets

To comprehensively evaluate the performance of our two-stage classification method enhanced with embedded semantic feature encoding for scientific texts, we select the public dataset DBPedia and our constructed scientific literature value sentence dataset as experimental datasets to validate the model's effectiveness at different stages. The specific dataset details are as follows:

(1) DBPedia Dataset. This dataset is a standard benchmark dataset for text classification tasks, sourced from Wikipedia article abstracts and containing 14 different thematic categories of text. The categories in the dataset cover 14 entity types including Company, Educational Institution, Artist, Athlete, etc. We select DBPedia as the benchmark dataset for evaluating text classification model performance. With its well-structured text and clear topics, it can effectively validate the performance of our proposed feature encoding embedding mechanism and semantic feature extraction methods on public datasets.

(2) Scientific Literature Value Sentence Dataset. This dataset is used to evaluate the performance of scientific text value sentence identification tasks, containing 23,912 scientific literature sentences composed of value sentences and non-value sentences with a positive-to-negative sample ratio of 1:1. The sentences in the dataset are sourced from academic papers in fields such as computer science and engineering technology, annotated by professional annotators to form a high-quality binary classification dataset. This dataset serves as a specialized dataset for evaluating value sentence recognition capability

and can be used to validate the practical application effectiveness of our method in the scientific text domain.

(3) Value Sentence Fine-grained Category Dataset. This dataset is used to evaluate the performance of value sentence multi-classification tasks, subdividing value sentences into three subcategories: academic value, application value, and innovation value. To validate model performance under different data distributions, we construct two experimental scenarios: balanced and imbalanced. The balanced dataset contains 1,200 samples per class, totaling 3,600 samples; the imbalanced dataset contains 11,570 academic value sentences, 1,297 application value sentences, and 1,430 innovation value sentences, totaling 14,297 samples. Through experiments on this dataset, we can validate the performance of our proposed semantic routing MOE method in handling class imbalance problems and compare it with classification effectiveness under balanced scenarios.

4.3 Experimental Setup

This dataset is a standard benchmark dataset for text classification tasks, sourced from Wikipedia article abstracts and containing 14 different thematic categories of text. The categories in the dataset cover 14 entity All experiments in this paper are implemented using PyCharm 2021.3.3 (Professional Edition) development tool and PyTorch deep learning framework. The main environment configuration is as follows: operating system Ubuntu 22.04, compilation environment Python 3.8.17, PyTorch=2.1.2; GPU acceleration environment CUDA 12.2, computer model DELL R740, CPU Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz, memory 1024GB, disk capacity 10TB, GPU A100 40G×2.

Our experiments adopt the commonly used accuracy (Accuracy, A), precision (Precision, P), recall (Recall, R), and F1-score (F1-Score, F1) as standard evaluation metrics in the self-attention mechanism research domain [37]. These metrics reflect the model's performance capability in sequence processing tasks from different perspectives and have been validated in numerous benchmark tests. The specific calculation formulas are as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{18}$$

$$Precision = \frac{TP}{TP+FP} \tag{19}$$

$$Recall = \frac{TP}{TP+FN} \tag{20}$$

$$F_1 = \frac{2\times Accuracy\times Precision}{Accuracy+Precision} \tag{21}$$

where TP (True Positive) represents correctly classified positive samples, i.e., a training instance that is positive and is also predicted as positive; FN (False Negative) represents incorrectly classified negative samples, i.e., a positive training instance that is predicted as negative; FP (False Positive) represents incorrectly classified positive samples, i.e., a negative training instance that is predicted as positive; TN (True Negative) represents correctly classified negative samples, i.e., a training instance that is negative and is also predicted as negative.

5. Experiments and Analysis

5.1. Stage 1: Comparative Experiments on LLMs Binary Classification with Embedded Universal Semantic Features

(1) Effectiveness Analysis of Semantic Feature Encoding

To validate the advantages of our proposed semantic feature encoding mechanism over traditional methods, we designed systematic comparative experiments on the DBPedia dataset. The experiment selected 20,000 texts each for two categories: Company and Educational Institution, forming a binary classification dataset with a positive-to-negative sample ratio of 1:1, totaling 40,000 samples. The dataset was divided into training set (32,000 samples), validation set (4,000 samples), and test set (4,000 samples) in an 8:1:1 ratio.

The experiments employed three pre-trained models—BERT, SciBERT, and RoBERTa—as base architectures, comparing them under both original encoding approaches and our proposed semantic feature encoding approach. The original encoding approach maintains the pre-trained models' [CLS] and [SEP] tokens unchanged, using traditional input sequence encoding; the semantic feature encoding

approach extracts topic semantic features and structural semantic features separately and replaces the embedding vectors at [CLS] and [SEP] positions. All experiments used identical hyperparameter settings: learning rate of 1e-5, batch size of 32, 5 training epochs, and AdamW optimizer [38]. The effectiveness analysis results of semantic feature encoding are shown in Table 1.

Table.1 The results of semantic feature encoding effectiveness analysis

Models	Embedding	A (%)	P (%)	R (%)	F1 (%)
BERT	original encoding	82.40	89.10	79.20	83.80
BERT	semantic feature encoding	**86.85 (+4.45)**	**91.20 (+2.10)**	**85.30 (+6.10)**	**88.17 (+4.37)**
SciBERT	original encoding	84.20	90.50	81.60	85.81
SciBERT	semantic feature encoding	**89.75 (+5.55)**	**92.80 (+2.30)**	**89.40 (+7.80)**	**91.07 (+5.26)**
RoBERTa	original encoding	83.60	89.80	80.40	84.85
RoBERTa	semantic feature encoding	**89.20 (+5.60)**	**92.50 (+2.70)**	**88.90 (+8.50)**	**90.67 (+5.82)**

The effectiveness analysis results of semantic feature encoding are shown in Table 7. The experimental results demonstrate that our proposed semantic feature encoding approach achieves significant performance improvements across all pre-trained models. On the BERT model, the semantic encoding approach improved accuracy, precision, recall, and F1-score by 4.45%, 2.10%, 6.10%, and 4.37% respectively compared to the original encoding approach. SciBERT model shows the best overall performance, with the semantic encoding approach achieving an F1-score of 91.07%, representing a 5.26% improvement over original encoding, and accuracy improving from 84.20% to 89.75%. The RoBERTa model also achieved significant improvements under the semantic encoding approach, with accuracy improving by 5.60% and F1-score improving by 5.82%, reaching 90.67%.

Notably, all models achieved the largest improvements in recall, with BERT, SciBERT, and RoBERTa improving by 6.10%, 7.80%, and 8.50% respectively, indicating that the semantic feature encoding mechanism has significant advantages in identifying positive samples. The improvements in precision were relatively stable, with the three models improving by 2.10%, 2.30%, and 2.70% respectively, demonstrating that this method effectively improves recall while maintaining high precision. This indicates that by replacing [CLS] tokens with topic semantic features and [SEP] tokens with structural semantic features, semantic information can be more deeply integrated into the attention computation process of PLMs. Compared to traditional feature concatenation or simple fusion approaches, our replacement strategy enables semantic features to play crucial roles at key positions: topic features at [CLS] positions can better aggregate global semantic representations, while structural features at [SEP] positions can more accurately model hierarchical structural information of texts. Furthermore, SciBERT's specialization in the scientific text domain makes it perform more prominently when combined with semantic features, further demonstrating the advantages of combining domain-adaptive pre-trained models with semantic feature encoding mechanisms.

(2) Effectiveness Analysis of Binary Classification Models with Embedded Universal Semantic Features

To validate the practical application effectiveness of our proposed semantic feature encoding mechanism in scientific text value sentence identification tasks, we conducted comparative experiments on the scientific literature value sentence dataset. We selected full texts from general domain scientific literature and constructed the corpus using manual annotation and iterative semi-automatic annotation methods. This dataset contains 23,912 scientific literature sentences with a positive-to-negative sample ratio of 1:1, divided into training set (19,130 samples), validation set (2,391 samples), and test set (2,391 samples) in an 8:1:1 ratio. The experiment designed 5 groups of comparative methods, covering different technical approaches including PLMs fine-tuning, LLMs zero-shot learning, semantic feature enhancement, and parameter-efficient fine-tuning, to comprehensively evaluate the effectiveness of our method. Among these, PLMs fine-tuning parameters are the same as in Section 3.5.1; LLMs fine-tuning adopts the QLora [39] parameter-efficient fine-tuning method with parameter settings: rank = 64, alpha = 16.

Considering the input requirements of different types of models, this experiment adopted two different fine-tuning data formats for PLMs and LLMs:

1) Fine-tuning Data Format for PLMs

For BERT-base, RoBERTa-base, and SciBERT, we adopt the standard classification task data format, containing Label and Sentence fields, where Label=0 indicates non-value sentences and Label=1 indicates value sentences. Specific data examples are shown in Table 2:

Table.2 PLMs fine-tuning data format

Label	Sentence
0	Deep learning algorithms have been widely applied in various domains and achieved remarkable success.
1	This study aims to develop a novel neural architecture that can significantly improve the accuracy of text classification tasks while reducing computational complexity by 40%.

2) Instruction Fine-tuning Dataset for LLMs

For LLMs, we adopt the instruction fine-tuning format, containing Instruction, Input, and Output fields. The instruction part provides detailed descriptions of the value sentence identification task definition and judgment criteria, with the specific format as follows:

"Instruction": "Determine if the following sentence is a research value sentence. Research value sentences in scientific literature are sentences that explicitly describe the specific contributions, significance, or potential impact of the research work. They clearly state the research value, importance, or benefits that the study provides to the academic field or practical applications. Output 'True' if it is a research value sentence, and 'False' if it is not.",

"Input": "Our proposed method demonstrates superior performance on benchmark datasets, achieving state-of-the-art results with 15% improvement in accuracy compared to existing approaches.",

"Output": "True"

The experimental results are shown in Table 3. The results demonstrate that our proposed semantic feature encoding mechanism achieves consistent performance improvements across models with different architectures. Whether it's the Transformer-based Qwen3-14B, LLaMa-based LLaMa4-17B, or GLM-based GLM4-9B, the performance improvement after adding semantic features ranges from 6%-8%. Particularly, Qwen3-14B shows the most outstanding performance when combined with semantic features and QLora fine-tuning, achieving an F1-score of 94.19%.

Table.3 The experimental results about instruction fine-tuning

Methods	Models	A (%)	P (%)	R (%)	F1 (%)
Fine-tuning PLMs	BERT-base	84.32	83.15	85.62	84.37
	RoBERTa-base	86.45	85.73	87.28	86.50
	SciBERT	88.76	88.42	89.15	88.78
Base-LLMs	Qwen3-14B-base	79.23	76.84	82.47	79.55
	LLaMa4-17B-base	81.67	80.15	83.52	81.80
	GLM4-9B-base	77.89	75.62	80.73	78.10
LLMs Encoding Semantic Feature	Qwen3-14B Encoding Semantic Feature	86.75	85.92	87.84	86.87
	LLaMa4-17B Encoding Semantic Feature	**89.34**	**88.67**	**90.15**	**89.40**
	GLM4-9B Encoding Semantic Feature	84.56	83.74	85.62	84.67
QLora LLMs	Qwen3-14B-QLora	90.12	89.75	90.58	90.16
	LLaMa4-17B-QLora	91.83	91.46	92.27	91.86
	GLM4-9B-QLora	88.94	88.31	89.67	88.98
QLora LLMs Encoding Semantic Feature	Qwen3-14B-QLora Encoding Semantic Feature	**94.15**	**93.82**	**94.56**	**94.19**
	LLaMa4-17B-QLora	92.67	92.34	93.12	92.73

	Encoding Semantic Feature				
	GLM4-9B-QLora Encoding Semantic Feature	91.28	90.85	91.84	91.34

1) Independent Effectiveness Analysis of Semantic Feature Encoding

In the value sentence identification task, incorporating semantic feature encoding significantly improves model recognition performance. Using LLaMa4-17B as the base model, the value sentence identification accuracy reaches 89.34% after incorporating semantic features, representing a 7.67% improvement over the base model without semantic features. Particularly, the recall reaches 90.15%, improving by 6.63% compared to the base model's recall, indicating that semantic feature encoding helps the model capture the vast majority of value sentences, thereby significantly improving recognition accuracy. Similarly, in the value sentence identification task using Qwen3-14B as the base, the model incorporating semantic features improves accuracy, precision, and recall by 7.52%, 9.08%, and 5.37% respectively compared to the base model. The reason is that topic semantic features and structural semantic features specific to value sentences help the model further capture the linguistic patterns and semantic structures of value sentences. By directly embedding semantic representations into [CLS] and [SEP] positions, the model is assisted to focus on semantic information most relevant to value sentence identification, compensating for deep semantic associations that traditional methods might overlook.

2) Synergistic Analysis of Instruction Fine-tuning and Semantic Feature Encoding

The synergistic mechanism between parameter-efficient fine-tuning and semantic feature encoding improves LLMs' performance in binary classification tasks. Using Qwen3-14B as the base model, the QLora fine-tuned version of LLaMa4-17B base model achieves an F1-score of 91.86%, which improves by 0.87% after combining with semantic feature encoding; the GLM4-9B base model's F1-score improves from 88.98% to 91.34% after incorporating semantic features on top of QLora fine-tuning. Particularly, Qwen3-14B-QLora + semantic features shows the best overall performance, achieving an accuracy of 94.19%, which is 2.55% higher than the QLora fine-tuned version. This indicates that parameter-efficient fine-tuning provides a more stable optimization foundation for semantic feature encoding. Through adaptive training on specific tasks, the model can more precisely utilize embedded semantic information for discrimination, forming a dual optimization mechanism of "fine-tuning adaptation + semantic enhancement" that effectively improves the accuracy and robustness of value sentence identification.

3) Comparative Effectiveness Analysis of LLMs vs. PLMs

It is noteworthy that untuned base LLMs perform slightly lower than fine-tuned PLMs in binary classification under zero-shot or few-shot settings. The reason is that PLMs can better adapt to the discriminative boundaries of binary classification tasks through supervised fine-tuning on specific tasks, while LLMs without targeted training have some differences between their generative pre-training objectives and discriminative classification tasks. However, once combined with semantic feature encoding, LLMs' performance rapidly surpasses traditional methods, indicating that the semantic feature encoding mechanism can effectively bridge the gap between generative pre-training and discriminative downstream tasks.

(3) Semantic Feature Ablation Experiments

To thoroughly validate the specific contributions of topic semantic features and structural semantic features in our proposed semantic feature encoding mechanism, we designed systematic feature ablation experiments on the DBPedia dataset. This experiment analyzes the independent contributions and synergistic effects of different features on model performance by progressively removing different semantic feature components. The experimental dataset is the same as in Section 3.5.1 Semantic Feature Encoding Effectiveness Analysis. The feature ablation experiment designed the following 4 configuration schemes, using SciBERT as the base model for comparative analysis: (1) Original SciBERT: maintains traditional [CLS] and [SEP] tokens as the baseline method; (2) SciBERT + Topic Semantic Features: only replaces [CLS] tokens with topic semantic features; (3) SciBERT + Structural Semantic Features: only replaces [SEP] tokens with structural semantic features; (4) SciBERT + Complete Semantic Features: simultaneously embeds both topic semantic features and structural

semantic features. All experiments used identical hyperparameter settings: learning rate of 1e-5, batch size of 32, 5 training epochs, and AdamW optimizer.

The semantic feature ablation experiment results are shown in Table 4. The experimental results demonstrate that each component in our proposed semantic feature encoding mechanism contributes positively to model performance, and there are obvious synergistic enhancement effects among components. The specific analysis is as follows:

Table.4 The results of semantic feature ablation experiments

Methods	A (%)	P (%)	R (%)	F1 (%)
Original-SciBERT	84.20	90.50	81.60	85.81
SciBERT Encoding Semantic Feature	87.45	91.75	85.20	88.35
SciBERT Encoding Semantic Feature	86.10	91.20	83.80	87.35
SciBERT Encoding Semantic Feature	**89.75**	**92.80**	**89.40**	**91.07**

1) Contribution Analysis about Topic Semantic Features

For topic semantic features, the SciBERT model with topic semantic features embedded alone achieves significant improvements across all evaluation metrics. Among these, SciBERT + Topic Semantic Features achieves an F1-score of 88.35%, representing a 2.54% improvement over the original SciBERT, with accuracy improving from 84.20% to 87.45%. Particularly noteworthy is that topic semantic features show the most significant improvement in recall, increasing from 81.60% to 85.20%, an improvement of 3.60%. This indicates that topic semantic features can effectively reduce the generation of false negative samples and improve the model's ability to identify positive samples. The reason is that topic semantic features, through the dynamic window-based local-global feature extraction method, can more precisely capture the core semantic content of texts. After replacing the [CLS] position, the global semantic representation becomes more focused on the thematic information of the text, thereby enhancing the model's ability to distinguish between different categories of texts.

2) Enhancement Analysis about Structural Semantic Features

For independent contributions of structural semantic features, embedding structural semantic features alone also brings significant performance improvement effects. SciBERT + Structural Semantic Features achieves an F1-score of 87.35%, representing a 1.54% improvement over the original SciBERT, with accuracy improving by 1.90%. Compared to topic semantic features, the improvement magnitude of structural semantic features is relatively smaller, but it shows stable performance in precision, improving from 90.50% to 91.20%, an increase of 0.70%. Experimental data show that structural semantic features primarily capture textual organizational structure information through hierarchical feature aggregation mechanisms. After replacing the [SEP] position, they can better model logical relationships between sentences, providing semantic understanding support at the text structure level for the model. Although the independent effect is not as significant as topic semantic features, they provide important structural information supplementation for complete semantic representation.

3) Synergistic Analysis about Complete Semantic Features

Analyzing from the perspective of synergistic effects between topic semantic features and structural semantic features, the complete semantic feature configuration achieves the best comprehensive performance. SciBERT + Complete Semantic Features achieves an F1-score of 91.07%, representing a 5.26% improvement over the original SciBERT. This improvement magnitude exceeds the simple additive effect of using topic semantic features alone (2.54% improvement) and structural semantic features alone (1.54% improvement), indicating significant synergistic enhancement mechanisms between the two types of semantic features. This indicates that topic semantic features and structural semantic features form effective functional complementarity: topic semantic features focus on capturing the content semantics of texts, while structural semantic features concentrate on modeling the organizational forms of texts. The combination of both can provide the model with more comprehensive semantic understanding capabilities.

5.2. Stage 2: MOE Multi-classification with Embedded Fine-grained Semantic Encoding Routing

To validate the advantages of our proposed two-stage classification method over traditional end-to-end multi-classification methods, we designed multi-classification comparative experiments based on the high-quality value sentence collection filtered in the first stage. The experiments use the value sentence fine-grained category dataset, subdividing value sentences into three subcategories: academic value, application value, and innovation value. By comparing the performance of the two-stage classification method with direct multi-classification methods, we validate the effectiveness of the "universal semantic understanding + fine-grained semantic differentiation" architecture. The experiment designed two groups of comparative studies: first, validating the classification advantages of the two-stage method under ideal data distribution on balanced datasets; second, focusing on evaluating the improvement effects of the semantic routing MOE mechanism on minority class sample recognition on imbalanced datasets.

(1) Balanced Dataset Multi-classification Experiments

To validate the advantages of our proposed two-stage classification method over traditional end-to-end multi-classification methods, we designed two core comparative approaches based on the high-quality value sentence collection filtered in the first stage: two-stage classification method and direct multi-classification method. Our constructed balanced dataset contains three subcategories: academic value, application value, and innovation value, with each category containing 1,200 high-quality annotated samples, totaling 3,600 value sentence samples, divided into training set (2,880 samples), validation set (360 samples), and test set (360 samples) in an 8:1:1 ratio. Considering the differences between the two classification paradigms, this experiment adopted corresponding data formats and task settings for different methods:

1) Two-stage Classification Method Data Format

For the two-stage classification method, the second stage input consists of value sentences filtered from the first stage, using a three-class classification task format containing Label and Sentence fields, where Label=0 indicates academic value sentences, Label=1 indicates application value sentences, and Label=2 indicates innovation value sentences. Specific data examples are shown in Table 5.

Table.5 Two-stage fine-tuning data format

Label	Sentence
0	This research contributes to the theoretical understanding of neural network optimization by providing rigorous mathematical proofs for convergence properties.
1	The proposed algorithm can be directly applied to real-time recommendation systems, reducing computational latency by 60% while maintaining accuracy.
2	We introduce a novel attention mechanism that fundamentally differs from existing approaches and opens new research directions in transformer architectures.

2) Direct Multi-classification Method Data Format

For the direct multi-classification method, we adopt a four-class classification task format, directly classifying from original scientific literature sentences, where Label=0 indicates non-value sentences, Label=1 indicates academic value sentences, Label=2 indicates application value sentences, and Label=3 indicates innovation value sentences. To ensure fair comparison, the training data in this format supplements equal amounts of non-value sentence samples based on value sentence samples. Specific data examples are shown in Table 6.

Table.6 Direct multi-classification fine-tuning data format

Label	Sentence
0	Machine learning has been widely adopted across various industries in recent years.
1	Our theoretical analysis reveals fundamental properties of gradient descent dynamics in high-dimensional spaces.
2	The developed system demonstrates practical utility in industrial quality control applications.
3	This work presents a breakthrough approach that challenges conventional assumptions in natural language processing.

The experimental results are shown in Table 7. The results demonstrate that our proposed two-stage classification method comprehensively outperforms traditional direct multi-classification methods on the balanced dataset. The best model, LLaMa4-17B-QLora + Semantic Routing MOE, achieves a macro-average F1-score of 93.35%, representing a 9.04% improvement over the corresponding direct multi-

classification method. The specific analysis is as follows:

1) Inter-category Performance Consistency Advantage Analysis

For performance distribution across categories, the two-stage classification method achieves balanced high-performance results across all value sentence categories. Taking LLaMa4-17B + Semantic Routing MOE as an example, the F1-scores for academic value, application value, and innovation value categories are 93.78%, 92.94%, and 93.33% respectively, with inter-category performance differences of only 0.84%, significantly better than the maximum 2.11% performance gap in direct multi-classification methods. In contrast, direct multi-classification methods show relatively weaker performance when handling application value sentences, with SciBERT achieving an F1-score of 79.55% on application value sentences, significantly lower than the 82.11% for academic value sentences. This indicates that the two-stage architecture effectively eliminates interference from non-value sentences on model judgment through first-stage value sentence filtering, enabling the second-stage MOE mechanism to focus on fine-grained differentiation between value sentence categories, achieving more balanced and stable classification performance.

2) Expert Division Effectiveness Analysis of Semantic Routing MOE Mechanism

For expert division, the MOE framework with embedded fine-grained semantic encoding routing demonstrates obvious specialization advantages in recognizing different types of value sentences. Qwen3-14B + Semantic Routing MOE achieves F1-scores of 91.84%, 90.22%, and 90.78% on the three categories respectively, representing improvements of 2.89%, 3.11%, and 2.94% compared to traditional SciBERT + MOE methods. Particularly, the semantic routing mechanism shows the most significant improvement in application value sentence recognition, with improvement magnitudes generally exceeding 3%. Experimental data show that fine-grained semantic feature encoding can effectively extract unique semantic patterns of different categories of value sentences: academic value sentences focus more on theoretical contributions and methodological innovations, application value sentences emphasize practical effects and performance improvements, and innovation value sentences highlight breakthrough nature and novelty. By embedding these category-specific semantic features into the MOE routing function, semantic similarity-based expert assignment is achieved, forming a precise division of labor pattern.

Table.7 The results about balanced dataset multi-classification method

Method	Models	Academic Value			Application Value			Innovation Value			Macro-average F1(%)
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Direct Multi-classification Method	BERT-base	76.78	78.56	77.66	74.23	76.89	75.54	75.67	77.34	76.49	76.56
	RoBERTa-base	78.89	80.23	79.55	76.45	78.67	77.54	77.23	79.12	78.16	78.42
	SciBERT	81.34	82.89	82.11	78.67	80.45	79.55	79.89	81.23	80.55	80.74
	Qwen3-14B-QLora	82.45	84.12	83.28	79.67	81.89	80.77	80.89	82.45	81.66	81.90
	LLaMa4-17B-QLora	83.67	85.23	84.44	81.23	83.45	82.33	82.78	84.12	83.44	83.40
	GLM4-9B-QLora	80.23	81.78	81	77.89	79.67	78.77	78.45	80.23	79.33	79.70
	BERT-base + MOE	83.45	85.12	84.28	80.67	82.34	81.49	81.89	83.67	82.77	82.85
	RoBERTa-base + MOE	86.23	87.45	86.84	83.67	85.12	84.39	84.78	86.23	85.5	85.58
	SciBERT + MOE	88.67	89.23	88.95	86.45	87.78	87.11	87.23	88.45	87.84	87.97

Our proposed Multi-classification Method	Qwen3-14B-QLora + Semantic Routing MOE	91.23	92.45	91.84	89.67	90.78	90.22	90.34	91.23	90.78	90.95
	LLaMa4-17B-QLora + Semantic Routing MOE	**93.45**	**94.12**	**93.78**	**92.23**	**93.67**	**92.94**	**92.78**	**93.89**	**93.33**	**93.35**
	GLM4-9B-QLora + Semantic Routing MOE	89.78	91.23	90.5	87.45	89.12	88.28	88.67	90.34	89.5	89.43

(2) Imbalanced Dataset Multi-classification Experiments

To validate the effectiveness of our proposed two-stage classification method in handling class distribution imbalance problems, we conducted in-depth comparative experiments on the imbalanced value sentence fine-grained category dataset. This dataset truly reflects the actual distribution of value sentence categories in scientific literature, containing 11,570 academic value sentences, 1,297 application value sentences, and 1,430 innovation value sentences, totaling 14,297 samples, with category distribution ratios of approximately 8.1:0.9:1.0, exhibiting significant imbalance characteristics. The dataset was divided into training set (11,438 samples), validation set (1,430 samples), and test set (1,429 samples) in an 8:1:1 ratio. The experiment focused on comparing the performance of two-stage classification methods and direct multi-classification methods under imbalanced scenarios, particularly focusing on the recognition effects of minority class samples (application value sentences and innovation value sentences), to validate the contribution of the semantic routing MOE mechanism in alleviating class imbalance problems. The experiment adopted the same model configurations and hyperparameter settings as in Section 3.6.1 to ensure comparability of experimental results.

The experimental results are shown in Table 8. The results demonstrate that our proposed two-stage classification method significantly outperforms traditional direct multi-classification methods on imbalanced datasets, particularly showing obvious advantages in minority class sample recognition. The best model, LLaMa4-17B-QLora + Semantic Routing MOE, achieves a macro-average F1-score of 88.05%, representing an 18.56% improvement over the corresponding direct multi-classification method, with a weighted average F1-score of 94.45%, improving by 9.11%. The specific analysis is as follows:

Table.8 The results about imbalanced dataset multi-classification

Method	Models	Academic Value			Application Value			Innovation Value			Macro-average F1(%)	Weighted AverageF1 (%)
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)		
Direct Multi-classification Method	BERT-base	85.67	92.34	88.87	52.34	38.45	44.32	54.78	41.23	47.12	60.1	78.45
	RoBERTa-base	87.23	93.12	90.04	55.67	42.18	48.04	57.89	44.67	50.45	62.84	80.23
	SciBERT	89.45	94.23	91.78	58.23	45.67	51.23	60.45	47.89	53.45	65.49	82.67
	Qwen3-14B-QLora	90.78	95.12	92.9	61.45	48.23	54.07	62.89	50.34	55.93	67.63	84.12
	LLaMa4-17B-	91.34	95.67	93.45	63.78	51.45	57.01	64.23	52.78	58.01	69.49	85.34

	QLora											
	GLM4-9B-QLora	88.67	93.78	91.16	59.12	46.89	52.34	61.45	49.23	54.71	66.07	82.89
Our proposed Multi-classification Method	BERT-base + MOE	89.23	94.56	91.82	68.45	59.78	63.81	70.12	61.34	65.44	73.69	84.23
	RoBERTa-base + MOE	91.67	95.23	93.42	72.34	64.12	68.02	73.45	65.89	69.48	76.97	86.78
	SciBERT + MOE	93.12	96.45	94.76	75.89	68.45	72.01	76.23	69.78	72.85	79.87	89.45
	Qwen3-14B-QLora + Semantic Routing MOE	95.45	97.23	96.33	82.67	76.89	79.67	83.45	78.23	80.77	85.59	92.78
	LLaMa4-17B-QLora + Semantic Routing MOE	**96.23**	**98.12**	**97.16**	**85.34**	**80.45**	**82.84**	**86.78**	**81.67**	**84.16**	**88.05**	**94.45**
	GLM4-9B-QLora + Semantic Routing MOE	94.78	96.89	95.82	80.45	74.23	77.21	81.89	76.45	79.08	84.04	91.67

1) Effectiveness Analysis of Minority Class Sample Recognition Capability

For minority class sample recognition effects, the two-stage classification method achieves significant performance improvements on the two minority categories of application value sentences and innovation value sentences. Taking application value sentences as an example, LLaMa4-17B + Semantic Routing MOE achieves an F1-score of 82.84%, representing a 25.83% improvement over direct multi-classification methods, with recall improving from 51.45% to 80.45%. The F1-score for innovation value sentences improves from 58.01% to 84.16%, an increase of 26.15%. Particularly, SciBERT + MOE shows F1-score improvements of 20.78% on application value sentences and 19.40% on innovation value sentences. The reason is that the semantic routing MOE mechanism can allocate specialized classifiers for minority class samples through expert division, avoiding the "suppression" effect of majority class samples on minority class samples in traditional methods.

2) Quantitative Analysis of Class Imbalance Mitigation

For inter-category performance gaps, the two-stage classification method significantly alleviates performance difference problems caused by class imbalance. In direct multi-classification methods, taking LLaMa4-17B-QLora as an example, the F1-score gap between academic value sentences and application value sentences reaches 36.44% (93.45% vs 57.01%), and the gap between academic value

sentences and innovation value sentences is 35.44%, showing extremely unbalanced inter-category performance. In contrast, the corresponding two-stage method reduces these gaps to 14.32% (97.16% vs 82.84%) and 13.00% (97.16% vs 84.16%) respectively, representing a reduction in performance gaps of over 60%. This indicates that the semantic routing MOE mechanism can provide sufficient attention to minority class samples based on semantic features of dfferent categories, effectively balancing the recognition capabilities across all categories.

(3) Computational Cost Analysis

Although our proposed two-stage classification framework introduces additional modules—such as semantic routing and expert selection—its computational overhead during inference remains minimal from a practical application perspective. This is because the semantic-aware routing mechanism only performs lightweight semantic matching to determine the optimal expert subset, and the expert models are activated sparsely rather than simultaneously. In other words, only a few relevant experts are triggered for each input based on semantic similarity, significantly reducing redundant computation.

Furthermore, compared to traditional dense models where the entire network processes every input instance, the MoE-based design offers computational advantages by distributing the classification task across specialized experts. While this routing mechanism adds a small fixed cost to inference, it is amortized by the efficiency gains from expert sparsity. Empirical results indicate that, although the overall architecture is more complex, the actual computational burden observed during inference is negligible from an application-layer standpoint, especially when deployed in systems with parallel inference capability.

Therefore, we argue that the slight increase in theoretical model complexity is outweighed by its practical benefits in accuracy and specialization, and does not constitute a bottleneck for real-world deployment.

6. Conclusion and Future Work

In this paper, we propose a two-stage classification method for LLMs based on embedded universal semantic feature encoding. The first stage achieves effective value sentence identification by embedding topic semantic features and structural semantic features, while the second stage employs a semantic routing-based MOE framework to accomplish fine-grained classification of value sentences. The specific contributions are summarized as follows:

(1) Based on pre-trained models, this paper proposes a universal text classification model that considers the weight vectors of topic semantic features and structural semantic features. The model constructs semantic feature sets for each type of value sentence from a linguistic perspective and considers the weight differences of different semantic features on the basis of syntactic structure features, calculating the weight vectors of different semantic features.

(2) Utilizing semantic feature vectors and weighting mechanisms, this paper proposes a method for embedding semantic feature vectors at the internal output stage of pre-trained models. By concatenating the weight vectors of semantic features with the output of pre-trained models for training, the method captures and strengthens the semantic expression and contextual information of value sentences, achieving direct interaction between semantic features and model outputs.

(3) In the self-attention layer, the absolute position and relative position information of semantic features are embedded. In the position-wise feed-forward encoding layer, dynamic position encoding is used to adjust the semantic changes of words, and the contextual information of semantic features is embedded to assist the model in capturing more subtle semantic differences.

(4) The MOE framework based on semantic routing indicates that hyperparameter settings are not simple linear relationships and that their interactions often exhibit nonlinear and complex characteristics. By introducing fine-grained semantic encoding routing mechanisms, the model can obtain optimal expert allocation strategies for multi-classification tasks while reducing computational costs.

(5) Results for pre-trained models indicate that: 1) Analysis of different base models shows that compared to BERT base model, SciBERT base model achieves 0.5% higher F1 score for value sentence identification tasks (increasing from 85.81% to 86.31%) and reaches 91.07% after combining with semantic features (5.26% improvement); 2) Analysis of the same base model reveals that incorporating semantic features in text classification tasks enables the model to better capture the linguistic patterns and structures of different sentence types, thereby improving the identification accuracy of models; 3)

Analysis of models with embedded semantic features demonstrates that our proposed method based on semantic feature encoding can more precisely capture subtle semantic variations of value sentences at different positions. By expressing the varying importance of different semantic features through weight differentials, it achieves superior performance across all value sentence identification tasks, particularly reaching an F1 score of 94.19% on the scientific literature value sentence dataset.

(6) Results for LLMs show that: 1) Analysis of different base LLMs reveals that compared to Qwen3-14B base model, LLaMa4-17B base model achieves higher F1 scores across all value sentence types: academic value sentences improving from 86.82% to 89.34%, with application value sentences and innovation value sentences showing improvements of 1.0% and 0.9% respectively; 2) Analysis of fine-tuning strategies indicates that task-specific fine-tuning significantly enhances LLMs' comprehension of scientific text structures, with all models showing an average F1 score improvement of approximately 6%-8% after fine-tuning. Notably, under the two-stage classification architecture, the best model achieved a macro-average F1 score of 93.35% on balanced datasets and 88.05% on imbalanced datasets.

Notably, this paper optimizes only the value sentence classification task for English scientific literature. Future research needs to focus on the effectiveness of the proposed method in multilingual text classification, as well as how to further improve model performance by integrating multimodal information such as text, images, and tables and extending to cross-domain application scenarios. Moreover, while our proposed taxonomy of value-bearing sentences is grounded in the conventions of scientific writing, future work may explore how this classification framework can be adapted to domains such as social sciences, where rhetorical structures and value expressions differ. The underlying semantic feature encoding and routing mechanisms are model-agnostic and potentially transferable, provided domain-specific value definitions are available.

Author Contributions

Meng Wang: Research proposition and design, initial draft writing, model experimentation;
Jing Xie: Experimental guidance, paper revision, innovation point refinement;
Yang Li: model experimentation, experimental result organization, paper revision;
Zhixiong Zhang: Paper revision, writing guidance;
Hanyu Li: Writing guidance, experimental improvement and overall control.

Acknowledgements

This study was funded by National Social Science Foundation of ChinaGrant No. 21&ZD329.

References

[1] Ping Feng, Xin Zhang, Jian Zhao, Yingying Wang, Biao Huang; Relation Extraction Based on Prompt Information and Feature Reuse. Data Intelligence 2023; 5 (3): 824–840.
[2] Li Y, Zhang M, Zhang Z, et al. Decoding the Essence of Scientific Knowledge Entity Extraction: An Innovative MRC Framework with Semantic Contrastive Learning and Boundary Perception[C]//Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries. 2024: 1-12.
[3] Wang M, Kim J, Yan Y. Syntactic-Aware Text Classification Method Embedding the Weight Vectors of Feature Words[J]. IEEE Access, 2025, 13: 37572-37590.
[4] CHANG Xuanwei, DUAN Liguo, CHEN Jiahao, et al. Sentiment triplet span-level extraction method with deep fusion of syntactic and semantic features[J/OL]. Computer Science, [2025-05-25].
[5] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
[6] Warner B, Chaffin A, Clavié B, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference[EB/OL]. [2024]. arXiv:2412.13663.
[7] Zhao J, Lan M, Niu Z Y, et al. Integrating word embeddings and traditional NLP features to measure textual entailment and semantic relatedness of sentence pairs[C]//Proc. Int. Joint Conf. Neural Netw. 2015: 1-7.
[8] YANG Dawei, XU Xihai, SONG Wei. Relation extraction method combining semantic enhancement and perceptual attention[J/OL]. Journal of Computer Applications, [2025-05-25].
[9] DU Qiliang, WANG Yimin, TIAN Lianfang. Attention module based on feature similarity and feature normalization[J]. Journal of South China University of Technology (Natural Science Edition), 2024, 52(7): 62-71.
[10] Wang M, Zhang Z, Li H, et al. An Improved Meta-Knowledge Prompt Engineering Approach for Generating Research

Questions in Scientific Literature[C]//Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR. 2024: 457-464.
[11] Achiam J. GPT-4 technical report[EB/OL]. [2023]. arXiv:2303.08774.
[12] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[EB/OL]. [2023]. arXiv:2302.13971.
[13] LI Jiangtao, MA Li, LI Yang. Medical data classification method based on large-small model fusion[J/OL]. Computer Engineering, [2025-05-25].
[14] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. [2019]. arXiv:1907.11692.
[15] LI Xuesi, ZHANG Zhixiong, LIU Yi, et al. Research on identification method of research problem sentences in scientific literature[J]. Library and Information Service, 2023, 67(9): 132-140.
[16] Liu P, Cao Y. A named entity recognition method for Chinese winter sports news based on RoBERTa-WWM[C]//Proc. 3rd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE). 2022: 785-790.
[17] Liu X, Zhao W, Ma H. Research on domain-specific knowledge graph based on the RoBERTa-wwm-ext pretraining model[J]. Comput. Intell. Neurosci., 2022: 1-11.
[18] Han Y. Advancing Text Analytics: Instruction Fine-Tuning of QianWen-7B for Sentiment Classification[C]//Proceedings of the 2023 4th International Conference on Big Data Economy and Information Management. 2023: 90-93.
[19] Zhang Y, Wang M, Ren C, et al. Pushing the limit of LLM capacity for text classification[EB/OL]. [2024]. arXiv:2402.07470.
[20] Chae Y, Davidson T. Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning[J]. Sociological Methods & Research, 2024: 00491241251325243.
[21] Fatemi S, Hu Y, Mousavi M. A Comparative Analysis of Instruction Fine-Tuning Large Language Models for Financial Text Classification[J]. Management Information Systems, 2025, 16(1).
[22] Peng L, Shang J. Incubating text classifiers following user instruction with nothing but LLM[EB/OL]. [2024]. arXiv:2404.10877.
[23] Meguellati E, Zeghina A, Sadiq S, et al. LLM-based Semantic Augmentation for Harmful Content Detection[EB/OL]. [2025]. arXiv:2504.15548.
[24] Guo Y, Ovadje A, Al-Garadi M A, et al. Evaluating large language models for health-related text classification tasks with public social media data[J]. Journal of the American Medical Informatics Association, 2024, 31(10): 2181-2189.
[25] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive Mixtures of Local Experts[J]. Neural Computation, 1991, 3.
[26] Peralta B, Soto A. Embedded Local Feature Selection Within Mixture of Experts[J]. Information Sciences, 2014.
[27] Le T D, Jouvet P, Noumeir R. Improving transformer performance for french clinical notes classification using mixture of experts on a limited dataset[EB/OL]. [2023]. arXiv:2303.12892.
[28] Chen S, Jie Z, Ma L. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms[EB/OL]. [2024]. arXiv:2401.16160.
[29] Chowdhury M N R, Wang M, Maghraoui K E, et al. A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts[EB/OL]. [2024]. arXiv:2405.16646.
[30] Wu H, Qiu Z, Wang Z, et al. GW-MoE: Resolving Uncertainty in MoE Router with Global Workspace Theory[EB/OL]. [2024]. arXiv:2406.12375.
[31] Lin X V, Shrivastava A, Luo L, et al. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts[EB/OL]. [2024]. arXiv:2407.21770.
[32] Lehmann J, Isele R, Jakob M, et al. DBpedia--a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.
[33] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text[EB/OL]. [2019]. arXiv:1903.10676.
[34] Yang A, Li A, Yang B, et al. Qwen3 technical report[EB/OL]. [2025]. arXiv:2505.09388.
[35] Meta A I. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation[EB/OL]. [2025-04-07].
[36] GLM T, Zeng A, Xu B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools[EB/OL]. [2024]. arXiv:2406.12793.
[37] Wang M, Shao T, Zhang Z, et al. Leveraging Weight Vectors of Feature Words for Research Question Identification in Scientific Articles[C]//2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2024.
[38] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]//International Conference on Learning Representations (ICLR). 2019.
[39] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: Efficient finetuning of quantized LLMs[C]//Advances in Neural Information Processing Systems (NeurIPS). 2023.

Author Biography

Meng Wang was born in 1992, holding a Ph.D. degree with research interests in natural language processing, large language model reasoning, and knowledge graphs.
Jing Xie was born in 1983, holding a master's degree with research interests in knowledge discovery services, AI semantic retrieval, and information extraction.
Yang Li was born in 2000 with research interests in natural language processing and multimodal information extraction.
Zhixiong Zhang was born in 1971, holding a Ph.D. degree with research interests in natural language processing and information extraction.
Hanyu Li was born in 1986, currently pursuing a Ph.D. degree with research interests in natural language processing and information extraction.