Multi-task Specialized Expert Model for Hierarchical Aspect-based Sentiment Analysis in Consumer Healthcare

Jiaxuan Li[a,b], Jielong Guo[a,b], Patrick Pang[a,c,*], Hugo Gonçalo Oliveira[b], Benjamin K. Ng[a] and Tao Tan[a]

[a]Faculty of Applied Sciences, Macao Polytechnic University, Macao, 918000, China
[b]Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, University of Coimbra, Coimbra, 3030-290, Portugal
[c]MPU-UC Joint Research Laboratory in Advanced Technologies for Smart Cities, Faculty of Applied Sciences, Macao Polytechnic University, Macao, 918000, China

ARTICLE INFO

Keywords:
Aspect-based sentiment analysis
Large language model
Expert model
Patient reviews
Affective computing
Healthcare management

ABSTRACT

Patient reviews offer valuable insights into patients' experiences and perceptions of healthcare services. However, sentiment analysis of these reviews remains challenging due to limited classification granularity, high annotation demands, and the suboptimal performance of existing Aspect-Based Sentiment Analysis (ABSA) methods. These constraints hinder the precise identification of clinical departments in need of improvement and limit the fine-grained understanding of patients' dissatisfaction and expectations, thereby restricting targeted healthcare management. To overcome these challenges, we propose a Hierarchical Aspect-Based Sentiment Analysis (H-ABSA) framework tailored to healthcare contexts. H-ABSA introduces a clinically-informed hierarchical labeling system comprising five primary aspects and twelve secondary sub-aspects, enabling multi-level sentiment analysis from both semantic and clinical perspectives. Furthermore, we develop a Multi-task Specialized Expert Model (MuSEM) that coordinates multiple Expert Models (EM) via a training-free directed routing mechanism. To enhance cross-task semantic understanding and feature sharing, MuSEM integrates Dual and Triple Cross-Attention (DCA and TCA) modules. Experimental results demonstrate that our approach notably improves both model performance and sentiment granularity. It achieves micro and macro $F1$ scores of 0.82 and 0.81, respectively, in end-to-end (E2E) ABSA tasks. This work establishes a practical and extensible framework for consumer-oriented feedback analysis, offering meaningful implications for data-driven healthcare management.

1. Introduction

Affective computing and sentiment analysis have made significant progress with the advancement of Deep Learning (DL), leading to an increasing number of applications across various fields [1], [2], [3], [4]. In healthcare, sentiment analysis emerged as a critical tool for processing patient reviews to extract actionable insights [5], [6]. For the consumer, this means their voice is not just heard, but systematically utilized to refine the very care they depend on. However, traditional ABSA methods, while effective in general domains, exhibit significant limitations in healthcare contexts. A key issue is that a single sentence may contain different sentiment subjects and outcomes, making a single label insufficient to capture the emotional nuance of the text [7]. For example, in the comment "The doctors are very patient, but the hospital has long waiting times," sentiments towards the particular type of staff (i.e., "doctors") and the medical institution (i.e., "hospital") are opposite. ABSA enables the division of text into distinct aspects for separate analysis [8].

On the other hand, the prevailing barrier to clinical applicability is the insufficient granularity of current ABSA approaches. Classifying sentiments only under broad aspect categories, such as "doctors" or "hospitals," fails to capture the multifaceted nature of patient experiences, which may pertain to clinician demeanor, treatment efficacy, facility cleanliness, or administrative processes [9]. This lack of details presents a major obstacle for healthcare management, as it prevents institutions from identifying specific areas for quality improvement, optimizing resource allocation, and ultimately enhancing patient satisfaction and care outcomes [6]. Moreover, attempts to expand the number of aspects significantly raise the complexity of ABSA and reduce accuracy [10]. These challenges are compounded by the complex clinical semantics and specialized terminology in healthcare reviews, which existing ABSA frameworks struggle to address.

The evolution of ABSA methods reflects these difficulties. Before the development of DL, machine learning approaches required extensive annotated data and multiple classifiers [11], [12]. The advent of DL and transformer-based models like BERT improved performance [13], [14], but subsequent enhancements eventually faced limits due to data scarcity [15], [16]. The emergence of Large Language Models (LLMs) has offered new solutions [17], [18], [19]; however, they still struggle with complex ABSA tasks [7] and raise privacy concerns when handling sensitive data, such as patient reviews [20]. This is because high-performance LLMs are typically proprietary and run on

*Corresponding author
jiaxuan.li@mpu.edu.mo (J. Li); jielong.guo@mpu.edu.mo (J. Guo); mail@patrickpang.net (P. Pang); hroliv@dei.uc.pt (H. Gonçalo Oliveira); bng@mpu.edu.mo (B.K. Ng); taotan@mpu.edu.mo (T. Tan)
ORCID(s): 0000-0001-7647-6839 (J. Li); 0000-0001-6779-3840 (J. Guo); 0000-0002-8820-5443 (P. Pang); 0000-0002-5779-8645 (H. Gonçalo Oliveira); 0000-0001-5901-5694 (B.K. Ng); 0000-0001-5403-0887 (T. Tan)

the infrastructure of the companies that own them, where data must be sent to [21]. While open-source LLMs can be deployed locally to address data privacy concerns [22], their computational demands are high, with smaller, more deployable LLMs (sLLM, fewer than 10B parameters) exhibiting lower accuracy [23], [24]. Therefore, exploring approaches that leverage sLLMs for complex problems remains a valuable research challenge [25]. To address these issues, researchers have proposed mixture expert models [26] and multi-LLM collaboration frameworks [27], [28]. Additionally, decomposing ABSA into sequential tasks, such as Aspect Category Detection (ACD) and Aspect Category Sentiment Analysis (ACSA), has been shown to improve accuracy [29]. Building on these advances, we propose a hierarchical ABSA framework tailored for healthcare, which constructs a multi-level sentiment classification system comprising five primary aspects and twelve sub-aspects. This enables fine-grained analysis aligned with clinical needs.

We further introduce the MuSEM, which decomposes ABSA into four sub-tasks: ACD-1, ACD-2, ACSA, and Intent Recognition (IR). Each sub-task is handled by a specialized EM fine-tuned using the Weight-Decomposed Low-Rank Adaptation (DoRA) method [30] and trained on our newly developed L_{3000} dataset (extended from LRS_{500} [31]). The MuSEM framework uses a specialized routing mechanism to pass features between experts, employing cross-task attention DCA and TCA for comprehensive reasoning. This reduces router parameters and improves efficiency compared to traditional Mixture-of-Experts (MoE) frameworks. Experimental results demonstrate that our approach significantly outperforms task-specific ABSA baselines and achieves a favorable performance–privacy–deployability trade-off compared with closed-source LLMs. After applying H-ABSA and MuSEM, $F1_{micro}$ increased by 34.7% and $F1_{macro}$ by 50.2% compared to baseline sLLMs. Ablation experiments confirm the effectiveness of few-shot and Chain-of-Thought (CoT) methods in our framework [32]. The main contributions of this work are summarized as follows:

- We propose a novel paradigm for healthcare ABSA by introducing a universal multi-sLLM framework, which reuses and orchestrates existing models through a systematic screening–fine-tuning–collaboration pipeline. This design eliminates the need for repeatedly training large models from scratch and significantly reduces computational cost.
- We design a hierarchical and clinically-grounded aspect taxonomy for H-ABSA, enabling fine-grained and interpretable modeling of patient reviews. Compared with existing coarse-grained schemes, our framework provides a more structured and clinically meaningful representation.
- We develop MuSEM, a multi-expert collaborative framework that dynamically integrates expert models via specialized routing and cross-task knowledge sharing. This design improves performance while maintaining efficient memory usage.
- Extensive experiments on multiple datasets demonstrate that our approach consistently outperforms task-specific ABSA baselines while maintaining practical efficiency, stability, and generalization ability.

2. Literature Review

Sentiment analysis represents a fundamental task within Natural Language Processing (NLP), with extensive applications across diverse domains [33], [19], [34]. Traditional machine learning approaches typically formulate sentiment analysis as a supervised text classification problem, learning complex mappings between linguistic features and sentiment polarity through annotated corpora. These models can recognize intricate linguistic phenomena such as context dependency, negation, and irony, though their performance is often constrained by feature engineering limitations [35]. The advent of deep learning has significantly advanced the capabilities of sentiment analysis. Recurrent Neural Networks (RNNs) and their variants, particularly Long Short-term Memory (LSTM) networks, demonstrated improved performance in capturing sequential dependencies in text [36]. Subsequent research introduced attention mechanisms to focus on sentiment-relevant portions of text, further enhancing model interpretability and accuracy [37]. More recently, transformer-based architectures have revolutionized the field, with models like BERT achieving state-of-the-art results across multiple sentiment analysis benchmarks [38]. The practical significance of these advancements lies in their ability to bridge the gap between consumer feedback and service optimization [6]. By accurately capturing the nuances of patient experiences, healthcare providers can identify systemic issues and service gaps in real-time [18].

Online reviews and consumer-generated feedback have become critical information sources for enterprises to gain insights into market demand and improve products and services, playing a vital role in understanding consumer sentiments, identifying key requirement shifts, and supporting data-driven decision-making [39], [34]. In this context, prior studies have developed dynamic customer requirement models based on Bidirectional LSTM and related approaches, integrating sentiment analysis to characterize the evolutionary patterns of consumer requirements [39]. Meanwhile, other research has incorporated models such as Multi-Layer Perceptron (MLP) and CRITIC to achieve quantitative consumer requirement modeling and risk assessment[34].

Within the healthcare domain, sentiment analysis presents unique challenges that distinguish it from general-domain applications. Patient reviews often contain complex clinical terminology, implicit sentiment expressions, and nuanced emotional states that require domain-specific understanding [40]. Existing healthcare sentiment analysis approaches primarily focus on document-level or sentence-level sentiment classification, lacking the granularity needed for actionable clinical insights [41]. While some studies have explored aspect-based approaches in healthcare contexts, they typically employ coarse-grained sentiment labels

that fail to capture the multidimensional nature of patient experiences [18]. Multi-task learning has gained increasing attention in sentiment analysis, with frameworks that jointly optimize related tasks to improve overall performance [42]. Recent approaches have explored hierarchical task decomposition, where complex sentiment analysis is broken down into specialized subtasks [8].

Recent studies have explored several advanced paradigms for medical ABSA, including prompt-based learning [18], knowledge-enhanced methods [43], and multi-modal approaches [44]. Prompt-based methods leverage large language models by designing task-specific prompts to guide the model in extracting aspect-level sentiment information [18]. These approaches benefit from the strong generalization ability of LLMs and can reduce the need for task-specific fine-tuning. However, their performance is often sensitive to prompt design and may lack stability across domains. Knowledge-enhanced methods incorporate external medical knowledge, such as clinical ontologies or domain-specific knowledge graphs, to improve semantic understanding and representation learning. By integrating structured knowledge, these methods can better capture domain-specific relationships and terminology [43]. Nevertheless, they typically rely on the availability and quality of external knowledge resources, which may limit scalability and increase system complexity. Multimodal approaches further extend ABSA by integrating heterogeneous data sources [44], such as textual reviews, medical images, and structured clinical data. These methods aim to provide a more comprehensive understanding of patient feedback by combining complementary modalities. Despite their effectiveness, multimodal models often require additional data alignment and incur higher computational and implementation costs. However, these methods often rely on resource-intensive model training and lack efficient mechanisms for cross-task feature sharing, limiting their practical deployment in healthcare settings.

The emerging field of lightweight language model collaboration offers promising directions for resource-constrained environments. Techniques such as model distillation and expert routing have demonstrated effectiveness in maintaining high performance while reducing computational requirements [45]. However, these approaches have not been systematically adapted to address the specific challenges of healthcare sentiment analysis, particularly in terms of hierarchical sentiment classification and cross-task semantic understanding. Our work bridges these research gaps by introducing H-ABSA, a hierarchical sentiment classification framework specifically designed for healthcare contexts, and MuSEM, a novel multi-task EM framework that efficiently coordinates specialized lightweight models through specialized routing and cross-task attention mechanisms.

Unlike conventional Mixture-of-Experts (MoE) frameworks that rely on probabilistic or learned gating functions, our MuSEM introduces a deterministic Specialized Routing Mechanism that eliminates the need for additional router training. This design not only prevents routing errors caused by suboptimal gating weights but also enables seamless feature reuse across subtasks through the DCA and TCA modules. In contrast to prior multi-LLM collaboration studies [26], [27], [28], our approach provides a mathematically grounded mechanism for task-level information sharing without increasing model parameters. In conclusion, our framework addresses both the granularity requirements of clinical sentiment analysis and the practical constraints of healthcare deployment environments.

3. Methods

In this section, we introduce the proposed framework, including the hierarchical aspect taxonomy and the MuSEM architecture. We first describe the overall design, followed by detailed explanations of each component.

3.1. Data

RateMDs[1] is one of the most influential medical review websites in Canada and the United States. It allows patients to review any healthcare provider, aiming to reduce information asymmetry between patients and providers by sharing patient experiences and helping the public make more informed healthcare choices. Since its creation in 2004, the platform has collected over three million patient reviews. Due to its valuable research potential, numerous researchers have conducted study using RateMDs data, [46]. In this study, we conducted an ABSA analysis using only patient review content. Our data originated from an open-source RateMDs dataset[2] shared by a researcher who previously analyzed whether online healthcare reviews exhibit bias against female healthcare providers [47]. This study compared ratings and linguistic features between male and female physicians using statistical regression and neural network embedding models [47]. The dataset contains over 30,000 patient reviews from evaluations of various hospitals and physicians.

3.2. ABSA problem definition

Traditional ABSA systems handle the sentiments of patient feedback too broadly, making it difficult for healthcare units to implement targeted improvements based on such coarse classifications. To better illustrate the issues of existing ABSA methods, we can have a look at the example shown in Figure 1. When classifying feedback based on physician-related sentiments, Patient 1 expressed dissatisfaction with the doctor's attitude, while Patients 2 and 3 were dissatisfied with the doctor's medical skills. However, both the first classification method [48] and the second classification method [49] grouped all three patients into the same aspect category related to a doctor.

In actual healthcare management, these three cases require distinct handling approaches. If poor physician attitude recurs (Patient 1), in-depth communication and service training should be provided to address the issue. If treatment

[1] https://www.ratemds.com

[2] Open-source dataset available at: https://github.com/avi-otterai/RateMDs

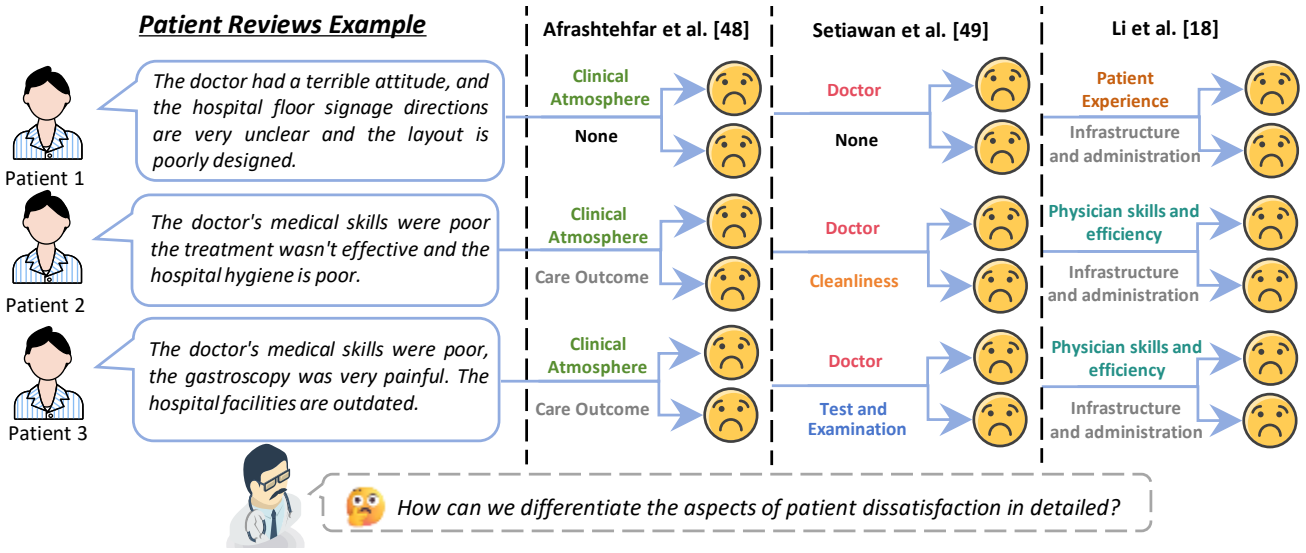

Figure 1: Existing classification of sentiment aspects in patient reviews using ABSA and their limitations.

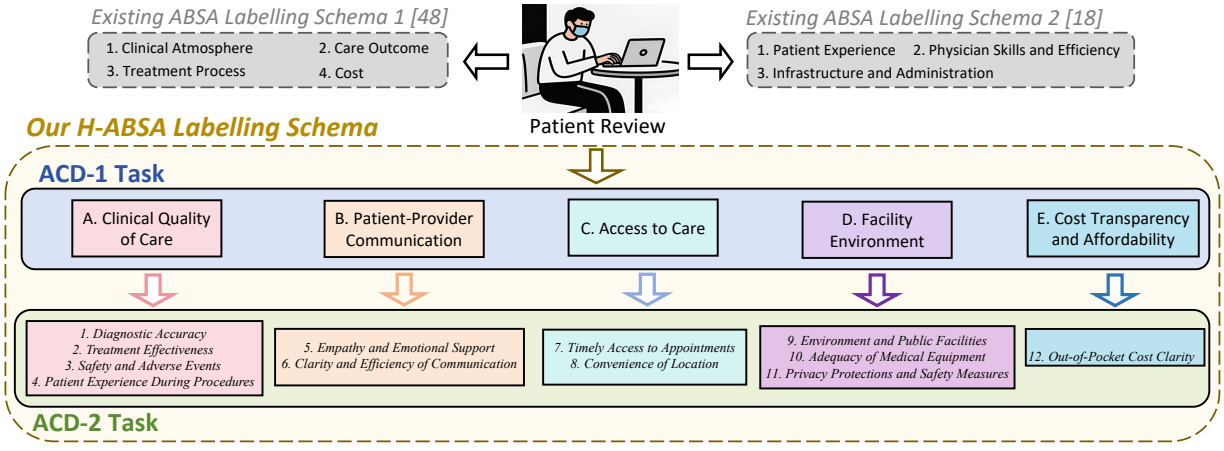

Figure 2: Hierarchical ABSA and Labelling Schema for Patient Reviews.

ineffectiveness persists (Patient 2), training and assessment on therapeutic medication should be implemented. If non-standard testing procedures occur repeatedly (Patient 3), training and assessment on laboratory workflow protocols should be conducted. Thus, overly simplistic ABSA classification standards prove inadequate for facilitating improvements by healthcare administrators. Moreover, the second classification correctly identifies the aspects "Cleanliness" and "Test and Examination". However, it does not include a category corresponding to floor guidance and planning, so this aspect is labeled as <none>. Finally, although the third classification [18] distinguishes between physician attitudes and medical skills, it groups floor guidance, poor hygiene, and outdated facilities into a single category, making differentiation impossible. In summary, while the aforementioned ABSA methods represent improvements over traditional classification approaches, they remain too simplistic for practical application in clinical and real-world healthcare settings. To address the issue of insufficient aspect granularity, we propose a multi-level hierarchical framework, as illustrated in Figure 2 and summarized in Table 1. Specifically, all patient review aspects are organized into five primary categories:

A. Clinical Quality of Care [50],
B. Patient–Provider Communication [51],
C. Access to Care [52],
D. Facility Environment [53], and
E. Cost Transparency and Affordability [54].

Each primary aspect is further decomposed into a set of clinically meaningful sub-aspects, as detailed below.

Table 1
Hierarchical aspect scheme with clinical interpretation.

Primary Aspect	Sub-aspect	Clinical Interpretation
A Clinical Quality of Care	A.1 Diagnostic Accuracy	Reflects the correctness of diagnosis and clinical judgment
	A.2 Treatment Effectiveness	Indicates whether medical interventions achieve expected outcomes
	A.3 Safety and Adverse Events	Captures potential risks, complications, and side effects during treatment
	A.4 Patient Experience During Procedures	Reflects patient comfort and experience throughout clinical procedures
B Patient–Provider Communication	B.5 Empathy and Emotional Support	Represents emotional care and psychological support provided by clinicians
	B.6 Clarity and Efficiency of Communication	Reflects the clarity and effectiveness of information delivery
C Access to Care	C.7 Timely Access to Appointments	Indicates waiting time and availability of medical services
	C.8 Convenience of Location	Reflects accessibility and geographical convenience of healthcare services
D Facility Environment	D.9 Environment and Public Facilities	Represents cleanliness and comfort of healthcare environments
	D.10 Adequacy of Medical Equipment	Reflects availability and quality of medical equipment
	D.11 Privacy Protections and Safety Measures	Indicates protection of patient privacy and safety standards
E Cost Transparency and Affordability	E.12 Out-of-Pocket Cost Clarity	Reflects transparency and reasonableness of medical expenses

Category A: Clinical Quality of Care. This category reflects the core dimension of healthcare delivery, including clinical effectiveness, safety, and patient experience. It consists of:

1. Diagnostic Accuracy [55],
2. Treatment Effectiveness [56, 57],
3. Safety and Adverse Events [58, 59], and
4. Patient Experience During Procedures [60].

Category B: Patient–Provider Communication. This category captures the interpersonal and informational aspects of care, emphasizing trust and effective communication. It includes:

5. Empathy and Emotional Support [61], and
6. Clarity and Efficiency of Communication [62].

Category C: Access to Care. This category focuses on healthcare accessibility from an organizational perspective, including temporal and spatial accessibility. It includes:

7. Timely Access to Appointments [63, 64], and
8. Convenience of Location [65].

Category D: Facility Environment. This category reflects the physical and institutional environment that shapes patient experience and perception. It consists of:

9. Environment and Public Facilities [66],
10. Adequacy of Medical Equipment [67], and
11. Privacy Protections and Safety Measures [68].

Category E: Cost Transparency and Affordability. This category addresses financial clarity and economic accessibility of healthcare. It includes:

12. Out-of-Pocket Cost Clarity [69].

In addition to aspect categorization, we further introduce an intent recognition (IR) layer to capture the underlying motivations of patient reviews, as shown in Figure 3. This layer consists of three categories: complaints and grievances (C&G), initiatives and suggestions (I&S), and confusion and inquiries (C&I) [70, 71, 72].

On the other hand, in order to validate the proposed aspect hierarchy, we conducted an expert evaluation involving ten licensed physicians with clinical experience. The experts were recruited from multiple tertiary hospitals, including West China Hospital of Sichuan University, The First Affiliated Hospital of Sun Yat-sen University, and Shenzhen People's Hospital, ensuring diverse clinical perspectives. Each expert independently assessed the rationality and clinical relevance of both primary and sub-aspects using a 5-point Likert scale. As shown in Table 2, all aspects achieved satisfactory scores, indicating that the proposed taxonomy is well-aligned with real-world medical analysis requirements. The inter-rater agreement, measured by Fleiss' kappa, ranges from 0.75 to 0.84, demonstrating substantial agreement among experts. These results confirm that the constructed aspect hierarchy is both clinically meaningful and reliable, and can effectively support downstream medical text analysis tasks.

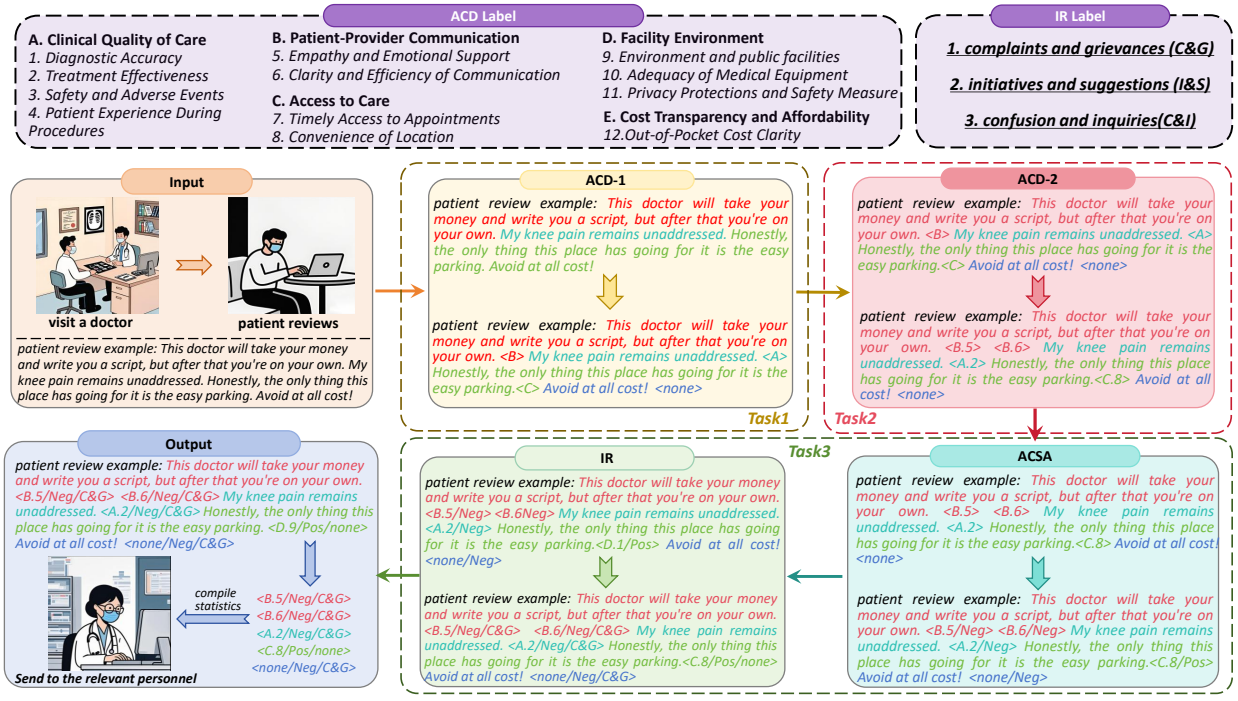

Figure 3: Flowchart of the proposed H-ABSA and associated label annotations.

Table 2
Doctor evaluation of the proposed aspect hierarchy.

Aspect	ID	Score	κ
Clinical Quality of Care	A	4.7	0.84
Diagnostic Accuracy	A.1	4.5	0.83
Treatment Effectiveness	A.2	4.6	0.83
Safety and Adverse Events	A.3	4.4	0.79
Patient Experience During Procedures	A.4	4.5	0.80
Patient–Provider Communication	B	4.6	0.83
Empathy and Emotional Support	B.5	4.4	0.78
Clarity and Efficiency of Communication	B.6	4.3	0.78
Access to Care	C	4.5	0.80
Timely Access to Appointments	C.7	4.4	0.80
Convenience of Location	C.8	4.1	0.78
Facility Environment	D	4.3	0.78
Environment and Public Facilities	D.9	4.2	0.77
Adequacy of Medical Equipment	D.10	4.0	0.75
Privacy Protections and Safety Measures	D.11	4.2	0.78
Cost Transparency and Affordability	E	4.3	0.76
Out-of-Pocket Cost Clarity	E.12	4.3	0.77

3.3. Hierarchical aspect-based sentiment analysis (H-ABSA)

To address the aforementioned issues and improve upon the limitations of previous ABSA research, the overall framework of our proposed H-ABSA is illustrated in the lower portion of Figure 3. Upon acquiring patient review data, this serves as the input to our framework, subsequently entering Task I. This task involves the first round of labeling within the ACD component, aiming to identify the primary aspect category from the five major aspects (A, B, C, D, E) outlined above. As illustrated in the Task I section below, our model classifies the first sentence of a patient review, "This doctor will take your money and write you a script, but after that you're on your own," and tags it with <B>, indicating that the model categorizes this sentence under "B. Patient-provider Communication" within the 5 major aspect categories. Subsequently, the model applies this labeling process to each sentence. If a sentence corresponds to multiple major aspects, the model assigns multiple labels to it. If a sentence does not fall within any category, the model labels it as <none>. The process then proceeds to Task II within the framework. Task II primarily performs a second round of detailed ACD categories, applying more granular labels based on the tags from the previous round. For example, the first sentence from the patient review receives the labels <B.5>and <B.6>in Task II. This indicates that the model identifies two aspects: "B.5 Empathy and Emotional Support" and "B.6 Clarity and Efficiency of Communication."

After completing the two-layer ACD process, the framework proceeds to Task III, which consists of two parts. The first part is ACSA, where the sentences with their assigned ACD labels undergo sentiment analysis to receive sentiment tags <Pos>or <Neg>, indicating whether the model perceives the sentence as positive or negative. The second part of Task III is our additional IR feature. It identifies the intent behind each patient statement to help us meet patient expectations. For example, the first patient comment in this section is labeled <C&G>, indicating that the model perceives the intent as complaints and grievances. Finally, our framework aggregates all labels for statistical analysis

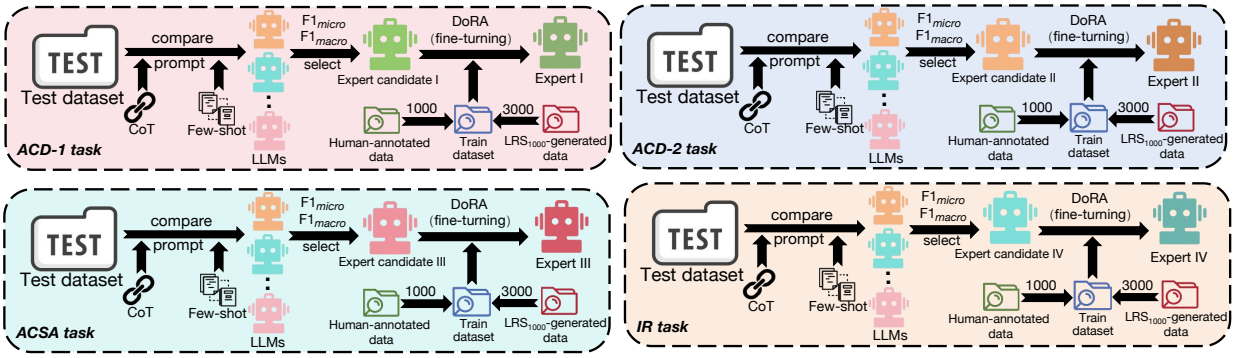

Figure 4: The screening and fine-tuning process of EM.

and distribution to relevant personnel, facilitating big data-driven healthcare quality management and monitoring.

To further illustrate the workflow of the proposed framework, we provide a real example below.

> **Input:** "This doctor will take your money and write you a script, but after that you're on your own. My knee pain remains unaddressed. Honestly, the only thing this place has going for it is the easy parking. Avoid at all cost!"
>
> **Output:** <B.5/Neg/C&G> <B.6/Neg/C&G>
> <A.2/Neg/C&G>
> <C.8/Pos/none>
> <none/Neg/C&G>

This example demonstrates how the proposed H-ABSA framework decomposes a patient review into hierarchical aspect categories, sentiment polarity, and intent labels in a step-by-step manner.

3.4. Multi-task Fine-tuning

Multiple studies indicate that LLMs have limited capabilities for handling complex tasks [73], and fine-tuning serves as an effective and cost-effective optimization approach [74], [75]. Therefore, for this task, we developed a multi-task, multi-expert training scheme tailored to its multi-task nature, ensuring each EM achieves outstanding performance on its specific task. Due to the immense difficulty of labeling this complex, multi-label task, we were unable to complete annotation for a large-scale fine-tuning dataset. Through the collaborative efforts of our annotation team, we successfully annotated 1,000 fine-tuning training examples, 400 test sets and 400 validation sets. Annotation was performed by teams of two: one annotator held a Ph.D. in Data Science, and the other was a clinician at First Affiliated Hospital of Sun Yat-sen University. The annotators worked independently and were blinded to each other's labels. Disagreements were adjudicated by a third expert from the Shenzhen People's Hospital. Since two annotators are involved, agreement is measured using Cohen's kappa ($\kappa = 0.83$). The results indicate substantial agreement, demonstrating the reliability of the annotation process. However, Hellwig et al. (2025) proposed the LRS_{500} method, which leverages LLMs and a low-resource annotated training set (500 examples) to generate new examples that are several times the number of original training examples [31]. Results showed that this approach outperformed training directly on the original annotated dataset. Building upon this research, we adopted the same experimental setup as the original study and proposed LRS_{1000}. This approach uses a 1000-example annotated dataset as the foundation, instructing the LLM to generate an additional 3,000 training examples (L_{3000}). As a result, our dataset is expanded from 1,000 to 4,000 examples. The combined 4,000-example dataset is then used for expert fine-tuning, thus minimizing the issue of insufficient fine-tuning data. Specifically, we first used a human-annotated dataset of 1,000 examples (H_{1000}). Similar to the original method, we randomly selected 50 examples from the 1,000-example dataset as Few-shot examples. We then provided the LLM with a complete answer template specifying the required aspects of the training examples, such as category and sentiment labels. The specific prompt is shown in the supplementary materials.

On the other hand, current smaller LLMs often exceed 5 billion parameters. Employing Full Fine-Tuning (FT) strategies thus demands substantial computational resources and time. In 2021, Microsoft introduced Low-Rank Adaptation (LoRA), a novel fine-tuning method for LLMs [76]. It effectively enhances model performance by updating only a small fraction of parameters, reducing computational requirements by over 10,000 times compared to FT. LoRA improves model performance by updating only a small fraction of parameters, reducing computational cost by over 10,000 times compared to FT. The key idea is to decompose the weight matrix into two low-rank matrices and perform updates only on these, avoiding full weight updates. However, LoRA has certain limitations, specifically:

1. Mapping high-dimensional weights to low-rank matrices may lead to the loss of information or fine-grained emotional features, which can reduce the model's capability in handling complex ABSA tasks.

2. It does not adjust the magnitude of weights, which may prevent the model from fully adapting to the data

distribution in complex ABSA tasks, thereby affecting performance.

To more clearly elucidate the limitations of LoRA in the ABSA task, we analyze its forward computation through Equation (1). In LoRA, matrix A is typically initialized from a random Gaussian distribution, while matrix B is initialized as a zero matrix, ensuring that the low-rank update $\Delta W = BA$ is initially zero and consistent with the pre-trained weights W_0. During the forward pass, both W_0 and ΔW are applied to the input vector x, and their outputs combined:

$$f(x) = W_0 x + \Delta W x = W_0 x + \frac{\alpha}{r} BAx. \tag{1}$$

This equation reveals two key limitations of LoRA. The low-rank decomposition may miss fine-grained or task-specific patterns in x, and the fixed scaling α/r limits the magnitude of $\Delta W x$, constraining adaptation to complex ABSA data. Therefore, we employ DoRA for fine-tuning [30]. Its core idea involves decomposing weights into magnitude and direction components, followed by utilizing LoRA to fine-tune the direction matrix. This approach enables more flexible control over the matrix update direction, aligning more closely with the principles of full fine-tuning (FT). Specifically, as shown in Equation (2) and Equation (3), DoRA decomposes the pre-trained weights $W \in \mathbb{R}^{d\times k}$ into a trainable magnitude vector $m \in \mathbb{R}^{1\times k}$ and a direction matrix $V \in \mathbb{R}^{d\times k}$, where $\odot$ denotes column-wise scaling, and $\|V\|_c$ represents the vector of column L_2-norms:

$$W = m \odot \frac{V}{\|V\|_c}, \tag{2}$$

$$\|V\|_c = [\, \|v_{:,1}\|_2, \|v_{:,2}\|_2, \cdots, \|v_{:,k}\|_2 \,]. \tag{3}$$

However, training both branches from scratch would make the model highly sensitive to initialization methods, thereby compromising training stability and final performance. To mitigate this issue, DoRA opts to initialize both components using pre-trained weights, effectively avoiding training fluctuations caused by suboptimal initialization. As shown in Equation (4), DoRA's initialization process proceeds as follows: First, the entire architecture is initialized using the pre-trained weight matrix W_0. The magnitude component m is initialized as the column norm of W_0 (i.e., $m = \|W_0\|_c$), while the direction component V is initialized as W_0 itself. During subsequent training, V is fixed as a non-trainable parameter, leaving only m as a trainable vector. Directional updates are achieved through LoRA, thereby enhancing stability while maintaining training efficiency.

$$W' = m\frac{V + \Delta V}{\|V + \Delta V\|_c} = m\frac{W_0 + BA}{\|W_0 + BA\|_c}, \tag{4}$$

According to Equation (5), the gradient of the loss function L with respect to m and $V' = V + \Delta V$ is given by:

$$\nabla_{V'}\mathcal{L} = \frac{m}{\|V'\|_c}\left(I - \frac{V'V'^{\mathrm{T}}}{\|V'\|_c^2}\right)\nabla_{W'}\mathcal{L}. \tag{5}$$

Equation (5) indicates that after scaling by $m\|V'\|_c$, the projection direction of the weight gradient $\nabla_{W'}\mathcal{L}$ deviates from that of the current weight matrix. These two effects combined cause the covariance matrix of the gradient to approach the identity matrix, thereby benefiting the optimization process. Furthermore, since $V' = V + \Delta V$, the gradient $\nabla_{V'}\mathcal{L}$ is effectively equivalent to $\nabla_{\Delta V}\mathcal{L}$. Consequently, the optimization advantage gained from this decomposition is fully transferred to ΔV, further enhancing the learning stability of the LoRA method. In summary, our proposed fine-tuning framework for the Multi-task EM is illustrated in Figure 4. We first compare and select all base LLMs using the Test dataset, where we employ identical prompts to describe the task and provide multiple examples as few-shot data for model learning. Finally, we incorporate the CoT approach to enable stepwise reasoning within the model.

To intuitively illustrate the multi-expert fine-tuning strategy for the H-ABSA task based on the DoRA framework, we present the architectural overview in Figure 5. Specifically, we propose a Multi-expert DoRA framework to efficiently fine-tune the pretrained ECM backbones while minimizing computational overhead. The core premise of this architecture is the structural decoupling of the pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ into two orthogonal components: a magnitude vector and a directional matrix. Specifically, the weight decomposition process initializes the magnitude vector $m_0 \in \mathbb{R}^{1\times k}$ by computing the column-wise Euclidean norm of the pretrained weights, denoted as $m_0 = \|W_0\|_c$. Similarly, the directional component is represented by the normalized matrix $V = W_0/\|W_0\|_c$, which is located on a unit hypersphere. During the training phase, the original backbone parameters V are frozen (indicated by the snowflake icon) to preserve the generalized representations of features learned during pretraining, while the magnitude vector m_0 is treated as a trainable parameter (indicated by the flame icon), allowing the model to dynamically rescale the importance of the features.

The adaptation of the directional component is achieved through a residual learning strategy inspired by LoRA. Instead of updating the high-dimensional direction matrix directly, we introduce a low-rank incremental update ΔV, parameterized by the product of two trainable matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, where the rank r is significantly smaller than the model dimensions ($r \ll d, k$). The updated direction is subsequently re-normalized to ensure strict directional alignment, formulated as $\hat{V} = (V + \Delta V)/\|V + \Delta V\|_c$, or equivalently, $\hat{V} = (W_0 + BA)/\|W_0 + BA\|_c$. This normalization step is critical as it decouples the directional adjustments from magnitude changes, stabilizing the optimization landscape. Finally, the adapted weight matrix W' is reconstructed by merging the optimized magnitude and the updated direction via a broadcasting operation $W' = m_0 \odot \hat{V}$. This decomposed adaptation mechanism enables the generation of multiple task-specific weight instances (e.g., $W'_1, \ldots, W'_4$), facilitating versatile model performance across different expectation scenarios (Expect I-IV) without catastrophic forgetting.

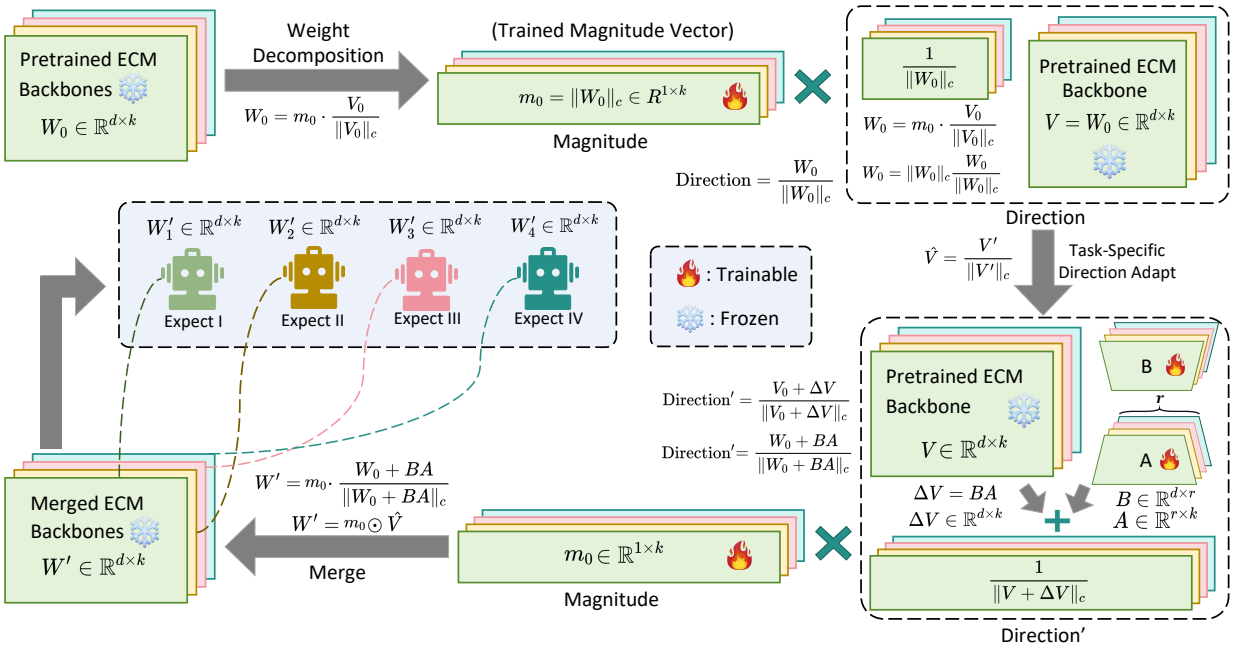

Figure 5: Multi-expert DoRA Adaptation of Multiple Pretrained ECMs for H-ABSA.

3.5. Evaluation Metrics

During this screening phase, our evaluation metrics employ both $F1_{micro}$ and $F1_{macro}$ for comprehensive assessment, ensuring the model maintains performance across both high-frequency and low-frequency aspects in multi-classification tasks. Specifically, $F1_{micro}$ calculates overall performance by aggregating confusion matrix statistics across all aspects, thus emphasizing accuracy in both overall and high-frequency aspects. As shown in Equation (6) and Equation (7), where TP_i denotes the number of samples with true label c_i that are correctly predicted, FP_i denotes the number of samples with true label non-c_i that are misclassified as c_i, and FN_i denotes the number of samples with true label category i that are missed by the model, we first compute the overall Precision and Recall for all aspects:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{12} \text{TP}_i}{\sum_{i=1}^{12}(\text{TP}_i + \text{FP}_i)}, \tag{6}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{12} \text{TP}_i}{\sum_{i=1}^{12}(\text{TP}_i + \text{FN}_i)}. \tag{7}$$

Then, as shown in Equation (8), the F1 score calculated through harmonic mean is referred to as $F1_{micro}$:

$$\text{F1}_{\text{micro}} = 2 \times \frac{\text{Precision}_{\text{micro}} \times \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}. \tag{8}$$

Unlike $F1_{micro}$, $F1_{macro}$ emphasizes equal contribution across all aspects. It calculates each sentiment aspect independently, thereby more accurately capturing the model's performance on low-frequency aspect. As shown in the Equation (9) and Equation (10), $\text{Precision}_{\text{macro}}$ and $\text{Recall}_{\text{macro}}$ represent the averages across all aspects:

$$\text{Precision}_{\text{macro}} = \frac{\sum_{i=1}^{12} \text{Precision}_i}{n}, \tag{9}$$

$$\text{Recall}_{\text{macro}} = \frac{\sum_{i=1}^{12} \text{Recall}_i}{n}. \tag{10}$$

Finally, similar to $F1_{micro}$, we can obtain the corresponding $F1_{macro}$ using the F1 calculation Equation:

$$F1_{\text{macro}} = 2 \cdot \frac{\text{Precision}_{\text{macro}} \cdot \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}}. \tag{11}$$

As shown in Figure 4, after screening the Expert Model Candidate (EMC) using $F1_{micro}$ and $F1_{macro}$, we fine-tuned the model on the training set generated by the aforementioned LRS_{1000} and a portion of manually annotated training data to obtain EM for different tasks. Additionally, for the ACSA task, we designed a Three-step Implicit Sentiment Reasoning (TISR) template. Given a patient comment x_i, the reasoning process for aspect t_i is as follows:

1. The mentioned medical entity or action is [*what*];
2. The implicit clinical impact or sentiment is [*what*];
3. Therefore, the final sentiment polarity is [*what*].

As shown in Figure 7, in the patient's first comment, the model thoroughly considers the medical entities as prescription and payment, provides its inferred implicit impact, and finally completes the overall sentiment polarity aspects based on the above analysis, thereby deeply mining the implicit sentiment within patient comments.

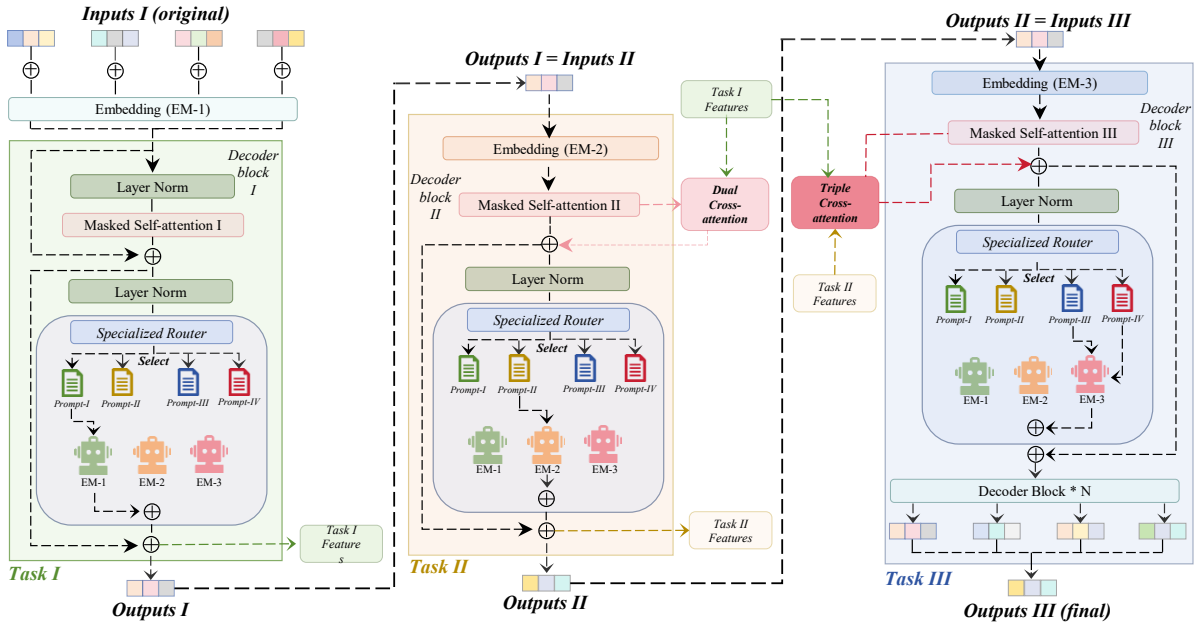

Figure 6: Schematic diagram of the internal structure of the MuSEM framework, illustrating the screening and fine-tuning process of EM.

"This doctor will take your money and write you a script, but after you're in your own."

- Medical Entity: [Prescription, Payment]
- Implicit clinical impact or sentiments: [Lack of follow-up care/support potentially leads to improper medication use, unresolved symptoms, or worsening condition.]
- Emotion Polarity: [neg] (Strongly negative tone: feeling abandoned after payment.)

Figure 7: Thinking process diagram utilising the TISR template in LLM.

3.6. Multi-task Specialized Expert Model Framework

Previous extensive research indicates that the accuracy of ABSA tasks diminishes as the number of aspect categories increases. To address the issue of low accuracy in multi-category ABSA, we propose a MuSEM framework. This framework divides our twelve-category ABSA task into multiple sub-tasks, with each sub-task framework corresponding to a pre-trained EM. Furthermore, different tasks can provide complementary feature support to one another, thereby enhancing the accuracy of complex ABSA tasks. As illustrated in Figure 6, our framework comprises three distinct tasks, detailed in Figure 3. Patient reviews serve as the initial input to the framework. In practice, each EM independently tokenizes and embeds its own task-specific input using its native tokenizer and embedding layer before the resulting hidden representations are processed by the corresponding decoder blocks. Subsequent operations, including layer norm and masked self-attention, resemble those in traditional MoE frameworks.

In MuSEM, each EM processes its task-specific input using its own native tokenizer, vocabulary, embedding layer, and decoder architecture. For Task I, EM-1 independently encodes the original patient review and produces both the ACD-1 prediction and its intermediate hidden-state representation, which is retained as Task I Features (T1F). For Task II, EM-2 encodes the Task-II input, which consists of the original review together with the structured ACD-1 prediction from Task I, and generates the current Task-II hidden states. Similarly, for Task III, EM-3 encodes the Task-III input based on the review and the structured outputs from previous tasks.

Therefore, MuSEM does not directly feed token embeddings from EM-1 into EM-2 or EM-3, nor does it require different experts to share the same tokenizer or vocabulary. Cross-expert interaction is performed only at the hidden-state level. Since different expert models may have different native hidden dimensions and produce sequences of different lengths, we introduce task-specific projection adapters before applying DCA and TCA. These adapters map heterogeneous hidden-state representations into a shared latent attention space, enabling dimensionally compatible cross-task attention without requiring token-level alignment across different expert models. This design preserves the independence of each expert model while allowing Task II and Task III to retrieve useful information from earlier task-level representations.

The distinction lies in our adoption of a Specialized Router (SR) within the router component. Having trained distinct EM through prior experimentation, each task corresponds to a specific EM. Consequently, our SR requires no additional training, solely responsible for selecting the appropriate prompt according to a predefined task sequence. This eliminates the need for the gating parameters used in traditional MoE routers to select experts, thereby avoiding

potential errors in EM selection inherent to that step. In summary, for Task I, our SR selects the corresponding Prompt-I and EM-1, subsequently yielding Task I's output I. Here, Prompt-I = Prompt-ACD-1 + CoT + Few-shot. For further details regarding the prompt, please refer to our supplementary materials.

Unlike running the three tasks as fully isolated pipelines, MuSEM establishes a feature-level collaboration mechanism among heterogeneous expert models. In this framework, each Expert Model (EM) independently processes its task-specific input using its own native tokenizer, embedding layer, and decoder architecture. The symbolic output produced by an earlier task is used as structured guidance for the subsequent task, while the intermediate hidden-state representation generated by that task is retained as task-level features for cross-task information transfer.

Specifically, after Task I, the predicted ACD-1 labels provide structured aspect-level guidance for Task II. Meanwhile, the hidden-state sequence generated by EM-1 is preserved as Task I Features (T1F). These features offer coarse-grained aspect information and help subsequent tasks with aspect localisation and semantic constraint. Similarly, after Task II, the hidden-state sequence generated by EM-2 is retained as Task II Features (T2F), which provide more fine-grained sub-aspect information for Task III. In this way, MuSEM enables cross-task knowledge transfer through hidden-state feature sharing rather than by directly reusing raw token embeddings or decoder states across different expert models.

Since the expert models may be initialized from different base models, they may have different tokenizers, vocabularies, embedding layers, hidden dimensions, and sequence lengths. Therefore, DCA and TCA do not operate on raw token embeddings from different models. Instead, cross-task interaction is performed only on task-level hidden representations after feature projection and alignment. Before applying DCA and TCA, the hidden representations from different EMs are mapped into a shared latent attention space through task-specific projection adapters, ensuring dimensional compatibility while avoiding the need for token-level alignment across heterogeneous expert models.

Let d_1, d_2, and d_3 denote the native hidden dimensions of EM-1, EM-2, and EM-3, respectively, while d_c denotes the shared latent dimension used inside the cross-attention modules. Before projection, the task-level representations are defined as $\mathrm{T1F} \in \mathbb{R}^{n\times d_1}$, $\mathrm{T2F} \in \mathbb{R}^{p\times d_2}$, $H_2 \in \mathbb{R}^{m\times d_2}$, and $H_3 \in \mathbb{R}^{q\times d_3}$, where n, p, m, and q denote the corresponding sequence lengths. Since the expert models may be initialized from different base models, these hidden dimensions and sequence lengths are not required to be identical. Before DCA and TCA are applied, all representations involved in cross-task attention are mapped into a common latent attention space with dimension d_c.

In our implementation, the selected EM-1, EM-2, and EM-3 all have a native hidden dimension of 4096 ($d_1 = d_2 = d_3 = 4096$). However, although these expert models have the same hidden dimensionality, their representation spaces are not assumed to be directly aligned, since they are initialized from different base models and further adapted for different subtasks. Therefore, direct cross-attention over their native hidden states may introduce semantic mismatch and unnecessary computational cost. To address this issue, we introduce task-specific projection adapters to map the hidden-state representations from different experts into a shared latent attention space. In our experiments, the shared latent dimension is set to $d_c = 1024$. This setting provides a compact cross-expert alignment space compared with the native 4096-dimensional hidden representations, reducing the computational cost of DCA and TCA while retaining sufficient capacity for cross-task semantic retrieval. Accordingly, $\mathrm{T1F} \in \mathbb{R}^{n\times 4096}$, $\mathrm{T2F} \in \mathbb{R}^{p\times 4096}$, $H_2 \in \mathbb{R}^{m\times 4096}$, and $H_3 \in \mathbb{R}^{q\times 4096}$ are first projected into $\mathbb{R}^{\cdot\times 1024}$ before cross-attention is computed. The resulting attention outputs are then projected back to the native hidden dimensions of the corresponding expert models for residual fusion. For notational simplicity, Equations (12)–(18) describe the attention operation using a single-head formulation. In practice, both the native Self-Attention layers and the proposed DCA module are implemented using standard multi-head attention, where multiple attention heads operate in parallel and their outputs are subsequently aggregated.

As shown in Equation 12, let $\mathrm{Attn}(Q, K, V)$ denote the standard scaled dot-product attention function computed in the common latent space:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_c}}\right)V, \tag{12}$$

where Q, K, and V represent the Query, Key, and Value matrices, respectively, and d_c is the dimensionality of the projected attention space.

To better leverage the information generated by Task I, we propose a Dual Cross-Attention (DCA) module. DCA allows Task II to retrieve relevant semantic information from T1F after its own masked self-attention operation. The module does not modify the internal computation of Task II self-attention itself. Instead, it introduces an auxiliary cross-task retrieval branch, whose output is subsequently fused with the original Task II representation through a residual connection. In this way, Task II can incorporate coarse-grained aspect information from Task I while preserving its own task-specific representation learning process.

For the DCA module, the intermediate representation $H_2 \in \mathbb{R}^{m\times d_2}$ from Decoder Block II is used as the Query source, while the Task I feature $\mathrm{T1F} \in \mathbb{R}^{n\times d_1}$ is used as the Key and Value source. To align their hidden dimensions, we introduce task-specific projection matrices before cross-attention:

$$\begin{aligned} \tilde{H}_2 &= \mathrm{LN}(H_2 W_2^{Q}), \\ \tilde{\mathrm{T1F}}^{K} &= \mathrm{LN}(\mathrm{T1F} W_1^{K}), \\ \tilde{\mathrm{T1F}}^{V} &= \mathrm{LN}(\mathrm{T1F} W_1^{V}), \end{aligned} \tag{13}$$

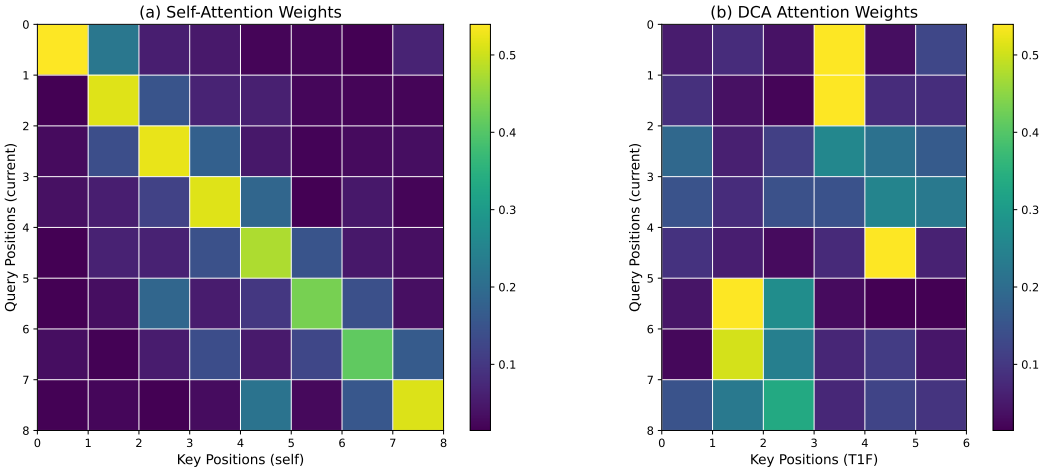

Figure 8: Visualization of attention weight distributions in Self-Attention and DCA modules from the fine-tuned EM-2. (a) Self-Attention weights (8×8) illustrate intra-task dependencies within the current sequence. (b) DCA attention weights (8×6) illustrate cross-task semantic retrieval between query tokens from the current task and key features from T1F.

where $W_2^Q \in \mathbb{R}^{d_2 \times d_c}$, and $W_1^K, W_1^V \in \mathbb{R}^{d_1 \times d_c}$. Therefore, $\tilde{H}_2$, $\tilde{\text{T1F}}^K$, and $\tilde{\text{T1F}}^V$ are mapped into the same latent dimension d_c, with shapes $\mathbb{R}^{m \times d_c}$, $\mathbb{R}^{n \times d_c}$, and $\mathbb{R}^{n \times d_c}$, respectively. The resulting attention score matrix has shape $\mathbb{R}^{m \times n}$, which allows each Task II hidden-state position to softly retrieve information from all Task I feature positions.

Although EM-1, EM-2, and EM-3 share the same native hidden dimensionality in our implementation ($d_1 = d_2 = d_3 = 4096$), their representation spaces are not assumed to be directly aligned because the experts are initialized from different base models and further adapted for different subtasks. Therefore, each expert's hidden representations are independently projected into a shared latent attention space with dimensionality $d_c = 1024$ before cross-attention is applied. Furthermore, the sequence lengths (n, p, m, q) may differ due to the use of heterogeneous tokenizers and task prompts. However, the proposed DCA and TCA mechanisms operate through soft semantic attention rather than token-level correspondence, allowing cross-task information retrieval without requiring aligned token positions across different expert models. The DCA operation is then defined as:

$$\text{DCA}(H_2, \text{T1F}) = \text{softmax}\left(\frac{\tilde{H}_2(\tilde{\text{T1F}}^K)^{\text{T}}}{\sqrt{d_c}}\right)\tilde{\text{T1F}}^V. \quad (14)$$

The output of the DCA module has the shape $\mathbb{R}^{m \times d_c}$. It is then projected back to the native hidden dimension of EM-2 before being added to the original representation H_2 through a residual connection:

$$H_2' = H_2 + \text{DCA}(H_2, \text{T1F})W_2^O, \quad (15)$$

where $W_2^O \in \mathbb{R}^{d_c \times d_2}$. This output projection ensures that the residual fusion is dimensionally valid.

After Task II, the symbolic ACD-2 prediction is used as structured sub-aspect guidance for Task III, while the hidden-state sequence generated by EM-2 is retained as Task II Features (T2F). Task III differs from Task II because it requires information from both previous stages: T1F provides coarse-grained primary-aspect information, whereas T2F provides fine-grained sub-aspect information. Therefore, we propose a Triple Cross-Attention (TCA) module to jointly retrieve information from T1F and T2F. Similar to DCA, TCA is applied after the masked self-attention operation of Task III and its retrieved cross-task information is fused with the current Task III representation through residual connection.

For the TCA module, the current representation $H_3 \in \mathbb{R}^{q \times d_3}$ is used as the Query source, while $\text{T1F} \in \mathbb{R}^{n \times d_1}$ and $\text{T2F} \in \mathbb{R}^{p \times d_2}$ are used to construct the Key and Value streams. These representations are first projected into the shared latent dimension d_c:

$$\begin{aligned}
\tilde{H}_3 &= \text{LN}(H_3 W_3^Q), \\
\tilde{\text{T1F}}^K &= \text{LN}(\text{T1F}W_1^K), \\
\tilde{\text{T1F}}^V &= \text{LN}(\text{T1F}W_1^V), \\
\tilde{\text{T2F}}^K &= \text{LN}(\text{T2F}W_2^K), \\
\tilde{\text{T2F}}^V &= \text{LN}(\text{T2F}W_2^V),
\end{aligned} \quad (16)$$

where $W_3^Q \in \mathbb{R}^{d_3 \times d_c}$, $W_1^K, W_1^V \in \mathbb{R}^{d_1 \times d_c}$, and $W_2^K, W_2^V \in \mathbb{R}^{d_2 \times d_c}$. Thus, $\tilde{H}_3 \in \mathbb{R}^{q \times d_c}$, $\tilde{\text{T1F}}^K, \tilde{\text{T1F}}^V \in \mathbb{R}^{n \times d_c}$, and $\tilde{\text{T2F}}^K, \tilde{\text{T2F}}^V \in \mathbb{R}^{p \times d_c}$. The Key and Value matrices for TCA are then constructed by concatenating the projected features from Task I and Task II along the sequence dimension:

$$\begin{aligned}
K_{\text{all}} &= \text{Concat}(\tilde{\text{T1F}}^K, \tilde{\text{T2F}}^K), \\
V_{\text{all}} &= \text{Concat}(\tilde{\text{T1F}}^V, \tilde{\text{T2F}}^V),
\end{aligned} \quad (17)$$

where $K_{\text{all}}, V_{\text{all}} \in \mathbb{R}^{(n+p) \times d_c}$. The resulting attention score matrix has shape $\mathbb{R}^{q \times (n+p)}$, enabling each Task III hidden-state position to retrieve information from both Task I and

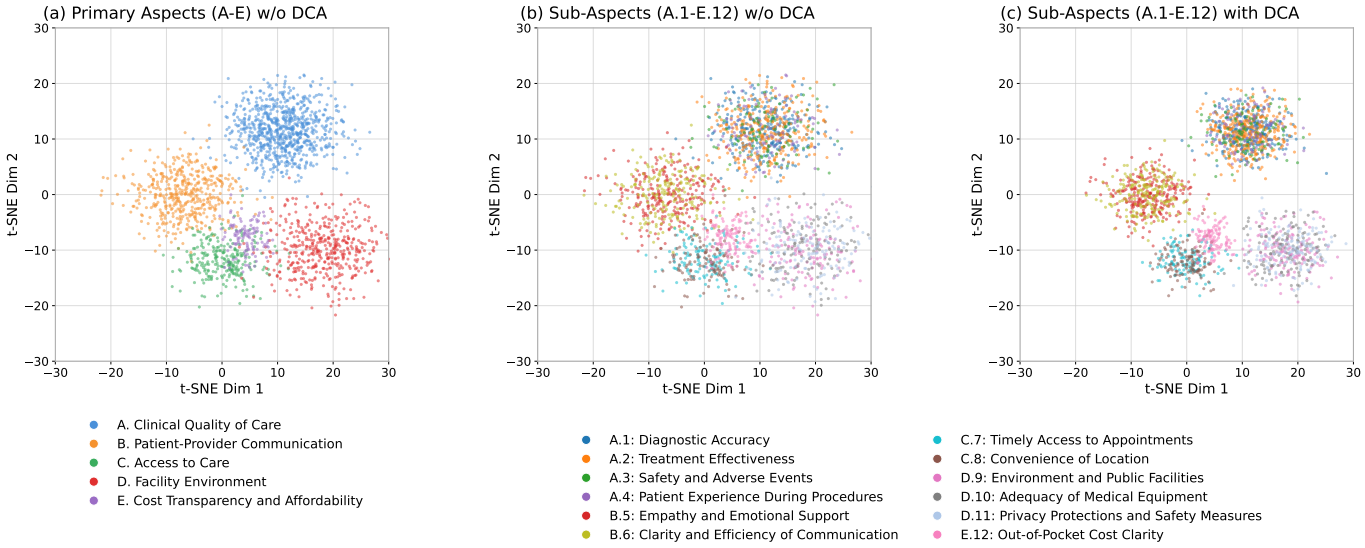

Figure 9: T-SNE visualization of hierarchical aspect representations before and after DCA fusion. (a) Primary Aspects (A-E) w/o DCA; (b) Sub-Aspects (A.1-E.12) w/o DCA. (c) Sub-Aspects (A.1-E.12) with DCA.

Task II feature sequences. The TCA operation is then defined as:

$$\mathrm{TCA}(H_3, \mathrm{T1F}, \mathrm{T2F}) = \mathrm{softmax}\left(\frac{\tilde{H}_3 K_{\mathrm{all}}^{\mathrm{T}}}{\sqrt{d_c}}\right) V_{\mathrm{all}}. \quad (18)$$

The output of TCA has the shape $\mathbb{R}^{q \times d_c}$. It is projected back to the native hidden dimension of EM-3 before residual fusion:

$$H_3' = H_3 + \mathrm{TCA}(H_3, \mathrm{T1F}, \mathrm{T2F}) W_3^O, \quad (19)$$

where $W_3^O \in \mathbb{R}^{d_c \times d_3}$. Although the sequence lengths n, p, m, and q may differ because different expert models use different tokenizers and task prompts, cross-attention does not require one-to-one token alignment. Instead, it performs soft semantic retrieval over prior-task feature sequences. Therefore, the projection-and-alignment procedure ensures that the cross-attention operations in DCA and TCA are dimensionally valid while preserving the intended cross-task knowledge transfer mechanism.

This design ensures that the primary information flow within each decoder block remains uncompromised. The cross-attention modules serve as supplemental mechanisms, enabling the model to intentionally "refer back" to the structured predictions and features from earlier stages, thereby enhancing coherence and accuracy in complex, hierarchical reasoning tasks. Moreover, the SR component differs somewhat from previous iterations. As the EM for both the ACSA and IR tasks is identical, Task III effectively encompasses two distinct tasks. Consequently, within Task III's SR, we have arranged tasks for both objectives, though they point to the same EM. The EM processes prompts sequentially, meaning the ACSA and IR tasks run concurrently yet independently. Upon task completion, all outputs are consolidated and processed through the decoder block before being sent to the final output.

To enable effective interaction across heterogeneous expert models, we perform feature-level alignment rather than directly sharing raw token embeddings. Specifically, each expert model independently processes the input using its native tokenizer and architecture, producing intermediate hidden representations. These representations capture higher-level semantic information and are therefore more comparable across models despite differences in tokenization and embedding spaces. The proposed cross-attention modules operate on these hidden representations instead of raw token-level inputs. This design avoids the need for strict alignment of vocabularies or embedding dimensions. In practice, the attention mechanism is applied to semantically-meaningful features, which allows information exchange across experts without requiring architectural homogeneity. This feature-level interaction enables the model to leverage complementary knowledge from different expert models while maintaining compatibility across heterogeneous representations.

To further analyze the operational characteristics of the proposed modules, we visualize the attention weight distributions in Figure 8. Each heatmap illustrates how the current-task query representations allocate attention to the corresponding key features. The attention maps are extracted from the final transformer block of the EM-2 and averaged across all attention heads using a randomly selected validation sample. The attention maps shown in Figure 8 are visualized by averaging the attention weights across all attention heads. Specifically, Self-Attention (Figure 8a) produces an 8×8 attention matrix within the same task domain. The prominent diagonal structure indicates that each query position primarily attends to its own representation, reflecting strong self-alignment and local contextual preservation. Meanwhile, several off-diagonal activations can also be observed in neighboring and semantically related

regions, suggesting that the model additionally captures contextual dependencies and semantic interactions among tokens within the current task.

In contrast, DCA (Figure 8b) generates an 8×6 cross-domain attention matrix, where the eight query positions from the current task interact with six transferred feature representations (T1F) derived from Task I. Unlike Self-Attention, the DCA map does not exhibit strong diagonal dominance because the query and key representations originate from different semantic spaces and hierarchical subtasks. Instead, the attention distribution forms several sparse yet concentrated activation regions, indicating selective retrieval of task-relevant prior knowledge. In particular, several query positions strongly attend to specific transferred features, while unrelated regions maintain relatively low attention weights. This behavior demonstrates that the DCA mechanism effectively performs focused cross-task semantic alignment rather than indiscriminate feature aggregation. Such selective sparsity further reflects the expert-oriented design of the proposed framework, where only task-relevant prior knowledge is activated during cross-task interaction.

Building upon the attention mechanism analysis above, we further visualize the representation distributions learned by the proposed DCA modules using t-SNE. Since the TCA module operates only in the downstream ACSA and IR tasks, whereas the sub-aspect representations visualized in Figure 9 are generated by EM-2, the t-SNE analysis focuses exclusively on the effect of the DCA module. Therefore, Figure 9 compares the representation distributions before and after the introduction of DCA.

To better reflect the fine-grained hierarchical classification task, the visualization is generated from the final-layer hidden representations of EM-2, i.e., the ACD-2 expert model responsible for identifying the twelve sub-aspect categories. The representations are extracted from the held-out test set. For each input, sequence-level embeddings are obtained by mean pooling over non-padding token hidden states. For multi-label samples, each aspect-specific instance is treated as a separate point according to its corresponding sub-aspect label. Figure 9 illustrates the embedding structures under different hierarchical settings and feature fusion strategies. Specifically, Figure 9(a) presents the coarse-grained distributions of the five primary aspects (A–E), where the major semantic categories form distinguishable yet partially overlapping clusters. Figure 9(b) further visualizes the fine-grained sub-aspect representations (A.1–E.12) before applying DCA, showing that several sub-aspect categories exhibit noticeable overlap and fragmented boundaries due to limited cross-task semantic interaction.

After incorporating the proposed DCA mechanisms, Figure 9(c) demonstrates an improved embedding structure for the 12 sub-aspect categories. Compared with the pre-fusion distributions, the post-fusion embeddings become more compact within the same class while exhibiting clearer inter-class separation across semantically related sub-aspects. This indicates that the cross-task attention mechanisms effectively enhance hierarchical semantic transfer and reduce representation ambiguity among fine-grained medical aspects. In particular, the improved cluster compactness suggests stronger intra-class semantic consistency, whereas the enlarged margins between clusters reflect better discriminative capability in the shared embedding space. These visualizations intuitively demonstrate that the dual- and triple-cross-attention mechanisms not only expand the receptive field at the attention level but also promote tighter intra-class cohesion and clearer inter-class boundaries in the shared embedding space.

Table 3
Results of the ACD-1 Task for Model Screening

Model	$F1_{\text{micro}}$	$F1_{\text{macro}}$
Apollo-7B	0.7483 ± 0.019	0.7276 ± 0.024
Gemma-2-9B-it	0.6995 ± 0.029	0.6758 ± 0.034
GLM-Z1-9B-0414	0.6716 ± 0.021	0.6529 ± 0.027
InternLM2.5-Chat-7B	0.5729 ± 0.033	0.5441 ± 0.035
Llama3.1-8B-Instruct	0.6610 ± 0.025	0.6374 ± 0.031
Ministral-8B-Instruct-2410	0.5762 ± 0.048	0.5638 ± 0.052
Phi-3-small-7B	0.6832 ± 0.023	0.6591 ± 0.029
Qwen3-8B	0.7054 ± 0.017	0.6913 ± 0.027
Starling-7B-LM-Beta	0.4998 ± 0.061	0.4832 ± 0.066
Yi-1.5-9B-Chat	0.6175 ± 0.032	0.5983 ± 0.037

Table 4
Results of the ACD-2 Task for Model Screening

Model	$F1_{\text{micro}}$	$F1_{\text{macro}}$
Apollo-7B	0.6631 ± 0.021	0.6349 ± 0.034
Gemma-2-9B-it	0.5888 ± 0.035	0.5632 ± 0.041
GLM-Z1-9B-0414	0.6320 ± 0.023	0.6055 ± 0.029
InternLM2.5-Chat-7B	0.5672 ± 0.045	0.5324 ± 0.052
Llama3.1-8B-Instruct	0.6263 ± 0.031	0.5978 ± 0.037
Ministral-8B-Instruct-2410	0.5801 ± 0.056	0.5704 ± 0.061
Phi-3-small-7B	0.6729 ± 0.028	0.6647 ± 0.025
Qwen3-8B	0.7009 ± 0.018	0.6926 ± 0.026
Starling-7B-LM-Beta	0.4937 ± 0.049	0.4692 ± 0.063
Yi-1.5-9B-Chat	0.5942 ± 0.032	0.5715 ± 0.038

4. Experiments

In this study, our experiments were primarily divided into two aspect categories: testing experiments and fine-tuning experiments, each with distinct environmental configurations. For the testing experiments, to replicate real-world medical scenarios and enhance the project's deployability, we selected the mainstream consumer-grade NVIDIA RTX 4090 GPU with 24GB of VRAM, using a single unit. For the fine-tuning experiments, we employed a single NVIDIA A100 GPU with 80GB of VRAM. Additionally, all hyperparameters for the LLMs were set to their default values without further modification. This approach ensured fair comparisons and demonstrated the generalizability of our experimental results. On the other hand, each of our experiments was run five times, yielding one $F1_{micro}$ and one $F1_{macro}$ score per run. The final results presented are

the average of the five $F1_{micro}$ and $F1_{macro}$. Furthermore, to enable a more precise assessment of model stability, we computed the sample standard deviation (σ) across five independent runs. A smaller sample std reflects greater consistency and robustness in model performance. At the same time, our fine-tuning dataset and test dataset are independent and non-overlapping, thus eliminating the possibility of data leakage. Details of the prompts used in all experiments are provided in our supplementary materials. In this section, we present the experimental setup and evaluation results, including comparisons with baseline methods and ablation studies.

4.1. ACD-1 Task Initial Screening

To screen EMs for different tasks, we referenced four distinct LLM leaderboards (HuggingFace LLM Leaderboard, LMSYS Leaderboard, OpenCompass, and LLM Leaderboard stats)[73], [74], [75]. We selected ten relatively high-performing and popular open-source models with parameters below 10 billion for experimentation. We deployed the different models on the same server and compared them using identical prompts, Few-shot training, and CoT instructions. As shown in Table 3, Apollo-7B achieved the best performance on the ACD-1 task for both $F1_{micro}$ (0.75) and $F1_{macro}$ (0.72), making it the EMC for EM-1. Qwen3-8B and Gemma2-9B-it ranked second and third, achieving $F1_{micro}$ and $F1_{macro}$ scores of 0.71 and 0.69, and 0.70 and 0.67, respectively. Additionally, Phi-3-small 7B, GLM-Z1-9B-0414, and Llama 3.1-8B-Instruct all exceeded 0.5 on both $F1_{micro}$ and $F1_{macro}$.

4.2. ACD-2 Task Initial Screening

As shown in Table 4, Qwen3-8B outperformed Apollo-7B in the ACD-2 task, achieving the highest performance with $F1_{micro}$ and $F1_{macro}$ scores of 0.70 and 0.69, respectively. Therefore, Qwen3-8B was selected as the EMC for EM-2. Apollo-7B followed closely behind, achieving $F1_{micro}$ and $F1_{macro}$ scores of 0.66 and 0.63, respectively. Both GLM-Z1-9B-0414 and Llama3.1-8B-Instruct achieved an $F1_{micro}$ score of 0.63, but their $F1_{macro}$ scores were only 0.60 and 0.59. Compared to the ACD-1 task, only Ministral-8B-Instruct-2410 showed improvement, with $F1_{micro}$ and $F1_{macro}$ increasing slightly from 0.58 and 0.56 to 0.58 and 0.57, respectively.

4.3. ACSA Task Initial Screening

In the ACSA task, as shown in Table 5, Phi-3-small 7B achieved the optimal performance with $F1_{micro}$ and $F1_{macro}$ scores of 0.81 and 0.76, respectively. Therefore, Phi-3-small 7B was selected as the EMC for EM-3. Additionally, Qwen3-8B demonstrated strong performance on this task, achieving an $F1_{micro}$ score of 0.80 and an $F1_{macro}$ score of 0.72, both of which ranked just behind Phi-3-small 7B. On the other hand, all models in this task achieved a score of over 60.00 in both $F1_{micro}$ and $F1_{macro}$.

Table 5
Results of the ACSA Task for Model Screening

Model	$F1_{micro}$	$F1_{macro}$
Apollo-7B	0.7724 ± 0.027	0.7138 ± 0.034
Gemma-2-9B-it	0.7831 ± 0.049	0.7306 ± 0.058
GLM-Z1-9B-0414	0.7559 ± 0.031	0.6719 ± 0.036
InternLM2.5-Chat-7B	0.7225 ± 0.039	0.6085 ± 0.046
Llama3.1-8B-Instruct	0.7338 ± 0.033	0.6214 ± 0.039
Ministral-8B-Instruct-2410	0.6870 ± 0.052	0.6157 ± 0.058
Phi-3-small-7B	0.8144 ± 0.028	0.7618 ± 0.032
Qwen3-8B	0.8036 ± 0.021	0.7285 ± 0.027
Starling-7B-LM-Beta	0.6591 ± 0.066	0.6292 ± 0.072
Yi-1.5-9B-Chat	0.7602 ± 0.028	0.6424 ± 0.033

Table 6
Results of the IR Task for Model Screening

Model	$F1_{micro}$	$F1_{macro}$
Apollo-7B	0.8097 ± 0.032	0.7721 ± 0.048
Gemma-2-9B-it	0.7495 ± 0.029	0.6883 ± 0.034
GLM-Z1-9B-0414	0.7256 ± 0.024	0.7134 ± 0.031
InternLM2.5-Chat-7B	0.7193 ± 0.038	0.6719 ± 0.043
Llama3.1-8B-Instruct	0.7380 ± 0.027	0.6718 ± 0.032
Ministral-8B-Instruct-2410	0.7361 ± 0.049	0.7063 ± 0.056
Phi-3-small-7B	0.8396 ± 0.020	0.8291 ± 0.026
Qwen3-8B	0.8250 ± 0.019	0.8013 ± 0.024
Starling-7B-LM-Beta	0.7415 ± 0.055	0.7281 ± 0.061
Yi-1.5-9B-Chat	0.7144 ± 0.031	0.6629 ± 0.037

Table 7
Manual evaluation of LRS-generated data quality.

Criterion	Score (%)	Agreement (κ)
Label Correctness	92.3	0.83
Semantic Consistency	89.7	0.78
Linguistic Naturalness	91.5	0.80

4.4. IR Task Initial Screening

As shown in Table 6, Phi-3-small 7B also achieved the best performance in the IR task, with both $F1_{micro}$ and $F1_{macro}$ exceeding 0.8000, at 0.8396 and 0.8283, respectively. Therefore, Phi-3-small 7B should be selected as the EMC for EM-4. However, since Phi-3-small 7B was previously chosen as the EMC for EM-3 in earlier tasks, we can retain the original naming convention and have this EM3 complete both tasks. Additionally, Qwen3-8B and Apollo-7B also exhibited competitive performance, achieving $F1_{micro}$ scores of 0.83 and 0.81, respectively, while their $F1_{macro}$ scores were 0.80 and 0.76, respectively. Finally, in this task, all models achieved $F1_{micro}$ and $F1_{macro}$ scores exceeding 0.6.

4.5. Expert Model Fine-tuning

To demonstrate that the DoRA-based fine-tuning approach is more suitable for our H-ABSA task, we compared DoRA and LoRA using three EMC experiments. We divided LoRA and DoRA into two control groups: one fine-tuned solely on the manually annotated training set (H_{1000}), and

Table 8
Comparison of LoRA and DoRA Fine-tuning Approaches

Model	ACD-1 Task		ACD-2 Task		ACSA Task		IR Task	
	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
Apollo-7B	0.7483	0.7205	0.6631	0.6266	0.7724	0.7040	0.8097	0.7631
+LoRA (H1000)	0.7882↑ 0.04	0.7524↑ 0.03	0.7173↑ 0.05	0.6841↑ 0.06	0.8334↑ 0.06	0.7318↑ 0.03	0.8360↑ 0.03	0.7957↑ 0.03
+LoRA (H1000+L3000)	0.8499↑ 0.10	0.8160↑ 0.10	0.7555↑ 0.09	0.6706↑ 0.04	0.8376↑ 0.07	0.7629↑ 0.06	0.8801↑ 0.07	0.8364↑ 0.07
+DoRA (H1000)	0.7945↑ 0.05	0.7538↑ 0.03	0.7205↑ 0.06	0.6932↑ 0.07	0.8494↑ 0.08	0.7320↑ 0.03	0.8363↑ 0.03	0.8083↑ 0.05
+DoRA (H1000+L3000)	0.8515↑ 0.10	0.8233↑ 0.10	0.7663↑ 0.10	0.7219↑ 0.10	0.8197↑ 0.05	0.7754↑ 0.07	0.8766↑ 0.07	0.8408↑ 0.08
Phi-3-small-7B	0.6832	0.6539	0.6729	0.6670	0.8144	0.7589	0.8396	0.8283
+LoRA (H1000)	0.7033↑ 0.02	0.6778↑ 0.02	0.7396↑ 0.07	0.7105↑ 0.04	0.8314↑ 0.02	0.8253↑ 0.07	0.8926↑ 0.05	0.8561↑ 0.03
+LoRA (H1000+L3000)	0.7446↑ 0.06	0.7025↑ 0.05	0.7431↑ 0.07	0.7268↑ 0.06	0.8777↑ 0.06	0.8790↑ 0.12	0.9139↑ 0.07	0.8818↑ 0.05
+DoRA (H1000)	0.7076↑ 0.02	0.6803↑ 0.03	0.7447↑ 0.07	0.7310↑ 0.06	0.8584↑ 0.04	0.8231↑ 0.06	0.8858↑ 0.05	0.8622↑ 0.03
+DoRA (H1000+L3000)	0.7486↑ 0.07	0.7083↑ 0.05	0.7747↑ 0.10	0.7503↑ 0.08	0.9036↑ 0.09	0.8837↑ 0.12	0.9229↑ 0.08	0.9023↑ 0.07
Qwen3-8B	0.7054	0.6861	0.7096	0.6941	0.8036	0.7208	0.8250	0.7962
+LoRA (H1000)	0.7307↑ 0.03	0.6989↑ 0.01	0.7756↑ 0.07	0.7674↑ 0.07	0.8591↑ 0.06	0.8164↑ 0.10	0.8358↑ 0.01	0.8622↑ 0.07
+LoRA (H1000+L3000)	0.7742↑ 0.07	0.7313↑ 0.05	0.7900↑ 0.08	0.7728↑ 0.08	0.8935↑ 0.09	0.8209↑ 0.10	0.8981↑ 0.07	0.8847↑ 0.09
+DoRA (H1000)	0.7593↑ 0.05	0.7006↑ 0.01	0.7749↑ 0.07	0.7682↑ 0.07	0.8427↑ 0.04	0.7954↑ 0.07	0.8411↑ 0.02	0.8568↑ 0.06
+DoRA (H1000+L3000)	0.8093↑ 0.10	0.7326↑ 0.05	0.8242↑ 0.11	0.8028↑ 0.11	0.8806↑ 0.08	0.8010↑ 0.08	0.8925↑ 0.07	0.8843↑ 0.09

Table 9
Comparison of Effects Between Two Distinct Fine-tuning Approaches

Model	ACD-1 Task		ACD-2 Task		ACSA Task		IR Task	
	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
Apollo-7B	0.7483	0.7205	0.6631	0.6266	0.7724	0.7040	0.8097	0.7631
+DoRA (H_{1000})	0.7945↑ 0.05	0.7538↑ 0.03	0.7205↑ 0.06	0.6932↑ 0.07	0.8494↑ 0.08	0.7320↑ 0.03	0.8363↑ 0.03	0.8083↑ 0.05
+DoRA (H_{1000}+L_{3000})	0.8515↑ 0.10	0.8233↑ 0.10	0.7663↑ 0.10	0.7219↑ 0.10	0.8197↑ 0.05	0.7754↑ 0.07	0.8766↑ 0.07	0.8408↑ 0.08
Gemma-2-9B-it	0.6995	0.6687	0.5888	0.5603	0.7831	0.7223	0.7495	0.6786
+DoRA (H_{1000})	0.7759↑ 0.08	0.7096↑ 0.04	0.6741↑ 0.09	0.6014↑ 0.04	0.8067↑ 0.02	0.7060↑ 0.04	0.8048↑ 0.06	0.7185↑ 0.04
+DoRA (H_{1000}+L_{3000})	0.7876↑ 0.09	0.7207↑ 0.05	0.7198↑ 0.13	0.6752↑ 0.12	0.8201↑ 0.04	0.7917↑ 0.07	0.8448↑ 0.10	0.7568↑ 0.08
GLM-Z1-9B-0414	0.6716	0.6521	0.6320	0.5997	0.7559	0.6657	0.7256	0.7102
+DoRA (H_{1000})	0.7489↑ 0.08	0.6901↑ 0.04	0.6946↑ 0.06	0.6433↑ 0.04	0.7877↑ 0.03	0.7012↑ 0.04	0.7958↑ 0.07	0.7594↑ 0.05
+DoRA (H_{1000}+L_{3000})	0.7628↑ 0.09	0.7204↑ 0.07	0.7565↑ 0.12	0.6520↑ 0.05	0.7948↑ 0.04	0.7329↑ 0.07	0.8412↑ 0.12	0.7936↑ 0.08
InternLM2.5-Chat-7B	0.5729	0.5401	0.5672	0.5258	0.7225	0.6018	0.7193	0.6645
+DoRA (H_{1000})	0.6637↑ 0.09	0.5880↑ 0.05	0.6442↑ 0.08	0.6051↑ 0.08	0.7695↑ 0.05	0.6960↑ 0.09	0.7672↑ 0.05	0.7119↑ 0.05
+DoRA (H_{1000}+L_{3000})	0.6828↑ 0.11	0.6599↑ 0.12	0.6928↑ 0.13	0.6943↑ 0.17	0.8053↑ 0.08	0.6913↑ 0.09	0.8093↑ 0.09	0.7537↑ 0.09
Llama3.1-8B-Instruct	0.6610	0.6317	0.6263	0.5881	0.7338	0.6141	0.7380	0.6638
+DoRA (H_{1000})	0.6921↑ 0.03	0.6713↑ 0.04	0.7048↑ 0.08	0.6450↑ 0.06	0.8031↑ 0.07	0.6523↑ 0.04	0.7927↑ 0.05	0.7732↑ 0.11
+DoRA (H_{1000}+L_{3000})	0.7433↑ 0.08	0.7062↑ 0.07	0.7465↑ 0.12	0.6772↑ 0.09	0.8853↑ 0.15	0.6819↑ 0.07	0.8545↑ 0.12	0.8072↑ 0.14
Ministral-8B-Instruct-2410	0.5762	0.5633	0.5801	0.5695	0.6870	0.6092	0.7361	0.7041
+DoRA (H_{1000})	0.6117↑ 0.04	0.6052↑ 0.04	0.7134↑ 0.13	0.6054↑ 0.04	0.7661↑ 0.08	0.6728↑ 0.06	0.7975↑ 0.06	0.8040↑ 0.10
+DoRA (H_{1000}+L_{3000})	0.6218↑ 0.05	0.7278↑ 0.16	0.6510↑ 0.07	0.7659↑ 0.20	0.7159↑ 0.03	0.6817↑ 0.07	0.8349↑ 0.10	0.8294↑ 0.13
Phi-3-small-7B	0.6832	0.6539	0.6729	0.6670	0.8144	0.7589	0.8396	0.8283
+DoRA (H_{1000})	0.7076↑ 0.02	0.6803↑ 0.03	0.7447↑ 0.07	0.7310↑ 0.06	0.8584↑ 0.04	0.8231↑ 0.06	0.8858↑ 0.05	0.8622↑ 0.03
+DoRA (H_{1000}+L_{3000})	0.7486↑ 0.07	0.7083↑ 0.05	0.7747↑ 0.10	0.7503↑ 0.08	0.9036↑ 0.09	0.8837↑ 0.12	0.9229↑ 0.08	0.9023↑ 0.07
Qwen3-8B	0.7054	0.6861	0.7096	0.6941	0.8036	0.7208	0.8250	0.7962
+DoRA (H_{1000})	0.7593↑ 0.05	0.7006↑ 0.01	0.7749↑ 0.07	0.7682↑ 0.07	0.8427↑ 0.04	0.7954↑ 0.07	0.8411↑ 0.02	0.8568↑ 0.06
+DoRA (H_{1000}+L_{3000})	0.8093↑ 0.10	0.7326↑ 0.05	0.8242↑ 0.11	0.8028↑ 0.11	0.8806↑ 0.08	0.8010↑ 0.08	0.8925↑ 0.07	0.8843↑ 0.09
Starling-7B-LM-Beta	0.4998	0.4785	0.4937	0.4629	0.6591	0.6235	0.7415	0.7244
+DoRA (H_{1000})	0.5710↑ 0.07	0.5736↑ 0.10	0.5183↑ 0.02	0.5057↑ 0.04	0.7274↑ 0.07	0.6521↑ 0.03	0.7856↑ 0.04	0.7702↑ 0.05
+DoRA (H_{1000}+L_{3000})	0.5933↑ 0.09	0.5947↑ 0.12	0.5824↑ 0.09	0.5752↑ 0.11	0.7387↑ 0.08	0.6980↑ 0.07	0.8326↑ 0.09	0.7528↑ 0.03
Yi-1.5-9B-Chat	0.6175	0.5968	0.5942	0.5654	0.7602	0.6353	0.7144	0.6549
+DoRA (H_{1000})	0.6415↑ 0.02	0.6372↑ 0.04	0.6238↑ 0.03	0.6190↑ 0.05	0.8104↑ 0.05	0.7057↑ 0.07	0.7362↑ 0.02	0.7026↑ 0.05
+DoRA (H_{1000}+L_{3000})	0.6600↑ 0.04	0.6551↑ 0.06	0.6984↑ 0.10	0.6321↑ 0.07	0.8258↑ 0.07	0.7451↑ 0.11	0.8159↑ 0.10	0.7738↑ 0.12

another fine-tuned on both the manually annotated training set and the training set generated by LRS_{1000} (H_{1000}+L_{3000}).

To evaluate the quality of the LRS-generated data, we randomly sampled 10% of the synthetic dataset for manual inspection. Three annotators independently evaluated each sample in terms of label correctness, semantic consistency, and linguistic naturalness. As summarized in Table 7, the generated data achieves high quality across all evaluation criteria, with 92% for label correctness, 89% for semantic consistency, and 91% for linguistic naturalness. The inter-annotator agreement, measured by Fleiss's kappa, reaches 0.83, 0.78, and 0.80 for the three criteria, respectively, indicating substantial agreement. Furthermore, no significant semantic drift or spurious patterns are identified during manual inspection. These findings suggest that the LRS-generated data maintains high quality and aligns well with the original data distribution, thereby effectively supporting model training.

In the comparative experiments between LoRA and DoRA, as shown in Table 8, both LoRA and DoRA achieved significant improvements over the baseline model, with (H_{1000}+L_{3000}) demonstrating a greater gain than (H_{1000}). In addition, DoRA generally outperforms LoRA in most tasks. Notably, DoRA (H_{1000}+L_{3000}) achieves the best performance across all specific EMC tasks (blue font).

On the other hand, to ensure that DoRA enhances model performance, we designed the following experiments for comprehensive comparison. They also examine whether EMC remains optimal after DoRA processing. Furthermore, as in previous experiments, we divided DoRA into two control groups: (H_{1000}), and (H_{1000}+L_{3000}). This aims to demonstrate whether the train set generated by LRS_{1000} contributes to the fine-tuning process. As shown in Table 9, we observe performance improvements across all models after DoRA. Specifically, on the ACD-1 task, the EMC Apollo-7B maintains the highest $F1_{micro}$ (0.85) and $F1_{macro}$ (0.82) scores after DoRA (H_{1000}+L_{3000}). Even when using only DoRA (H_{1000}), Apollo-7B demonstrated outstanding performance, achieving 0.79 and 0.75, respectively. In the ACD-2 task, the EMC Qwen3-8B achieved significant improvements after DoRA (H_{1000}), with $F1_{micro}$ and $F1_{macro}$ rising from 0.70 to 0.77 and from 0.69 to 0.77, respectively. Although performance improved further after DoRA (H_{1000}+L_{3000}), the gains were modest: $F1_{micro}$ and $F1_{macro}$ increased only from 0.77 to 0.82 and from 0.77 to 0.80, respectively. This model remains the only one in this task with both $F1_{micro}$ and $F1_{macro}$ scores above 0.8.

As shown in Table 9, the EMC Phi-3-small 7B achieved substantial gains on the ACSA task after training with DoRA (H_{1000}+L_{3000}). Its $F1_{micro}$ and $F1_{macro}$ scores rose from 0.81 to 0.95 and from 0.76 to 0.88 respectively, representing the highest values among all models. In some cases, $F1_{micro}$ saw a notable increase from 0.86 on DoRA (H_{1000}) to 0.95 on DoRA (H_{1000}+L_{3000}), further validating the effectiveness of the LRS_{1000}-generated training set. Moreover, Apollo-7B demonstrated substantial performance gains on this task after processing through DoRA (H_{1000}+L_{3000}): $F1_{micro}$ improved from 0.77 to 0.89, while $F1_{macro}$ rose from 0.70 to 0.78. Finally, in the IR task, the EMC Phi-3-small 7B maintained optimal performance after DoRA (H_{1000}+L_{3000}), achieving $F1_{micro}$ and $F1_{macro}$ scores of 0.92 and 0.90, respectively. Additionally, for this task, all models achieved $F1_{micro}$ and $F1_{macro}$ scores exceeding 0.7 only on DoRA (H_{1000}).

4.6. Dataset

To ensure the transparency of the experiment, we have provided detailed statistics and segmentation information for the dataset. The dataset is divided into training, test, and validation sets, with sizes of 4000, 400, and 400, respectively. The training set comprises 1,000 manually-labelled samples and 3,000 samples generated by LRS (4,000 samples in total), whilst the 400 samples in the test set and the 400 samples in the validation set consist entirely of manually labelled data.The test set is strictly disjoint from both the human-annotated training data and the generated samples, ensuring no data leakage.

To illustrate the distribution of labels within the dataset, we have calculated the label distributions for each subclass. Since all sub-aspects are mutually independent and the number of sub-aspects varies across different primary aspects, we analyze the data distribution at the sub-aspect level. As shown in Table 10, most sub-aspects are distributed within a relatively stable range between 5% and 10%, indicating a generally balanced distribution. A few sub-aspects, such as A.1 (Diagnostic Accuracy), A.2 (Treatment Effectiveness), and B.1 (Empathy and Emotional Support), exhibit higher proportions, reflecting that these dimensions are more frequently discussed in patient reviews. This distribution pattern is consistent with real-world scenarios, where clinical outcomes and doctor–patient communication are the primary concerns of patients. Despite these variations, all aspects are adequately represented, and no extreme imbalance is observed. This distribution is consistent with real-world medical review scenarios and provides a reliable basis for model training and evaluation.

4.7. H-ABSA

To further demonstrate the effectiveness of our framework, we designed the following comparative experiments to comprehensively and intuitively showcase the performance of all models and methods across multiple tasks in ABSA. Among these, we introduced E2E tasks to test models' ability to directly complete the entire complex ABSA task. As shown in Table 11, the three EM demonstrate significant improvements over all baseline LLMs across all tasks. EM-3 achieves the best performance in the E2E task, attaining an $F1_{micro}$ score of 0.81. On the other hand, our proposed MuSEM exhibits noticeable improvements in all tasks except ACD-1, where it performs slightly worse than EM-1. Leveraging its cross-task feature sharing advantage, MuSEM elevated the $F1_{macro}$ score for E2E tasks from EM-3's peak of 0.66 to 0.81, enabling the model to exhibit sensitivity even for extremely scarce cases. Since ACD-1 is the first task in MuSEM, it cannot benefit from cross-task feature

Table 10
Label distribution across primary aspects and sub-aspects (%).

Aspect	ID	Percentage (%)
Clinical Quality of Care	A	38.4
Diagnostic Accuracy	A.1	13.2
Treatment Effectiveness	A.2	15.1
Safety and Adverse Events	A.3	5.5
Patient Experience During Procedures	A.4	4.6
Patient–Provider Communication	B	22.6
Empathy and Emotional Support	B.5	12.8
Clarity and Efficiency of Communication	B.6	9.8
Access to Care	C	12.3
Timely Access to Appointments	C.7	6.5
Convenience of Location	C.8	5.8
Facility Environment	D	21.1
Environment and Public Facilities	D.9	7.3
Adequacy of Medical Equipment	D.10	6.9
Privacy Protections and Safety Measures	D.11	6.9
Cost Transparency and Affordability	E	5.6
Out-of-Pocket Cost Clarity	E.12	5.6

sharing. Its performance is therefore nearly identical to that of the underlying EM-1. This demonstrates that both DCA and TCA within our framework enhance model performance across remaining tasks, with ACD-2 achieving $F1_{micro}$ and $F1_{macro}$ scores of 0.87 and 0.84, respectively. Particularly in the ACSA and IR tasks, both $F1_{micro}$ (0.96, 0.95) and $F1_{macro}$ (0.90, 0.92) achieved scores exceeding 0.9.

Table 12 presents a comparison of the performance of the proposed method against representative ABSA methods and proprietary LLM on E2E tasks. To ensure consistency with the comparative experiments involving the 10 baseline models in Table 11, we did not integrate the H-ABSA architecture into Bert-ABSA [77], Prompt-based ABSA (P-ABSA) [18], Lego-ABSA [78], and T5-ABSA [79]. All these baseline models were evaluated on E2E tasks under a unified flat-labelling setting. This setting allows for a direct comparison of existing models' capabilities on end-to-end tasks within the same label space, whilst avoiding confounding factors introduced by additional architectural modifications.

Compared with these task-specific ABSA baselines, our method achieves consistently superior performance, particularly in terms of $F1_{macro}$, demonstrating its effectiveness in handling imbalanced and fine-grained aspect distributions. Closed-source LLMs, including Claude 4 [80], Gemini 3 [81] and ChatGPT-5 [82], achieve the highest absolute F1 scores on the E2E task. This result is expected given their substantially larger model scales, broader pre-training corpora, and stronger general-purpose reasoning capabilities. However, MuSEM achieves competitive performance while relying on smaller expert models that can be deployed locally. In addition, MuSEM shows lower standard deviations than the closed-source LLMs, suggesting more stable performance under repeated evaluation. These results indicate that MuSEM provides a favorable trade-off among performance, stability, privacy, and deployability, rather than aiming to surpass closed-source LLMs in absolute accuracy.

To evaluate the generalization capability of the proposed model, we further conduct cross-dataset experiments on HealthTap and WebMD. We use the same annotation protocol as for the RateMDS dataset, labeling 300 patient reviews from each dataset, resulting in a total of 600 annotated samples. As shown in Table 13, the model maintains comparable performance across all tasks when transferred to unseen datasets. Although a slight performance degradation is observed compared to the source dataset, the overall results remain stable, demonstrating the robustness of the proposed framework under domain shifts. The relatively small performance gaps across datasets indicate that the model is capable of capturing domain-invariant features and generalizing effectively across different medical platforms. This also suggests a certain level of consistency in the expression patterns of patient reviews across platforms.

To demonstrate the reproducibility of our experiments, we report the running time and GPU memory consumption of all models under a unified hardware setting, as shown in Table 14. All efficiency measurements are conducted on a single NVIDIA RTX 4090 GPU (24 GB VRAM), and the latency is measured per sample. In MuSEM, the EMs are executed sequentially rather than simultaneously, and thus the memory consumption does not scale linearly with the number of experts. The slightly higher VRAM usage mainly arises from additional modules such as DCA and TCA. Despite the increased latency, MuSEM achieves significantly better performance while remaining deployable on a single GPU. Furthermore, the runtime can be further reduced with multi-GPU deployment or parallelization.

4.8. Ablation Experiments

To thoroughly investigate the specific contributions of Few-shot examples and CoT prompts to this research, we systematically designed and conducted a series of ablation experiments. The specific settings are as follows: We performed three sets of retraining on three EMs, sequentially excluding Few-shot examples, CoT prompts, and both together, and compared the results with the model performance under the full configuration. Detailed configurations include:

- w/o Few-shot: The model receives only the task prompt and CoT inference prompt without any Few-shot examples, followed by DoRA fine-tuning training.
- w/o CoT: The model receives only the task prompt and Few-shot examples without introducing CoT instructions, proceeding directly to DoRA fine-tuning training.
- w/o Few-shot and w/o CoT: The model undergoes fine-tuning based solely on the task prompt, without providing Few-shot examples or introducing CoT instructions, directly executing DoRA fine-tuning.

Table 11
Performance Results of Base Model, EMs, and MuSEM Across Tasks

Model	E2E Task		ACD-1 Task		ACD-2 Task		ACSA Task		IR Task	
	$F1_{\text{micro}}$	$F1_{\text{macro}}$	$F1_{\text{micro}}$	$F1_{\text{macro}}$	$F1_{\text{micro}}$	$F1_{\text{macro}}$	$F1_{\text{micro}}$	$F1_{\text{macro}}$	$F1_{\text{micro}}$	$F1_{\text{macro}}$
Apollo-7B (BM-E1)	0.6071	0.4842	0.7483	0.7205	0.6631	0.6266	0.7724	0.7040	0.8097	0.7631
Gemma-2-9B-it	0.5223	0.4753	0.6995	0.6687	0.5888	0.5603	0.7831	0.7223	0.7495	0.6786
GLM-Z1-9B-0414	0.5493	0.4536	0.6716	0.6521	0.6320	0.5997	0.7559	0.6657	0.7256	0.7102
InternLM2.5-Chat-7B	0.5048	0.4399	0.5729	0.5401	0.5672	0.5258	0.7225	0.6018	0.7193	0.6645
Llama3.1-8B-Instruct	0.5725	0.5167	0.6610	0.6317	0.6263	0.5881	0.7338	0.6141	0.7380	0.6638
Ministral-8B-Instruct-2410	0.5231	0.4424	0.5762	0.5633	0.5801	0.5695	0.6870	0.6092	0.7361	0.7041
Phi-3-small-7B (BM-E3)	0.6404	0.5578	0.6832	0.6539	0.6729	0.6670	0.8144	0.7589	0.8396	0.8283
Qwen3-8B (BM-E2)	0.6333	0.5659	0.7054	0.6861	0.7009	0.6914	0.8036	0.7208	0.8250	0.7962
Starling-7B-LM-Beta	0.4650	0.4028	0.4998	0.4785	0.4937	0.4629	0.6591	0.6235	0.7415	0.7244
Yi-1.5-9B-Chat	0.5516	0.4872	0.6175	0.5968	0.5942	0.5654	0.7602	0.6353	0.7144	0.6549
EM-1	0.7492	0.5903	0.8515	0.8233	0.6727	0.6560	0.9246	0.8395	0.9071	0.8812
EM-2	0.7051	0.6336	0.7780	0.6717	0.8242	0.8028	0.8529	0.7453	0.8897	0.8113
EM-3	0.8093	0.6631	0.7971	0.7592	0.7533	0.7246	0.9503	0.8837	0.9229	0.9031
Ours	0.8224	0.8097	0.8509	0.8267	0.8666	0.8434	0.9638	0.9022	0.9459	0.9176

(BM-E1): The Base Model of EM-1
(BM-E2): The Base Model of EM-2
(BM-E3): The Base Model of EM-3

Table 12
Comparison of the Proposed Method with Existing ABSA Approaches and Closed-source LLMs on the E2E Task

Model	$F1_{\text{micro}}$	$F1_{\text{macro}}$
Bert-ABSA[77]	0.6744 ± 0.021	0.6180 ± 0.025
P-ABSA[18]	0.7565 ± 0.029	0.7238 ± 0.033
Lego-ABSA[78]	0.7896 ± 0.026	0.7612 ± 0.028
T5-ABSA[79]	0.7632 ± 0.032	0.7204 ± 0.038
Claude 4 Opus[80]	0.8437 ± 0.030	0.8326 ± 0.033
Gemini 3[81]	0.8619 ± 0.028	0.8581 ± 0.032
ChatGPT-5[82]	0.8575 ± 0.027	0.8408 ± 0.029
EM-1	0.7492 ± 0.024	0.5903 ± 0.031
EM-2	0.7051 ± 0.029	0.6336 ± 0.033
EM-3	0.8093 ± 0.026	0.6631 ± 0.031
Ours	0.8224 ± 0.019	0.8097 ± 0.024

The ablation results are presented in Table 15. Consistent and significant performance trends can be observed across multiple tasks. Taking EM-1 as an example, when the Few-shot setting is removed, the model's $F1_{macro}$ and $F1_{macro}$ scores decrease by over 0.08 on average across the ACD-1, ACD-2, ACSA, and IR tasks. Specifically, in the ACD-1 task, the $F1_{macro}$ score drops from 0.82 to 0.75, indicating slight numerical fluctuation but an overall degradation in performance. Removing only CoT also leads to a noticeable decline, with an average $F1_{macro}$ reduction of approximately 0.03. This drop is particularly evident in the ACD-2 task, where performance decreases from 0.66 to 0.62, underscoring the crucial role of CoT in complex semantic classification tasks. The most substantial degradation occurs when both Few-shot and CoT are removed simultaneously, resulting in average declines of over 0.11 in both $F1_{micro}$ and $F1_{macro}$. This finding indicates that the absence of both severely undermines the model's generalization and reasoning abilities.

In EM-2, we observed a similar pattern: without the Few-shot setting, the F1 scores declined by approximately 0.06 on average, while without CoT, they decreased by about 0.03 on average. When both Few-shot and CoT were removed, $F1_{macro}$ and $F1_{micro}$ also declined by an average of more than 0.08. Notably, the ACD-2 task exhibited a drop from 0.80 to 0.70, further validating the necessity of combining CoT and Few-shot learning. Additionally, EM-3 showed particularly pronounced performance shifts in the ACSA task. While the full model achieved an $F1_{micro}$ score of 0.95, it dropped to 0.82 without both Few-shot and CoT instructions. Meanwhile, the model's $F1_{macro}$ score on the IR task decreased from 0.90 to 0.80. In summary, the ablation results demonstrate that both Few-shot examples and CoT reasoning mechanisms independently and synergistically enhance model performance. They not only improve the model's generalization capabilities across multiple tasks but also partially compensate for the limited guidance capacity of either method alone. This ablation study confirms their importance as effective prompting components in specialized model fine-tuning and provides evidence for future research.

To further evaluate the contributions of DCA, TCA, and DoRA, we conducted a series of ablation experiments. Since TCA is constructed on top of DCA, it is not feasible to retain TCA while removing DCA. Therefore, we only consider configurations where TCA is removed together with or after DCA. As shown in Table 16, the full model consistently achieves the best performance, demonstrating the effectiveness of integrating all components. Removing TCA results in slight performance degradation across most tasks, as the ACD-1 and ACD-2 missions took place prior to the TCA module, they were not affected by it. When both DCA and TCA are removed, performance drops more noticeably on ACD-2 task and downstream tasks, suggesting

Table 13
Cross-dataset evaluation on RateMDS, HealthTap, and WebMD datasets.

Dataset	ACD-1 Task		ACD-2 Task		ACSA Task		IR Task		E2E Task	
	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
RateMDS	0.8509	0.8267	0.8666	0.8434	0.9638	0.9022	0.9459	0.9176	0.8224	0.8097
HealthTap	0.8421	0.8164	0.8516	0.8348	0.9385	0.8717	0.9314	0.8905	0.8403	0.8081
WebMD	0.8386	0.8037	0.8579	0.8302	0.9423	0.8976	0.9412	0.9148	0.8356	0.8179

Table 14
Inference efficiency and E2E performance comparison.

Model	Time(s)	VRAM (GB)	$F1_{micro}$	$F1_{macro}$
BERT-ABSA	0.4	1.2	0.674	0.618
Apollo-7B	6.6	15.5	0.607	0.484
Phi-3-small-7B	5.2	15.9	0.640	0.558
Qwen3-8B	10.3	17.3	0.633	0.566
MuSEM (Ours)	24.8	19.7	0.822	0.810

that these two modules provide complementary benefits for hierarchical feature interaction. In contrast, removing DoRA leads to substantial and statistically significant performance declines across all tasks, especially on ACD-1 and ACSA, highlighting its critical role in effective representation adaptation. These results confirm that while DCA and TCA contribute to performance improvements, DoRA is the most essential part in the proposed framework.

5. Discussion

In this section, we further analyze the proposed approach, including its advantages, limitations, and comparison with existing methods. Furthermore, we highlight the strengths of H-ABSA in application management, suggesting that future research should focus on validating its generalizability across diverse medical and cultural contexts. Finally, we acknowledge the current limitations regarding runtime performance, and we propose that future efforts should emphasize improving both inference efficiency and overall computational performance.

5.1. Multi-task Specialized Expert Model Framework (MuSEM)

We employ the MuSEM framework to efficiently distribute complex ABSA tasks among three EMs for sequential collaborative processing. We introduce DCA and TCA to enable subsequent tasks to leverage feature results from preceding tasks. The results demonstrate that subsequent tasks enhance their respective accuracy by sharing task features from preceding tasks, with the final end-to-end task performance significantly outperforming that of a single sLLM. Moreover, one reason for MuSEM's improved accuracy may lie in its avoidance of multi-step complex tasks. Previous research has demonstrated that LLMs exhibit lower proficiency in handling intricate tasks [73]. Consequently, our MuSEM approach effectively assigns each LLM to perform a single task, thereby enabling collaborative completion of a complex task. We shall test this hypothesis in other domains in the future.

However, this framework still holds room for improvement. As shown in Table 11, during the first task (ACD-1), the initial EM lacks feature input from preceding tasks. Consequently, its accuracy shows minimal improvement over EM-1 and even slightly underperforms. Therefore, enhancing the accuracy of the first task is crucial. This significantly aids in improving overall E2E accuracy, as boosting the first task's accuracy elevates T1F's accuracy. T1F's improved accuracy then positively impacts the accuracy of every subsequent task, creating a virtuous cycle. Researchers introduced the concept of continuous learning as early as 2020 [83]. This approach involves continuously converting the latest runtime results into trainable datasets while the model operates. This not only persistently improves the accuracy of the first task but also enhances overall accuracy, enabling the model to continually learn new task features.

During this process, to avoid the accumulation of errors, noise, or biases that may emerge as the model autonomously generates and reuses its own training data, manual verification of selected portions of the newly constructed datasets is often required. This human-in-the-loop validation ensures data reliability and prevents the model from reinforcing incorrect representations or learning spurious task features that could compromise overall performance. Separately, researchers designed reward functions to leverage reinforcement learning for Continuous Reinforcement Learning (CRL). CRL addresses limitations of traditional continuous learning by enabling agents to persistently learn, adapt to new tasks, and retain prior knowledge. Thus, integrating CRL could empower MuSEM to perform efficient autonomous continuous learning while executing complex ABSA tasks, thereby optimizing accuracy across branch tasks. Finally, while MuSEM successfully establishes connections across different branch tasks, relationships between distinct E2E tasks (i.e., different patient reviews) remain relatively independent. Therefore, our plan involves connecting diverse patient reviews into a unified corpus. Subsequent sub-tasks will not only extract task features from the current humble task but also leverage techniques like text similarity detection to access desired features from all prior tasks. This will create a "massive task feature repository" for utilization by all subsequent E2E tasks and sub-tasks.

5.2. H-ABSA

To address the issue of overly coarse sentiment category in current ABSA tasks for patient reviews, we propose

Table 15
Ablation Experiments for Few-shot and CoT Directive. Statistical significance is evaluated against the full model (* $p < 0.05$, ** $p < 0.01$).

	ACD-1 Task		ACD-2 Task		ACSA Task		IR Task	
	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
EM-1	0.8515	0.8233	0.6727	0.6560	0.9246	0.8395	0.9071	0.8812
w/o Few-shot	$0.7628^{\downarrow 0.09^{**}}$	$0.7453^{\downarrow 0.08^{**}}$	$0.5997^{\downarrow 0.07^{**}}$	$0.6412^{\downarrow 0.01}$	$0.7965^{\downarrow 0.13^{**}}$	$0.7440^{\downarrow 0.10^{**}}$	$0.8479^{\downarrow 0.06^{*}}$	$0.7806^{\downarrow 0.10^{**}}$
w/o CoT	$0.8282^{\downarrow 0.02^{*}}$	$0.7927^{\downarrow 0.03^{*}}$	$0.6363^{\downarrow 0.04^{*}}$	$0.6219^{\downarrow 0.03^{*}}$	$0.8812^{\downarrow 0.04^{*}}$	$0.7954^{\downarrow 0.04^{*}}$	$0.8886^{\downarrow 0.02}$	$0.8408^{\downarrow 0.04^{*}}$
w/o Few-shot and CoT	$0.7483^{\downarrow 0.10^{**}}$	$0.7005^{\downarrow 0.12^{**}}$	$0.5831^{\downarrow 0.09^{**}}$	$0.5766^{\downarrow 0.08^{**}}$	$0.7724^{\downarrow 0.15^{**}}$	$0.7140^{\downarrow 0.13^{**}}$	$0.8097^{\downarrow 0.10^{**}}$	$0.7731^{\downarrow 0.11^{**}}$
EM-2	0.7780	0.6717	0.8242	0.8028	0.8529	0.7453	0.8897	0.8113
w/o Few-shot	$0.7093^{\downarrow 0.07^{**}}$	$0.6209^{\downarrow 0.05^{*}}$	$0.7648^{\downarrow 0.06^{*}}$	$0.7482^{\downarrow 0.05^{*}}$	$0.7966^{\downarrow 0.06^{*}}$	$0.7010^{\downarrow 0.04^{*}}$	$0.8525^{\downarrow 0.04^{*}}$	$0.7743^{\downarrow 0.04^{*}}$
w/o CoT	$0.7434^{\downarrow 0.03^{*}}$	$0.6471^{\downarrow 0.02}$	$0.7808^{\downarrow 0.04^{*}}$	$0.7991^{\downarrow 0.00}$	$0.8345^{\downarrow 0.02}$	$0.7222^{\downarrow 0.02}$	$0.8559^{\downarrow 0.03^{*}}$	$0.8086^{\downarrow 0.00}$
w/o Few-shot and CoT	$0.6854^{\downarrow 0.09^{**}}$	$0.5961^{\downarrow 0.08^{**}}$	$0.7209^{\downarrow 0.10^{**}}$	$0.7014^{\downarrow 0.10^{**}}$	$0.7736^{\downarrow 0.08^{**}}$	$0.6708^{\downarrow 0.07^{**}}$	$0.8350^{\downarrow 0.05^{*}}$	$0.7562^{\downarrow 0.05^{*}}$
EM-3	0.7971	0.7592	0.7533	0.7246	0.9503	0.8837	0.9229	0.9031
w/o Few-shot	$0.7217^{\downarrow 0.08^{**}}$	$0.7062^{\downarrow 0.05^{*}}$	$0.6839^{\downarrow 0.07^{**}}$	$0.6484^{\downarrow 0.08^{**}}$	$0.8703^{\downarrow 0.08^{**}}$	$0.8650^{\downarrow 0.02}$	$0.8727^{\downarrow 0.05^{*}}$	$0.8574^{\downarrow 0.05^{*}}$
w/o CoT	$0.7700^{\downarrow 0.03^{*}}$	$0.7147^{\downarrow 0.04^{*}}$	$0.7338^{\downarrow 0.02}$	$0.6918^{\downarrow 0.03^{*}}$	$0.9127^{\downarrow 0.04^{*}}$	$0.8632^{\downarrow 0.02}$	$0.8818^{\downarrow 0.04^{*}}$	$0.8764^{\downarrow 0.03^{*}}$
w/o Few-shot and CoT	$0.7032^{\downarrow 0.09^{**}}$	$0.6639^{\downarrow 0.10^{**}}$	$0.6529^{\downarrow 0.10^{**}}$	$0.6070^{\downarrow 0.12^{**}}$	$0.8244^{\downarrow 0.13^{**}}$	$0.8089^{\downarrow 0.08^{**}}$	$0.8396^{\downarrow 0.08^{**}}$	$0.7983^{\downarrow 0.10^{**}}$

Table 16
Ablation Study of DCA, TCA, and DoRA Modules. Statistical significance is evaluated against the full model (* $p < 0.05$, ** $p < 0.01$).

	ACD-1 Task		ACD-2 Task		ACSA Task		IR Task	
	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
Ours	0.8509	0.8267	0.8666	0.8434	0.9638	0.9022	0.9459	0.9176
w/o TCA	$0.8442^{\downarrow 0.01}$	$0.8214^{\downarrow 0.01}$	$0.8661^{\downarrow 0.00}$	$0.8430^{\downarrow 0.00}$	$0.9443^{\downarrow 0.02}$	$0.8615^{\downarrow 0.04^{*}}$	$0.9187^{\downarrow 0.03^{*}}$	$0.8894^{\downarrow 0.03^{*}}$
w/o DCA and TCA	$0.8594^{\uparrow 0.01}$	$0.8377^{\uparrow 0.01}$	$0.8114^{\downarrow 0.06^{*}}$	$0.7931^{\downarrow 0.05^{*}}$	$0.9405^{\downarrow 0.02}$	$0.8588^{\downarrow 0.04^{*}}$	$0.9109^{\downarrow 0.04^{*}}$	$0.8851^{\downarrow 0.03^{*}}$
w/o DoRA	$0.6404^{\downarrow 0.21^{**}}$	$0.5578^{\downarrow 0.27^{**}}$	$0.7437^{\downarrow 0.12^{**}}$	$0.7188^{\downarrow 0.12^{**}}$	$0.8614^{\downarrow 0.10^{**}}$	$0.8155^{\downarrow 0.09^{**}}$	$0.8713^{\downarrow 0.07^{*}}$	$0.8590^{\downarrow 0.06^{*}}$

H-ABSA, which enables LLMs to perform ABSA tasks with greater precision. However, researchers have noted that different countries possess distinct healthcare environments, leading to varying priorities in medical focus [84]. For instance, in some regions, medical communication or efficacy may be emphasized, in others, physician expertise or cost may be prioritized, and in yet others, different aspects may take precedence. Importantly, the diversity and flexibility of our model allow it to accommodate variations in healthcare focus across different medical systems. Since clinical priorities, patient expectations, and service delivery models differ substantially between countries and even institutions, H-ABSA's modular layered architecture can be readily adapted to real-world scenarios. Specifically, each aspect layer can be expanded, merged, or fine-tuned to reflect specific healthcare practices, ensuring consistent analytical performance across heterogeneous datasets. This flexibility not only enhances cross-domain generalizability but also reinforces the model's applicability in diverse cultural and clinical contexts.

On another front, as illustrated in Figure 3, we innovatively incorporated an IR task into the ABSA framework. The IR task enhances management across any domain. For instance, it helps determine the urgency level of a patient's comment based on whether they intend to suggest or complain. For patients who express strong complaints or use abusive language, administrators should prioritize analyzing the complaint content and resolving the matter promptly. Compared to patients offering calm suggestions, those voicing complaints or abuse require more immediate attention due to the higher urgency of their concerns. Therefore, using the IR task to prioritize patients based on intent category is highly efficient and scientifically sound for administrators.

Moreover, within the realm of medical applications, ABSA has consistently been a prominent research focus, encompassing multiple facets of healthcare. For instance, researchers employed ABSA to investigate public acceptance and attitudes towards COVID-19 vaccines [85]. Similarly, other studies utilised ABSA to explore public perceptions of government public health policies during the COVID-19 pandemic [86]. Furthermore, researchers have employed ABSA to analyse drug reviews, thereby evaluating drug efficacy and safety to enhance healthcare quality and rational medication use [87]. These studies demonstrate sufficient accuracy and adequate aspect categories at the application level. However, the application of ABSA to enhance healthcare quality through patient reviews necessitates continuous refinement of sentiment granularity. As shown in Figure 1, insufficient aspect categories hinder the identification of specific clinical departments responsible for improvements and impede the precise identification of patient dissatisfaction and needs, thereby limiting targeted healthcare management.

On the other hand, the outputs of the proposed framework are inherently interpretable due to the hierarchical structure of the H-ABSA schema. In practical scenarios, these structured outputs can support multiple downstream applications. First, at the operational level, healthcare providers can use ABSA to identify specific service deficiencies. For instance, negative sentiment associated with sub-aspects such as "waiting time" or "communication clarity" can highlight bottlenecks in service delivery and guide targeted improvements.

Second, at the management level, the results can be aggregated and visualized in interactive dashboards, enabling administrators to monitor patient satisfaction trends across different aspects, departments, or time periods. Such dashboards can provide actionable insights for resource allocation, service optimization, and policy adjustments. Third, the integration with decision-support systems allows automatic detection of critical issues and prioritization of interventions. For example, frequent complaints related to safety or treatment effectiveness can trigger alerts for further investigation, supporting data-driven clinical governance. Moreover, the modular design of the proposed framework facilitates customization for different healthcare systems. The hierarchical aspect schema can be adapted to reflect local clinical practices, while the model outputs remain interpretable and consistent across different deployment scenarios.

Therefore, this paper proposes a fine-grained aspect classification framework based on H-ABSA to assist in accurately identifying department-level issues and capturing nuanced patient sentiments and intentions, thereby supporting more precise and practically healthcare management.

5.3. Inference Efficiency and Practical Applicability

The per-sample latency of MuSEM should be considered when interpreting its practical applicability. As reported in Table 14, MuSEM requires 24.8 seconds per sample on a single NVIDIA RTX 4090 GPU, which is higher than that of the single-model baselines. This increased inference time is mainly caused by the sequential execution of multiple expert models across the hierarchical subtasks. At the same time, MuSEM remains deployable on a single consumer-grade GPU, requiring 19.7 GB of VRAM, while achieving substantially stronger E2E performance. This result indicates a clear trade-off between inference efficiency and fine-grained analytical performance. Therefore, MuSEM is more suitable for offline or batch-oriented healthcare feedback analysis, where aspect-level accuracy, interpretability, and clinical actionability are prioritized over real-time response speed. For example, hospitals or healthcare platforms may periodically analyze accumulated patient reviews to identify recurring issues in communication, treatment effectiveness, access to care, or facility environments.

This also clarifies the practical deployment scope of the proposed framework. MuSEM is not intended as a real-time conversational system or instant clinical decision-support tool in its current form. Instead, it provides a deployable and privacy-preserving analytical framework for structured patient feedback mining. Future deployment-oriented optimization could further reduce inference latency through expert parallelization, multi-GPU execution, model compression, quantization, or early-exit routing.

5.4. Compared with existing approaches

Compared with existing ABSA methods, the proposed framework has certain advantages and limitations. First, unlike traditional coarse-grained labeling schemes [18], our hierarchical aspect taxonomy enables multi-level modeling of patient reviews, which better aligns with real-world clinical evaluation processes. This design allows the model to capture both high-level categories and fine-grained sub-aspects, resulting in improved interpretability and more informative analysis outcomes. Second, instead of training a single large model from scratch, our multi-sLLM paradigm leverages and orchestrates existing models through a structured pipeline. This design not only reduces training cost but also enables flexible integration of heterogeneous expert models, making the framework more adaptable to different tasks and domains.

Third, in contrast to existing LLM-based approaches that primarily rely on scaling model parameters to improve performance, our MuSEM framework adopts a multi-expert collaboration strategy based on small-scale LLMs. By introducing specialized expert routing and cross-task feature sharing, the framework effectively integrates complementary knowledge across tasks, achieving strong empirical performance without relying on large-scale models. However, these advantages come with certain trade-offs. Due to the sequential invocation of multiple expert models, the inference process incurs higher latency compared to single-model approaches. In addition, the construction of the hierarchical aspect taxonomy relies on domain expertise, which may introduce additional annotation costs when extending to new domains. Finally, although the model demonstrates good generalization across datasets, its performance may still be influenced by domain-specific variations in patient reviews.

5.5. Limitations

Although MuSEM demonstrates strong performance and flexibility in complex ABSA tasks, the model still has certain limitations that need to be addressed sequentially in future work. First, although MuSEM achieves improved performance through multi-expert collaboration, this design also introduces additional inference latency. As shown in Table 14, MuSEM requires 24.8 seconds per inference, which is higher than the baseline LLMs, whose inference times range from approximately 5.2 to 10.3 seconds. This is acceptable for offline patient review analysis and batch-level healthcare management, where real-time response is not always required. However, the increased latency may limit its applicability in real-time or interactive systems, such as online decision-support dashboards or live feedback monitoring platforms. Future work will therefore explore inference acceleration strategies, such as parallel expert

execution, model compression, and more efficient routing mechanisms.

Second, the methods and models developed in this study were trained and optimized exclusively for English text, without extension to other languages. Their performance and generalization capabilities remain unverified in multilingual environments, particularly for low-resource languages. But the hierarchical aspect schema can be adapted through aspect re-mapping, where primary aspects remain stable while sub-aspects are adjusted to reflect domain-specific or cultural differences. In addition, the modular multi-expert design allows efficient adaptation via few-shot fine-tuning, enabling the framework to generalize to new settings with limited additional data. Future work will focus on optimizing inference efficiency and exploring training and adaptation on multilingual corpora to enhance the model's practicality and applicability.

6. Conclusions

This study addresses the limitations of existing ABSA methods in patient review processing, including insufficient classification granularity, challenging data annotation, high training difficulty, and poor accuracy, by proposing the H-ABSA and MuSEM frameworks. Results demonstrate that our approach significantly enhances model performance and sentiment granularity for ABSA tasks targeting patient reviews. Specifically, H-ABSA constructs a hierarchical labeling system supports multi-level analysis from semantics to sentiment. Meanwhile, MuSEM decomposes the complex ABSA task into four subtasks and synergizes multiple lightweight sLLM EM. This achieves high accuracy with low-cost local deployment. In the end-to-end ABSA task, it achieves $F1_{micro}$ and $F1_{macro}$ scores of 0.82 and 0.81, respectively, significantly outperforming base models. This research not only contributes to improved analytical performance but also proposes a sustainable and computationally efficient solution pathway for complex NLP tasks. Future research will focus on enhancing MuSEM's cross-lingual generalization capabilities, exploring CRL mechanisms to strengthen the model's adaptive optimization and enhance the overall operational efficiency of the model. These efforts aim to further elevate the system's practical value in real-world healthcare management scenarios.

7. Acknowledgements

This work was supported in part by the Macao Science and Technology Development Fund (FDCT) under Grant 0029/2025/AIJ, in part by Macao Polytechnic University under Grant RP/FCA-24/2025, and fca.3d39.03f5.e, as well as in part by the Fundação para a Ciência e a Tecnologia (FCT), Portugal, through the Project UIDB/00326/2025. We would also like to thank the doctors from West China Hospital of Sichuan University, the First Affiliated Hospital of Sun Yat-sen University, and Shenzhen People's Hospital for their contributions and guidance regarding the dataset used in this study.

CRediT authorship contribution statement

Jiaxuan Li: Writing – original draft, Methodology, Software, Validation. **Jielong Guo:** Methodology, Software, Validation. **Patrick Pang:** Supervision, Project administration, Writing – review & editing, Funding acquisition, Conceptualization. **Hugo Gonçalo Oliveira:** Supervision, Writing – review & editing, Funding acquisition, Conceptualization. **Benjamin K. Ng:** Supervision, Writing – review & editing. **Tao Tan:** Supervision, Conceptualization.

References

[1] Ali Nikseresht, Sajjad Shokouhyar, Erfan Babaee Tirkolaee, Sina Shokoohyar, Sadia Samar Ali, and Mohammad Zoynul Abedin. An intelligent predictive framework for consumer returns forecasting: Leveraging social media data in the electronics service industry. *Advanced Engineering Informatics*, 66:103433, 2025.

[2] Ravi Shankar, Qian Xu, and Anjali Bundele. Patient voices in dialysis care: Sentiment analysis and topic modeling study of social media discourse. *Journal of Medical Internet Research*, 27:e70128, 2025.

[3] Yong-Hai Li, Wei-Wei Wang, Shan-Tao Yue, Jing-Mei Wang, and Bing Lei. A new product development method to incorporating customer sensory preferences in food product design. *Advanced Engineering Informatics*, 62:102769, 2024.

[4] Hyeongchan Cho, Kyu-Min Kim, Jee-Young Kim, and Bo-Young Youn. Twitter discussions on# digitaldementia: content and sentiment analysis. *Journal of Medical Internet Research*, 26:e59546, 2024.

[5] Lin Gui and Yulan He. Understanding patient reviews with minimum supervision. *Artificial Intelligence in Medicine*, 120:102160, 2021.

[6] Jiaxuan Li, Yunchu Yang, Rong Chen, Dashun Zheng, Patrick Cheong-Iao Pang, Chi Kin Lam, Dennis Wong, and Yapeng Wang. Identifying healthcare needs with patient experience reviews using chatgpt. *PLoS One*, 20(3):e0313442, 2025.

[7] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038, 2022.

[8] Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11):296, 2024.

[9] Yuehua Zhao, Linyi Zhang, Chenxi Zeng, Wenrui Lu, Yidan Chen, and Tao Fan. Construction of an aspect-level sentiment analysis model for online medical reviews. *Information Processing & Management*, 60(6):103513, 2023.

[10] Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863, 2022.

[11] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830, 2015.

[12] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49, 2016.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[14] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Nordic Conference on Computational Linguistics*, pages 187–196, 2019.

[15] Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems With Applications*, 118:272–299, 2019.

[16] Ganpat Singh Chauhan, Ravi Nahta, Yogesh Kumar Meena, and Dinesh Gopalani. Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, 49:100576, 2023.

[17] Hao Zhang, Yu-N Cheah, Osamah Mohammed Alyasiri, and Jieyu An. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and chatgpt: a comprehensive survey. *Artificial Intelligence Review*, 57(2):17, 2024.

[18] Jiaxuan Li, Yunchu Yang, Chao Mao, Patrick Cheong-Iao Pang, Quanjing Zhu, Dejian Xu, and Yapeng Wang. Revealing patient dissatisfaction with health care resource allocation in multiple dimensions using large language models and the international classification of diseases 11th revision: aspect-based sentiment analysis. *Journal of Medical Internet Research*, 27:e66344, 2025.

[19] Wei Wei, Chenliang Hao, and Zixin Wang. User needs insights from ugc based on large language model. *Advanced Engineering Informatics*, 65:103268, 2025.

[20] Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digital Medicine*, 8(1):143, 2025.

[21] Xiaodong Wu, Ran Duan, and Jianbing Ni. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of information and intelligence*, 2(2):102–115, 2024.

[22] A Dorca Josa and Marc Bleda-Bejar. Local llms: Safeguarding data privacy in the age of generative ai. a case study at the university of andorra. In *ICERI 2024 Proceedings*, pages 7879–7888. IATED, 2024.

[23] Yunchu Yang, Jiaxuan Li, Jielong Guo, Patrick Cheong-Iao Pang, Yapeng Wang, Xu Yang, and Sio-Kei Im. Performance evaluation and application potential of small large language models in complex sentiment analysis tasks. *IEEE Access*, 2025.

[24] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[25] Minyoung Kyoung, Joon-Ho Lim, and Youngsoo Kim. Reasoning beyond length limits: Improving accuracy in long-context question answering with small-scale language models. *IEEE Access*, 2025.

[26] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[27] Haoxiang Luo, Yinqiu Liu, Ruichen Zhang, Jiacheng Wang, Gang Sun, Dusit Niyato, Hongfang Yu, Zehui Xiong, Xianbin Wang, and Xuemin Shen. Toward edge general intelligence with multiple-large language model (multi-llm): architecture, trust, and orchestration. *IEEE Transactions on Cognitive Communications and Networking*, 2025.

[28] Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, et al. When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2502.04506*, 2025.

[29] Zengzhi Wang, Rui Xia, and Jianfei Yu. Unified absa via annotation-decoupled multi-task instruction tuning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7242–7254, 2024.

[30] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[31] Nils Constantin Hellwig, Jakob Fehle, and Christian Wolff. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, 261:125514, 2025.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[33] Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003, 2023.

[34] Yuanrong Zhang, Wei Guo, Zhixing Chang, Jian Ma, Zhonglin Fu, Lei Wang, and Hongyu Shao. User requirement modeling and evolutionary analysis based on review data: Supporting the design upgrade of product attributes. *Advanced Engineering Informatics*, 62:102861, 2024.

[35] Jyothis Joseph, S Vineetha, and NV Sobhana. A survey on deep learning based sentiment analysis. *Materials Today: Proceedings*, 58:456–460, 2022.

[36] Minghui Huang, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, and Fu Lee Wang. Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*, 13(3):1337–1348, 2020.

[37] Arwa Diwali, Kawther Saeedi, Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis. *IEEE Transactions on Affective Computing*, 2023.

[38] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. Bert for stock market sentiment analysis. In *IEEE International Conference on Tools with Artificial Intelligence*, pages 1597–1601. IEEE, 2019.

[39] Kaixin Sha, Yupeng Li, Yanan Dong, and Na Zhang. Modelling the dynamics of customer requirements considering their lability and sensitivity in product development. *Advanced Engineering Informatics*, 59:102296, 2024.

[40] Wenna Lai, Haoran Xie, Guandong Xu, and Qing Li. Rvisa: reasoning and verification for implicit sentiment analysis. *IEEE Transactions on Affective Computing*, 2025.

[41] Jiaxuan Li, Patrick Cheong-Iao Pang, Yundan Xiao, and Dennis Wong. Changes in doctor–patient relationships in china during covid-19: a text mining analysis. *International Journal of Environmental Research and Public Health*, 19(20):13446, 2022.

[42] Duan Chen, Huang Zhengwei, Tan Yiting, Min Jintao, and Ribesh Khanal. Emotion and sentiment analysis for intelligent customer service conversation using a multi-task ensemble framework. *Cluster Computing*, 27(2):2099–2115, 2024.

[43] Yupeng Li, Zilu Su, Ke Chen, and Kexin Jiang. Aspect-based sentiment analysis via knowledge enhancement. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2024.

[44] Usha Divakarla, Harshaditya Sharma, and K Chandrasekarana. A systematic literature review on multimodal aspect-based sentiment analysis. In *Eighth International Conference on Parallel, Distributed and Grid Computing*, pages 78–84. IEEE, 2024.

[45] Dashun Zheng, Jiaxuan Li, Yunchu Yang, Yapeng Wang, and Patrick Cheong-Iao Pang. Microbert: Distilling moe-based knowledge from bert into a lighter model. *Applied Sciences*, 14(14):6171, 2024.

[46] Zohair Ahmed, Junwen Duan, FangXiang Wu, and Jianxin Wang. Efca: An extended formal concept analysis method for aspect extraction in healthcare informatics. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1241–1244. IEEE, 2021.

[47] Avijit Thawani, Michael J Paul, Urmimala Sarkar, and Byron C Wallace. Are online reviews of physicians biased against female providers? In *Machine Learning for Healthcare Conference*, pages 406–423. PMLR, 2019.

[48] Kelvin I Afrashtehfar, Mansour KA Assery, and S Ross Bryant. Patient satisfaction in medicine and dentistry. *International Journal of Dentistry*, 2020(1):6621848, 2020.

[49] Esther Irawati Setiawan, Patrick Tjendika, Joan Santoso, FX Ferdinandus, Gunawan Gunawan, and Kimiya Fujisawa. Aspect-based sentiment analysis of healthcare reviews from indonesian hospitals based on weighted average ensemble. *Journal of Applied Data Sciences*, 5(4):1579–1596, 2024.

[50] Avedis Donabedian. The quality of care: how can it be assessed? *Jama*, 260(12):1743–1748, 1988.

[51] David Riedl and Gerhard Schüßler. The influence of doctor-patient communication on health outcomes: a systematic review. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 63(2):131–150, 2017.

[52] Jean-Frederic Levesque, Mark F Harris, and Grant Russell. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. *International Journal for Equity in Health*, 12(1):18, 2013.

[53] Tracey S Dagger, Jillian C Sweeney, and Lester W Johnson. A hierarchical model of health service quality: scale development and investigation of an integrated model. *Journal of Service Research*, 10(2):123–142, 2007.

[54] Kathryn R Tringale and Jona A Hattangadi-Gluth. Truth, trust, and transparency—the highly complex nature of patients' perceptions of conflicts of interest in medicine. *JAMA Network Open*, 2(4):e191929–e191929, 2019.

[55] David E Newman-Toker, Najlla Nassery, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Zheyu Wang, Yuxin Zhu, Ali S Saber Tehrani, Mehdi Fanai, Ahmed Hassoon, et al. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety*, 33(2):109–120, 2024.

[56] Antonio Passaro, Maise Al Bakir, Emily G Hamilton, Maximilian Diehn, Fabrice André, Sinchita Roy-Chowdhuri, Giannis Mountzios, Ignacio I Wistuba, Charles Swanton, and Solange Peters. Cancer biomarkers: Emerging trends and clinical implications for personalized treatment. *Cell*, 187(7):1617–1635, 2024.

[57] Manuel Au-Yong-Oliveira, Antonio Pesqueira, Maria José Sousa, Francesca Dal Mas, and Mohammad Soliman. The potential of big data research in healthcare for medical doctors' learning. *Journal of Medical Systems*, 45(1):13, 2021.

[58] Edward Fernandez, Jordon Jaggers, Allison E Norton, Cosby Stone, and Elizabeth Phillips. Adverse events following vaccines: From detection to research translation. *Annual Review of Public Health*, 46, 2024.

[59] Magnhild Vikan, Arvid Steinar Haugen, Ann Kristin Bjørnnes, Berit Taraldsen Valeberg, Ellen Catharina Tveter Deilkås, and Stein Ove Danielsen. The association between patient safety culture and adverse events–a scoping review. *BMC Health Services Research*, 23(1):300, 2023.

[60] Raffaella Gualandi, Cristina Masella, Michela Piredda, Matteo Ercoli, and Daniela Tartaglini. What does the patient have to say? valuing the patient experience to improve the patient journey. *BMC Health Services Research*, 21(1):347, 2021.

[61] Ronald M Epstein and Richard L Street. The values and value of patient-centered care, 2011.

[62] Jacquelin Forsey, Stella Ng, Paula Rowland, Risa Freeman, Connie Li, and Nicole N Woods. The basic science of patient–physician communication: a critical scoping review. *Academic Medicine*, 96(11S):S109–S118, 2021.

[63] Joana Lopes, Tiago Miranda, Regina Sousa, and José Machado. Primary health care appointments and hospital stay: an impact analysis. *Procedia Computer Science*, 251:696–702, 2024.

[64] Joanne Parsons, Carol Bryce, and Helen Atherton. Which patients miss appointments with general practice and why? a systematic review. *British Journal of General Practice*, 2021.

[65] Chuanyao Li and Junren Wang. A hierarchical two-step floating catchment area analysis for high-tier hospital accessibility in an urban agglomeration region. *Journal of Transport Geography*, 102:103369, 2022.

[66] Zishan K Siddiqui, Rebecca Zuccarelli, Nowella Durkin, Albert W Wu, and Daniel J Brotman. Changes in patient satisfaction related to hospital renovation: experience with a new clinical building. *Journal of Hospital Medicine*, 10(3):165–171, 2015.

[67] Aizat Hilmi Zamzam, Ahmad Khairi Abdul Wahab, Muhammad Mokhzaini Azizan, Suresh Chandra Satapathy, Khin Wee Lai, and Khairunnisa Hasikin. A systematic review of medical equipment reliability assessment in improving the quality of healthcare services. *Frontiers in Public Health*, 9:753951, 2021.

[68] Jahanpour Alipour, Yousef Mehdipour, Afsaneh Karimi, Mohadeseh Khorashadizadeh, and Maryam Akbarpour. Security, confidentiality, privacy and patient safety in the hospital information systems from the users' perspective: A cross-sectional study. *International Journal of Medical Informatics*, 175:105066, 2023.

[69] Caroline E Sloan, Lorena Millo, Sophia Gutterman, and Peter A Ubel. Accuracy of physician estimates of out-of-pocket costs for medication filling. *JAMA network open*, 4(11):e2133188–e2133188, 2021.

[70] Traber D Giardina, Saritha Korukonda, Umber Shahid, Viralkumar Vaghani, Divvy K Upadhyay, Greg F Burke, and Hardeep Singh. Use of patient complaints to identify diagnosis-related safety concerns: a mixed-method evaluation. *BMJ Quality & Safety*, 30(12):996–1001, 2021.

[71] Angela Coulter, Louise Locock, Sue Ziebland, and Joe Calabrese. Collecting data on patient experience is not enough: they must be used to improve care. *BMJ*, 348, 2014.

[72] RE Davis, M Koutantji, and CA Vincent. How willing are patients to question healthcare staff on issues related to the quality and safety of their healthcare? an exploratory study. *BMJ Quality & Safety*, 17(2):90–96, 2008.

[73] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[74] Fei Ding, Xin Kang, and Fuji Ren. Neuro or symbolic? fine-tuned transformer with unsupervised lda topic clustering for text sentiment analysis. *IEEE Transactions on Affective Computing*, 15(2):493–507, 2023.

[75] TR Mahesh, R Sivakami, Arastu Thakur, Achyut Shankar, and Fayez Alqahtani. Fine tuned llm with lora-q for enhanced health literacy. *IEEE Transactions on Consumer Electronics*, 2025.

[76] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[77] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 380–385, 2019.

[78] Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. Lego-absa: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *International Conference on Computational Linguistics*, pages 7002–7012, 2022.

[79] Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. Exploring conditional text generation for aspect-based sentiment analysis. In *Pacific Asia Conference on Language, Information and Computation*, pages 119–129, 2021.

[80] Ali Salbas and Rasit Eren Buyuktoka. Performance of large language models in recognizing brain mri sequences: a comparative analysis of chatgpt-4o, claude 4 opus, and gemini 2.5 pro. *Diagnostics*, 15(15):1919, 2025.

[81] Maikel Leon. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620, 2025.

[82] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with gemini 3. *Google Blog, November*, 18, 2025.

[83] Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

[84] L. M. Sabik and R. K. Lie. Priority setting in health care: Lessons from the experiences of eight countries. *International Journal for Equity in Health*, 7(1):4, 2008.

[85] Linlin Hou, Wenhui Tu, Ting Yu, Ting Jiang, Mohamed Bah, Zenghui Xu, Yu Zhang, Gaoming Yang, and Ji Zhang. Aspect-based sentiment analysis for covid-19: A heterogeneous graph convolutional network

approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(6):1–26, 2025.

[86] Wei Zhang, Nian-xi Yang, Chen-guang Li, and Jing Li. Analyzing influence of epidemic policy adjustment on public concerns and emotional feedback using the absa approach. In *Wuhan International Conference on E-business*, pages 1–13. Springer, 2024.

[87] Somiya Rani and Amita Jain. Aspect-based sentiment analysis of drug reviews using multi-task learning based dual bilstm model. *Multimedia Tools and Applications*, 83(8):22473–22501, 2024.