



MEDAIDE: Information Fusion and Anatomy of Medical Intents via LLM-based Agent Collaboration

Dingkang Yang^{1,2,*†}, Jinjie Wei^{1,2,3,*}, Mingcheng Li^{1,2,*}, Jiyao Liu^{3,4}, Lihao Liu³, Ming Hu³, Junjun He³, Yakun Ju⁵, Wei Zhou⁶, Yang Liu^{7,§}, Lihua Zhang^{1,8,9,10,§}

¹College of Intelligent Robotics and Advanced Manufacturing, Fudan University ²Multimodal Foundation Research Team, CIT Lab

³Shanghai Artificial Intelligence Laboratory ⁴Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University

⁵School of Computing and Mathematical Sciences, University of Leicester ⁶Cardiff University

⁷Department of Computer Science, The University of Toronto ⁸Institute of Metaverse & Intelligent Medicine

⁹Jilin Provincial Key Laboratory of Intelligence Science and Engineering

¹⁰Engineering Research Center of AI and Robotics, Ministry of Education

*Equal contribution, †Project lead, §Corresponding Author

Abstract

In healthcare intelligence, the ability to fuse heterogeneous, multi-intent information from diverse clinical sources is fundamental to building reliable decision-making systems. Large Language Model (LLM)-driven information interaction systems currently showing potential promise in the healthcare domain. Nevertheless, they often suffer from information redundancy and coupling when dealing with complex medical intents, leading to severe hallucinations and performance bottlenecks. To this end, we propose MEDAIDE, an LLM-based medical multi-agent collaboration framework designed to enable intent-aware information fusion and coordinated reasoning across specialized healthcare domains. Specifically, we introduce a regularization-guided module that combines syntactic constraints with retrieval-augmented generation to decompose complex queries into structured representations, facilitating fine-grained clinical information fusion and intent resolution. Additionally, a dynamic intent prototype matching module is proposed to utilize dynamic prototype representation with a semantic similarity matching mechanism to achieve adaptive recognition and updating of the agent's intent in multi-round healthcare dialogues. Ultimately, we design a rotation agent collaboration mechanism that introduces dynamic role rotation and decision-level information fusion across specialized medical agents. Extensive experiments are conducted on four medical benchmarks with composite intents. Experimental results from automated metrics and expert doctor evaluations show that MEDAIDE outperforms current LLMs and improves their medical proficiency and strategic reasoning.

Date: July 4, 2025

Corresponding: dicken@fysics.ai

Project Page: <https://github.com/ydk122024/MedaIDE>



1 Introduction

The emergence of information fusion technologies has revolutionized healthcare systems by enabling the integration of multi-source medical data for enhanced decision-making capabilities [33]. Contemporary multi-source information fusion approaches have proven crucial for processing diverse healthcare data modalities, enabling comprehensive patient

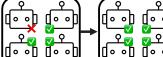
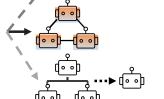
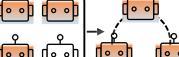
Method	Single	MedAgents	MDAgents	MEDAIDE (Ours)
Interaction Type				
Multiple Roles	✗	✓	✓	✓
Query Decomposition	✗	✗	✗	✓
Multi-party Chat	✗	✗	✓	✓
Complexity Check	✗	✗	✓	✓
Intention Extraction	✗	✗	✗	✓

Figure 1 Comparison between our framework and previous methods. Among these works, MEDAIDE is the only one with intent awareness and comprehensive analysis capabilities.

assessment and personalized treatment strategies [23]. In the healthcare domain, the integration of artificial intelligence with information fusion techniques has demonstrated significant potential for smart healthcare applications [3]. The advancement of multimodal medical signal fusion has played a crucial role in smart healthcare systems, enabling real-time analysis and providing medical professionals with enhanced insights for disease detection and prediction accuracy [27]. Furthermore, multi-modal lifelog data fusion approaches have shown promising results in improving human activity recognition and health monitoring applications [15]. Advanced deep learning technologies [20, 25, 26, 29, 30, 35, 37, 40] have shown exceptional generalization capabilities in general scenarios for multiple purposes within human-machine interactions. Developing goal-oriented conversation systems [8] has received increasing attention in recent years, particularly in medical contexts where information fusion enables sophisticated reasoning across heterogeneous data sources. In this context, LLM-centered interactive medical assistants [7, 40, 41] have become research hotspots that promise to improve diagnosis efficiency and promote service automation. The advancement of multimodal medical signal fusion has played a crucial role in smart healthcare systems, enabling real-time analysis and providing medical professionals with enhanced insights for disease detection and prediction accuracy [27]. Previous attempts have infused LLMs with healthcare-specific knowledge through relevant corpus construction and multi-stage training procedures . However, traditional approaches often overlook the potential of multi-source information fusion for achieving a comprehensive understanding of complex medical scenarios [15]. Although these strategies enhance the models’ comprehension of medically relevant intents, bottlenecks remain when faced with real-world applications that require sophisticated reasoning and accurate feedback, particularly in scenarios that demand the integration of diverse information sources for optimal clinical decision-making.

Given the mimicking of learned feedback in human behaviors, automated medical agent construction promises to enhance LLMs’ instruction following and logical profiling capabilities. The collaboration among multiple agents to handle different patient inquiries and symptomatic caseloads facilitates accurate dialog goal fulfillment while accounting for individual differences [12]. Despite significant advancements, current efforts focus primarily on medical education training [36] or selective question-answering [34], lacking a comprehensive understanding when dealing with the sophisticated intents behind user queries in real-world diagnosis and treatment scenarios. In addition, the agents in most collaborative frameworks are usually familiar with only limited intents, making it difficult to provide systematic recommendations.

To address these issues, we propose MEDAIDE, an LLM-based medical multi-agent collaboration framework designed to enable intent-aware information fusion and coordinated reasoning across specialized healthcare domains. There are three core contributions in MEDAIDE based on the tailored components. Specifically, we propose a regularization-guided information extraction module, which fuses syntactic constraints with retrieval-augmented generation to transform compositional medical queries into structured semantic representations. This structured decomposition enhances the granularity and reliability of clinical knowledge extraction, facilitating downstream intent interpretation. In addition, a dynamic intent prototype matching mechanism is introduced to adaptively recognize and update agent intent through semantic similarity-based alignment during multi-turn healthcare dialogues. Ultimately, we design a rotation-based agent collaboration mechanism that enables dynamic role-switching and decision-level information fusion across agents, further

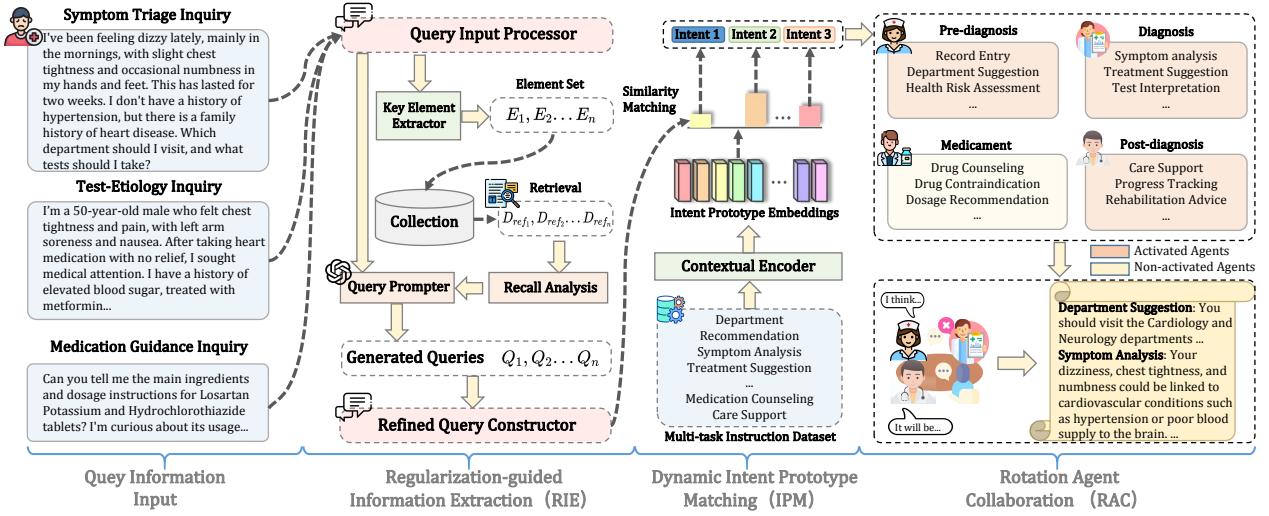


Figure 2 Illustration of the MEDAIDE framework. MEDAIDE consists of three major phases: (i) Regularization-guided Information Extraction (RIE) employs syntactic parsing and retrieval-augmented generation to reformulate complex medical queries and extract clinical elements. (ii) Dynamic Intent Prototype Matching (IPM) utilizes a BioBERT-based encoder to generate intent embeddings for precise classification across 17 medical intent categories. (iii) Rotation Agent Collaboration (RAC) implements a polling-based mechanism where specialized agents (pre-diagnosis, diagnosis, medicament, post-diagnosis) dynamically rotate as main contacts, coordinating multi-source information fusion through cross-agent knowledge integration.

ensuring coherent and globally optimized decision-making. Experimental results on four medical benchmarks with composite intents, evaluated by both automated metrics and expert physicians, demonstrate that MEDAIDE significantly improves medical reasoning accuracy, intent alignment, and agent coordination over strong LLM baselines.

2 Related Work

2.1 LLM-enhanced Information Fusion

The emergence of Large Language Models (LLMs) has introduced revolutionary paradigms for information fusion, transforming the integration and processing of heterogeneous data sources. Recent advancements demonstrate that LLMs serve as sophisticated fusion engines capable of harmonizing diverse information types and enabling semantic understanding across different data domains [28]. Contemporary research has demonstrated several key approaches to LLM-based information fusion, including context-based fine-tuning frameworks that leverage information fusion principles to enhance model adaptability while preserving pre-trained knowledge [28]. Novel techniques such as knowledge distillation and attention-based fusion mechanisms enable seamless integration of heterogeneous information sources into LLMs, representing significant advancement in intelligent data processing capabilities [43]. Various application domains have particularly benefited from LLM-enhanced information fusion approaches, where LLMs demonstrate exceptional capabilities in understanding, reasoning, and generation, introducing transformative paradigms for integrating artificial intelligence into complex decision-making systems [39]. These developments encompass fusion strategies operating at various architectural levels, from feature-level integration to decision-level fusion, with emerging hybrid approaches that leverage the representational power of LLMs to bridge semantic gaps effectively.

2.2 LLMs in Healthcare Domains

Large Language Models (LLMs), exemplified by ChatGPT [29], demonstrate excellent performance in multidisciplinary applications. Although current LLMs [1, 2] with certain medical knowledge benefit from large-scale corpus support, they lack specialized medical proficiency and have significant performance bottlenecks in domain-specific scenarios. Recently, several attempts [6, 7, 40] have begun to build medically customized LLM assistants to fulfill the diagnostic and consultative demands. For instance, HuatuoGPT series [6] have shown promising results in bridging generic-medical knowledge gaps by absorbing real doctor-patient conversations. ZhongJing [41] improves Chinese medical capabilities by introducing expert feedback and multi-round medical instructions. In addition, PediatricsGPT [40] proposes a systematic

Algorithm 1 Syntactic Parsing Algorithm

Require: Input query Q
Ensure: Parsed syntactic tree T

Step 1: Tokenize the query
Let $Q = w_1, w_2, \dots, w_n$ be the sequence of words (tokens) in the query.

Step 2: Initialize parsing
Initialize an empty syntactic tree $T = \emptyset$.

Step 3: Apply CFG (Context-Free Grammar)
for each non-terminal symbol A in grammar G **do**
 Find a rule $A \rightarrow \alpha$, where α is a string of terminals or non-terminals.
 if $A \rightarrow \alpha$ matches a substring of Q **then**
 Add this production rule to the tree T .
 Replace the matched string with non-terminal A .
 end if
end for

Step 4: Recursion
while there are still non-terminals to be parsed **do**
 Repeat step 3 until T covers the entire query Q .
end while

Step 5: Output the syntactic tree
return T

training framework to construct interactive healthcare systems for pediatric specialists and medical generalists. Unlike previous studies, our framework aims to more fully recognize medical intents and refine the models' reasoning abilities in complex scenarios through the LLM-based multi-agent collaboration.

2.3 LLM-based Multi-agent Collaboration

With the focus of researchers on sophisticated goal-oriented dialog generation [12], the inherent dilemmas of hallucinatory responses and weak comprehension in LLMs have been gradually exposed. In this context, LLM-based automated agents are proposed to provide effective perception and decision-making skilfulness by incorporating external tools and databases [22]. By mimicking human behavioral logic, multiple agents perform feedback and collaboration to enhance diverse intent understanding tasks, including educational training [36] and emotional comfort [8]. For example, MEDCO [36] enables LLMs to simulate interactions between patients and doctors, enhancing the practice performance of virtual students in interactive environments. MedAgents [34] improves the performance of medical assistants in zero-shot settings through the role-playing strategy. MDAgents [18] flexibly select the corresponding processing method according to the complexity of the inquiry. A detailed comparison between our framework and previous methods is provided in Figure 1. In comparison, the proposed MEDAIDE focuses more on mining profound healthcare intents and moving towards robust healthcare practices.

3 Methodology

Figure 2 illustrates the overall architecture of the proposed MEDAIDE framework. The workflow comprises three key components: Regularization-guided Information Extraction (RIE) module, dynamic Intent Prototype Matching (IPM) module, and Rotation Agent Collaboration (RAC) mechanism.

3.1 Regularization-guided Information Extraction

Due to the inherent complexity and ambiguity of natural language expressions in medical contexts, the precise identification of complex medical intents and information extraction presents significant challenges [33]. Previous works have utilized rule-based approaches [5] and traditional natural language processing techniques to achieve certain levels of medical text understanding, but they suffer from the following limitations: (1) inadequate handling of composite medical intents that contain multiple interconnected clinical concerns; (2) insufficient semantic understanding of domain-specific medical terminology and contextual relationships. To address these challenges, we propose a Regularization-guided

Algorithm 2 LLM-Based Input Standardization

Input: User query Q_{input} , Large Language Model \mathcal{M} , Set of rules \mathcal{R}
Output: Standardized query Q_{std}
Initialize: $Q_{cur} \leftarrow Q_{input}$,
 $converged \leftarrow \text{False}$
while not $converged$ **do**
 $converged \leftarrow \text{True}$
 for each rule $r \in \mathcal{R}$ **do**
 $Q_{new} \leftarrow \mathcal{M}(Q_{cur}, r)$
 if $Q_{new} \neq Q_{cur}$ **then**
 $Q_{cur} \leftarrow Q_{new}$
 $converged \leftarrow \text{False}$
 end if
 end for
 end while
return $Q_{std} \leftarrow Q_{cur}$

Information Extraction (RIE) module that leverages syntactic regularization algorithms combined with large language models to achieve robust medical query understanding and intent decomposition.

In this phase, we first process the initial query through the query input processor. The processor is based on a set of syntactic regularization algorithms that combine LLMs with a predefined ruleset. In this case, the core concept of RIE is to check, optimize, and normalize user input to Q_{std} . Algorithm 1 shows the corresponding program. First, the algorithm disambiguates the input query and divides the query into individual words or tokens. Next, the algorithm initializes an empty syntax tree and matches substrings in the query by context-free grammar rules, adding the conforming rules to the syntax tree. The algorithm keeps applying the rules recursively until the syntax tree covers the entire query and finally outputs the complete syntax tree.

Then, a key element extractor distils critical information from Q_{std} , such as symptoms, condition descriptions, and medical histories, to form the element set E_i . In the Retrieval-Augmented Generation (RAG), we build an indexed database of 1,095 expert-reviewed medical guidelines and retrieve documents related to these elements using a semantic retrieval method [21], forming the document set D_{ref} . Subsequently, the recall analysis module identifies relevant documents through prompt-guided selection based on BERT-Score [44], which is fed into the LLM-based query prompter along with Q_{std} . The prompter optimizes the empirical information and efficient decomposition of composite intents. Eventually, we design a refined query constructor to merge and integrate the generated subqueries. To ensure that the generated subqueries maintain high quality in both semantic meaning and formal structure, we implement a set of rules (as shown in Figure 3) in the refined query constructor using Algorithm 2. These rules encompass subquery filtering, semantic overlap removal, query consolidation, grammatical normalization, intent prioritization, and format standardization. Specifically, the constructor filters and reconstructs multiple subqueries Q_{gen} with these predefined rulesets to ensure semantically complete and formally uniform output. These processes streamline the output, optimize the query structure, and prioritize critical medical intents, ensuring that the final query is semantically coherent and structurally consistent.

3.2 Dynamic Intent Prototype Matching

Medical intent recognition in healthcare systems faces significant challenges due to the diverse expressions of similar clinical concerns and the subtle semantic differences between related medical intents [23, 27]. Previous approaches for intent classification have employed traditional machine learning methods [45], achieving reasonable performance in general domains but exhibiting limitations in medical contexts: (1) insufficient capture of domain-specific semantic nuances in medical terminology and clinical expressions; (2) inadequate handling of fine-grained intent distinctions that are crucial for accurate medical decision-making [42]. To address these challenges, we propose a dynamic Intent Prototype Matching (IPM) module that leverages contextualized medical embeddings to achieve robust semantic understanding and precise intent classification in healthcare applications.

After the RIE module, the optimized query Q_{opt} is matched with a set of intent prototype embeddings E_i generated by

Rule	Description
Subquery Filtering	Only retain subqueries that are relevant to the user's intent, discarding unrelated ones.
Semantic Overlap Removal	Remove redundant subqueries by detecting overlapping information to streamline the output.
Query Consolidation	Merge subqueries with similar content or meaning into a single, coherent query.
Grammatical Normalization	Maintain grammatical consistency across merged subqueries for natural language flow.
Intent Prioritization	Prioritize subqueries by medical context and urgency, processing critical intents first.
Format Standardization	Standardize query structure to follow consistent formatting, including punctuation and phrasing

Figure 3 Illustration of the rule sets.

Models	BLEU-1	BLEU-2	Meteor	BERT-Score	ROUGE-L	GLEU
ZhongJing/ + MEDAIDE	10.24/ 12.44	3.74/ 4.56	14.51/ 15.78	47.54/ 48.06	11.12/ 12.31	3.87/ 4.69
Meditron-7B/ + MEDAIDE	3.76/ 4.94	0.91/ 1.30	6.63/ 8.51	42.64/ 42.75	4.52/ 5.21	1.74/ 2.07
HuatuoGPT-II/ + MEDAIDE	14.6/ 15.18	5.87/ 6.63	18.81/ 21.84	49.26/ 51.88	13.57/ 19.04	6.43/ 7.25
Baichuan4/ + MEDAIDE	12.97/ 15.95	5.39/ 7.85	16.68/ 21.60	48.68/ 52.69	13.33/ 20.75	5.57/ 7.98
LLama-3.1-8B/ + MEDAIDE	10.95/ 15.36	2.13/ 6.78	15.47/ 19.54	47.62/ 51.41	11.45/ 19.29	4.37/ 6.45
GPT-4o/ + MEDAIDE	15.28/ 15.93	6.33/ 7.56	19.89/ 21.78	51.95/ 53.65	14.15/ 17.64	6.34/ 7.31
Claude 3.7 Sonnet/ + MEDAIDE	15.67/ 16.42	6.55/ 7.89	20.31/ 22.47	52.34/ 57.88	14.62/ 18.21	6.51/ 7.65
DeepSeek-R1/ + MEDAIDE	16.03/ 16.78	6.82/ 8.14	20.85/ 23.15	52.86/ 54.92	15.04/ 18.73	6.73/ 7.93

Table 1 Comparison results on the Pre-Diagnosis benchmark.

a contextual encoder, which is designed to capture the semantic features of different medical intents. The encoder is constructed on top of BioBERT [19] to learn prototype representations by performing fine-grained intent classification. We add a fully connected layer after the embedding layer with output dimensions aligned to the 17 medical intent categories and generate the corresponding probability distributions via a softmax activation function. Specifically, the contextual encoder maps the optimized query together with the intent embeddings into a 768-dimensional embedding space. It computes the cosine similarity S_{ij} between the query and each intent embedding E_i with the following formula:

$$S_{ij} = \frac{Q_{opt} \cdot E_i}{\|Q_{opt}\| \|E_i\|}. \quad (1)$$

Subsequently, the probability distribution α_{ij} for each intent after the softmax is expressed as:

$$\alpha_{ij} = \frac{\exp(S_{ij})}{\sum_{l=1}^{17} \exp(S_{il})}. \quad (2)$$

If the probability α_{ij} of an intent i exceeds a predetermined threshold, the intent will be activated, triggering the corresponding agent:

$$\text{Activated Intent}_i = \begin{cases} 1 & \text{if } \alpha_{ij} > \text{Threshold}, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In this way, our framework can automatically activate the most compliant medical intent based on the optimized query, directing it to the corresponding agent to perform subsequent operations.

Models	BLEU-1	BLEU-2	Meteor	BERT-Score	ROUGE-L	GLEU
ZhongJing/ + MEDAIDE	11.21/ 12.33	5.12/ 6.25	26.05/ 27.13	5.23/ 5.67	11.34/10.87	6.14/5.49
Meditron-7B/ + MEDAIDE	12.01/ 13.93	3.13/ 6.03	19.10/ 26.53	3.60/ 6.12	8.60/ 11.67	3.99/ 6.48
HuatuoGPT-II/ + MEDAIDE	22.65/ 25.26	10.21/ 12.11	34.51/ 40.32	8.36/ 11.73	15.77/ 19.11	8.95/ 10.65
Baichuan4/ + MEDAIDE	22.58/ 25.51	10.98/ 13.61	36.26/ 42.77	10.30/ 15.00	17.03/ 22.43	9.61/ 11.95
LLama-3.1-8B/ + MEDAIDE	16.92/ 22.75	6.56/ 9.97	26.39/ 38.52	4.78/ 10.41	12.79/ 18.68	6.21/ 8.94
GPT-4o/ + MEDAIDE	21.67/ 26.80	9.73/ 14.56	32.67/ 44.11	7.95/ 15.81	16.20/ 23.67	8.41/ 12.69
Claude 3.7 Sonnet/ + MEDAIDE	22.14/ 27.41	10.36/ 15.02	33.85/ 45.23	8.64/ 16.32	16.78/ 24.14	8.92/ 13.08
DeepSeek-R1/ + MEDAIDE	22.83/ 28.16	10.89/ 15.67	34.91/ 46.54	9.23/ 17.05	17.45/ 24.88	9.38/ 13.52

Table 2 Comparison results on the Diagnosis benchmark.

3.3 Rotation Agent Collaboration

Effective multi-agent collaboration in healthcare systems requires sophisticated coordination mechanisms to ensure coherent and comprehensive patient care. Traditional multi-agent approaches often suffer from information fragmentation and a lack of systematic integration across different medical specialties [10]. We propose a Rotation Agent Collaboration (RAC) mechanism that systematically coordinates agents through sequential leadership transitions, ensuring comprehensive coverage of healthcare services from pre-diagnosis to post-treatment.

The multi-agent collaboration framework operates through the dynamic activation of specialized agents based on IPM. Our polling-based information fusion mechanism enables systematic knowledge integration across four distinct medical domains: pre-diagnosis assessment, diagnostic reasoning, medication management, and post-treatment care. Each domain has a designated primary contact agent that coordinates the fusion process for domain-specific decision-making.

Pre-Diagnosis Assessment with Multi-Source Fusion. When pre-diagnosis intents are detected, the pre-diagnosis agent serves as the primary contact for orchestrating the synthesis of patient information. The agent maintains a relational database repository [9] and implements cross-agent polling to fuse comprehensive patient profiles. The fusion mechanism integrates the diagnostic agent output for the preliminary interpretation of symptoms with the analysis of the medication agent’s patterns of medication history. This multi-source approach enables enhanced risk stratification and informed departmental routing decisions through the systematic integration of knowledge.

Diagnostic Reasoning through Hybrid Knowledge Fusion. For diagnostic intents, the diagnosis agent serves as the main contact, coordinating information fusion across 506 high-quality medical cases from the medical record database [12]. The agent employs a hybrid retrieval-fusion mechanism [32] enhanced by cross-agent knowledge integration. The polling process incorporates pre-diagnosis agent outputs for patient historical context and medicament agent analyses for drug-symptom interaction patterns.

The hybrid information retrieval operates through parallel keyword and semantic fusion channels. The keyword-based channel extracts document subsets D_{slice} through lexical matching:

$$D_{\text{slice}} = \{d \in D \mid \text{KeywordMatch}(Q, d) = \text{True}\}. \quad (4)$$

The semantic channel computes similarity scores $S(Q, d)$ between queries and documents using embedding-based fusion:

$$S(Q, d) = \frac{E_Q \cdot E_D}{\|E_Q\| \|E_D\|}, \quad (5)$$

$$D_{\text{match}} = \{d \in D \mid S(Q, d) > \tau\}. \quad (6)$$

The final information fusion combines both retrieval channels:

$$D_{\text{final}} = D_{\text{slice}} \cup D_{\text{match}}. \quad (7)$$

Components	Prompt Template
Key Element Extractor	Analyze the user's input, which may include symptoms, condition descriptions, medical history, and other relevant health details. Identify the key elements essential for further diagnostic analysis. Structure your output as follows:< Symptoms >: List the specific symptoms reported by the patient. < Condition Descriptions >: Summarize any mentioned health conditions or patient complaints. < Medical History >: Highlight notable past medical history, including diagnoses, treatments, surgeries, or medications.< Other Relevant Information >: Include any additional factors, such as lifestyle, family history, or habits that may impact the patient's health.
Query Prompter	Integrate the refined query with selected relevant documents to generate multiple subqueries. Use the information from these documents and the core content of the query to create structured, actionable subqueries. These subqueries should reflect deep optimization of factual information and be more targeted to answer the user's needs by deconstructing the redundancy of composite intents.

Figure 4 Illustration of the prompt template for the RIE.

Medication Management via Cross-Agent Integration. When medicament intents are activated, the medicament agent coordinates knowledge fusion across 26,684 medication entries from PubMed [13]. The polling-based integration mechanism fuses diagnostic agent outputs for confirmed conditions with pre-diagnosis agent data regarding allergy histories and contraindications. This multi-source fusion enables dynamic dosage optimization and interaction detection tailored to individual patient profiles through systematic information synthesis.

Post-Treatment care through Clinical Data Integration. For post-diagnosis intents involving care support and rehabilitation advice, the post-diagnosis agent coordinates information synthesis from clinical decision outputs. The integration process fuses diagnostic agent prognosis data with medicament agent adherence to develop personalized rehabilitation protocols and recovery milestone frameworks. This approach ensures continuity between active treatment and long-term care management through comprehensive data fusion.

Polling-Based Coordination Protocol. The polling mechanism ensures systematic information flow between main contact transitions. For each stage $s \in \{1, 2, 3, 4\}$, the main contact agent A_s^{mc} coordinates with supporting agents $\{A_1, A_2, \dots, A_n\} \setminus A_s^{mc}$:

$$O_s^{(0)} = A_s^{mc}(P_s, D_{\text{input}}^{(s)}), \quad (8)$$

where P_s represents the stage-specific prompt and $D_{\text{input}}^{(s)}$ includes both patient data and outputs from previous stages. Each supporting agent contributes specialized knowledge:

$$C_i^{(s)} = A_i(P_s, D_{\text{input}}^{(s)}, O_s^{(0)}), \quad i \neq s, \quad (9)$$

The main contact agent then integrates all contributions to produce the stage output, and the final comprehensive output combines all stage results:

$$O_s^{final} = \text{Integrate}_s(O_s^{(0)}, \{C_i^{(s)} \mid i \neq s\}), \quad (10)$$

$$\text{Output}_{\text{final}} = \text{Synthesize}(O_1^{final}, O_2^{final}, O_3^{final}, O_4^{final}). \quad (11)$$

This polling-based approach ensures that each healthcare stage receives focused attention from the most qualified agent while maintaining systematic integration across the entire care continuum.

4 Experiments

4.1 Datasets and Implementation Details

All experiments are conducted in the zero-shot setting. To reflect the real-world demands for healthcare services, we conduct four different benchmarks across 17 types of medical intents, including Pre-Diagnosis, Diagnosis, Medicament, and Post-Diagnosis benchmarks. Each benchmark contains 500 composite intent instances. We employ structured prompt templates across four distinct agents within our framework. Each prompt is designed to guide the language model in generating contextually relevant and accurate responses tailored to specific medical tasks. These templates are instrumental in ensuring that the interactions between the agents and users are effective and medically informative. The

Dataset	Intent	Prompt Template
Pre-Diagnosis	Department Suggestion	Please recommend the appropriate department and suggest appropriate tests for the user based on their described symptoms and medical history. Please provide your suggestions according to the following requirements:1. Department Recommendation: Specify the department to be registered and explain the reason for the recommendation.2. Priority Tests: List the priority tests for each department and briefly explain their purpose.
	Health Risk Assessment	You are an experienced medical assistant specializing in health risk assessment, guiding users to identify potential health risks based on their symptoms, lifestyle and medical history. Please perform the assessment according to the following requirements:1. Risk Assessment: Evaluates the user's overall health risk based on the information provided and identifies any potential problems or red flags. Briefly explain the factors contributing to the risk.2. Preventive Measures: Recommend lifestyle changes, habits, or preventive strategies to help reduce the identified risks. Explain how each advice could positively impact the user's health. 3. Recommendations for Further Evaluation: List any tests or screenings needed for further evaluation based on the user's risk factors. Include the purpose of each recommended test and its relevance to the identified risk.

Figure 5 Illustration of the prompt template for the Pre-Diagnosis.

Models	BLEU-1	BLEU-2	Meteor	BERT-Score	ROUGE-L	GLEU
ZhongJing/ + MEDAIDE	13.14/ 13.87	5.02/ 5.34	22.68/ 25.09	5.48/ 5.99	14.01/ 14.85	6.02/ 6.43
Meditron-7B/ + MEDAIDE	5.76/ 19.43	2.04/ 17.50	12.29/ 34.84	2.12/ 29.43	6.31/ 32.57	2.79/ 24.47
HuatuoGPT-II/ + MEDAIDE	17.14/ 37.89	9.61/ 27.47	30.72/ 58.06	11.96/ 33.62	19.96/ 45.39	9.57/ 24.23
Baichuan4/ + MEDAIDE	14.64/ 44.63	7.12/ 36.34	26.26/ 62.11	8.09/ 42.56	15.47/ 49.91	7.08/ 35.99
LLama-3.1-8B/ + MEDAIDE	13.80/ 23.87	5.90/ 16.75	27.28/ 47.39	7.50/ 26.75	16.05/ 38.30	6.37/ 15.77
GPT-4o/ + MEDAIDE	16.23/ 45.13	8.16/ 38.96	29.24/ 65.09	9.78/ 40.53	17.53/ 52.71	8.61/ 37.09
Claude 3.7 Sonnet/ + MEDAIDE	16.88/ 46.24	8.67/ 39.85	30.41/ 66.27	10.52/ 41.86	18.24/ 53.94	9.15/ 38.16
DeepSeek-R1/ + MEDAIDE	17.42/ 47.38	9.21/ 40.92	31.63/ 67.83	11.27/ 43.15	19.06/ 55.18	9.74/ 39.28

Table 3 Comparison results on the Medicament benchmark.

prompt templates for the various components of the RIE phase are detailed in Figure 4. The specifics for each agent’s prompt templates are depicted in Figures 5, 6, 7, and 8.

To enable fine-grained intent prototype matching, we develop a contextual encoder training dataset consisting of 2,800 training samples, 300 validation samples, and 300 testing samples. For the medical intent classification, we use the BioBERT-based contextual encoder from the HuggingFace library¹ to map the input sentences to a 768-dimensional embedding space. For each intent classification task, we add a fully connected layer after the embedding layer with an output dimension of 17 and generate probability distributions corresponding to the 17 intents via a softmax activation function. To handle these tasks, we deploy four V100 GPUs (32 GB VRAM each) to provide sufficient computational resources and graphics memory. The Adam optimizer is used in the training process with an initial learning rate of 1e-6 and 5 training rounds, and a learning rate scheduling strategy is applied in the validation set for dynamic adjustment. The loss function is the cross-entropy loss function. We evaluate the model using accuracy and F1 score to ensure its performance on the validation set.

4.2 Model Zoo

We compare a series of state-of-the-art (SOTA) models for comprehensive evaluations. Among medical LLMs, **ZhongJing** [41] is a traditional Chinese medicine model based on the Ziya-LLaMA model [14] with complete training procedures. **Meditron-7B** [7] performs medically relevant continuous pre-training on Llama-2-7B to extend the breadth and depth of the model in medical knowledge. **HuatuoGPT-II (13B)** [6] employs a one-stage unified approach for domain

¹<https://huggingface.co/dmis-lab/biobert-v1.1>

Dataset	Intent	Prompt Template
Diagnosis	Symptom Analysis	Generate a symptom analysis report based on the patient's description. The report should include: 1. Preliminary Diagnosis: Based on the patient's medical history, symptoms, physical signs, and test results. 2. Diagnostic Basis: Provide a detailed justification using the patient's history, symptoms, signs, and test results. 3. Differential Diagnosis: Identify conditions similar to the diagnosis and describe how to distinguish them using clinical and test findings.
	Treatment Suggestion	Please generate a treatment recommendation report based on the patient's detailed history and diagnosis. The report should contain the following: 1. Recommended treatment plan: Based on the current diagnosis, provide specific treatment recommendations, including medication, surgery, or other treatments. 2. Treatment purpose: Specify the purpose and expected effect of each treatment measure. 3. Potential risks and side effects: List the risks and side effects that may be associated with the treatment, as well as how risk management and monitoring will be carried out.
	Etiology Detection	Generate a cause-specific test report based on the patient's complaints and medical history information. The report should include 1. Possible etiologies: provide a comprehensive list of potential etiologies describing the symptoms, with a detailed description of each etiology. 2. Disease screening: identify and list diseases that need to be ruled out or potential complications associated with the initial diagnosis, and clearly state the reasons for considering or ruling out these diseases.
	Second Opinion	Please provide a structured second opinion by evaluating the patient's current diagnosis against their comprehensive medical history and initial diagnostic findings. This systematic review is intended to confirm or adjust the initial diagnosis, thus supporting healthcare providers and patients in making informed medical decisions.
	Test Interpretation	Using the detailed medical examination results provided, generate a comprehensive interpretation that includes the following components: 1. Explanation of Findings: Describe the significance of key results from the medical examinations, detailing what each finding suggests about the patient's health. 2. Implications for Diagnosis: Relate the examination results to potential medical conditions, explaining how specific findings contribute to or rule out certain diagnoses. 3. Recommendations for Further Testing: If applicable, suggest additional tests or follow-up procedures that might be necessary to provide more clarity on ambiguous results.
	Prevention Guidelines	Based on the patient's current health status, medical history, and lifestyle, generate a detailed set of prevention guidelines that aim to mitigate the risk of future health issues. Ensure your response includes the following elements: 1. Risk Factors Identification: Highlight key risk factors specific to the patient, based on their medical and family history, lifestyle, and any existing conditions. 2. Lifestyle Recommendations: Provide actionable advice on lifestyle modifications such as diet, exercise, and stress management that are tailored to reduce the identified risks. 3. Monitoring and Screening: Suggest appropriate regular screening tests and health monitoring practices to detect early signs of potential health issues. 4. Vaccination and Prophylaxis: Recommend any relevant vaccinations or prophylactic treatments that can prevent the onset of diseases the patient may be particularly susceptible to.

Figure 6 Illustration of the prompt template for the Diagnosis.

adaptation to improve medical expertise. **Baichuan4** [2] optimizes long texts by scaling law and combines reinforcement learning techniques to improve reasoning ability with enhanced medical knowledge representation. In general-purpose models, **Llama-3.1-8B** [11] relies on the grouped query attention mechanism to enhance the inference efficiency and performs well on multilingual tasks. **GPT-4o** [17] shows excellent language comprehension and generation capabilities and excels in handling complex tasks. **Claude 3.7 Sonnet** [4] is designed for advanced reasoning with balanced performance and speed, featuring enhanced contextual understanding capabilities. **DeepSeek-R1** [16] incorporates advanced multi-stage reasoning techniques with recursive refinement, demonstrating superior performance on complex reasoning tasks and knowledge integration.

4.3 Comparison with SOTA Methods

As a plug-and-play framework, we combine MEDAIDE with the baseline models to provide comprehensive evaluations by different metrics, including BLEU-1/2 (%) [31], ROUGE-1/2/L (%) [24], GLEU (%) [44], Meteor (%) [44], BERT-Score (%) [38].

Results on Pre-Diagnosis Benchmark. As Table 1 shows, our framework consistently improves the performance across metrics for all models. (i) With MEDAIDE support, HuatuoGPT-II and Baichuan4 achieve significant relative improvements in BERT score, indicating stronger diagnostic semantic alignment. (ii) Additionally, larger frontier models like DeepSeek-R1 and Claude 3.7 Sonnet show substantial improvements with MEDAIDE, with DeepSeek-R1 showing an 11.0% increase in Meteor scores and Claude 3.7 Sonnet achieving a 10.6% improvement in BERT-Score, highlighting the framework's effectiveness even with SOTA models. (iii) Furthermore, MEDAIDE helps general LLama-3.1-8B to improve content quality with a notable Meteor score increase from 15.47% to 19.54%, demonstrating better precision in medical terminology and symptom descriptions. (iv) Despite the performance constraints of ZhongJing and Meditron-7B in tackling complex medical scenarios due to the scaling law, our framework still provides performance gains across all evaluated metrics.

Results on Diagnosis Benchmark. Table 2 shows the diagnosis task results containing six composite intent scenarios. (i) Combined with our framework, Meditron-7B and HuatuoGPT-II outperform vanilla baselines in response accuracy and completeness, as evidenced by significant improvements across metrics. Meditron-7B shows a remarkable 70.0% improvement in BERT-Score and a 38.8% increase in Meteor scores, indicating MEDAIDE's ability to enhance specialized

Dataset	Intent	Prompt Template
Medicament	Drug Counseling	Provide comprehensive information about a specific medication, detailing its uses, benefits, and mechanisms of action. Explain how this medication interacts with other treatments, and outline potential side effects and how they can be managed. Ensure the patient understands why this medication is prescribed, its role in their overall treatment plan, and any necessary lifestyle adjustments or monitoring that should accompany its use.
	Dosage Recommendation	Based on the patient's medical history, current condition, and any concurrent medications, recommend an appropriate dosage for the prescribed drug. Include details on dosage adjustments for specific populations, such as the elderly or those with renal impairment.
	Drug Interaction	Identify and describe potential interactions between the prescribed medication and other drugs the patient is currently taking. Include interactions with over-the-counter drugs, herbal supplements, and food that could affect the drug's efficacy or safety.
	Drug Contraindication	List and explain any contraindications for the medication in question. Detail conditions or factors (like allergies or pregnancy) that would prohibit the use of the drug, emphasizing the risks associated with these contraindications.
	Adverse Effects	Detail the common and serious adverse effects associated with the medication. Provide guidance on how to manage these effects and under what circumstances the patient should seek immediate medical attention.
	Optimal Timing	Please provide the advise on the best time of day to take the prescribed medication to maximize its effectiveness and minimize side effects. Include considerations for interactions with food or other medications that could influence the timing.

Figure 7 Illustration of the prompt template for the Medicament.

Dataset	Intent	Prompt Template
Post-Diagnosis	Rehabilitation Advice	Generate a personalized post-operative rehabilitation plan based on the patient's specific surgery details and current health status. Include recommendations for physical activities, dietary suggestions, and any necessary lifestyle adjustments. Highlight the importance of each element in the recovery process and any signs that might indicate a need for medical reassessment.
	Progress Tracking	Create a detailed framework for monitoring the patient's recovery from a specific condition or post-surgical recovery. List the key health indicators that need regular monitoring, the frequency of assessments, and the criteria for evaluating progress. Explain how variations in these indicators might inform changes in treatment or care strategies, ensuring the patient understands the importance of each metric in their overall recovery.
	Care Support	Provide a comprehensive support plan for a patient, focusing on both physical and emotional aspects of healthcare. Detail support resources available, such as counseling services, support groups, and family engagement strategies. Discuss how these resources can be integrated into the patient's daily life to improve their overall well-being and assist in managing any health-related stress or anxiety.

Figure 8 Illustration of the prompt template for the Post-Diagnosis.

medical knowledge in smaller models. (ii) Among frontier models, DeepSeek-R1 with MEDAIDE achieves the highest performance with Meteor scores of 46.54% and BERT-Score of 17.05%, outperforming both Claude 3.7 Sonnet and GPT-4o. This suggests that DeepSeek-R1's superior capability in medical knowledge integration is enhanced when augmented with our structured reasoning framework. (iii) We observe slight performance drops in ZhongJing on ROUGE-L and GLEU, likely due to its limited capacity to handle the RIE phase needed to integrate multifaceted knowledge from retrieval-augmented generation, resulting in sub-optimal prevention guidelines and second opinions.

Results on Medicament Benchmark. Table 3 provides the results of the study of different models on the intentions of the composition of the drugs. (i) Attributed to the drug knowledge injection and prototype-guided embedding in MEDAIDE, the reliability of medication counselling responses is consistently improved across all baselines. DeepSeek-R1 demonstrates significant performance improvements when augmented with MEDAIDE, with its Meteor score increasing from 31.63% to 67.83% and its BERT-Score rising from 11.27% to 43.15%, showcasing its enhanced ability to integrate structured medication knowledge with recursive reasoning. (ii) Similarly, Baichuan4 achieves a substantial boost, with its GLEU score improving from 7.08% to 35.99%, while Claude 3.7 Sonnet shows strong progress in BERT-Score, increasing from 10.52% to 41.86%, highlighting MEDAIDE's effectiveness in improving drug information comprehension

Models	BLEU-1	BLEU-2	Meteor	BERT-Score	ROUGE-L	GLEU
ZhongJing/ + MEDAIDE	13.03/ 17.83	5.08/ 5.94	26.22/ 32.14	4.23/ 5.22	13.48/ 15.34	6.21/ 7.59
Meditron-7B/ + MEDAIDE	8.72/ 12.78	2.46/ 3.06	18.75/ 20.21	1.77/ 4.32	7.28/ 8.58	3.84/ 4.40
HuatuoGPT-II/ + MEDAIDE	21.39/ 26.10	12.24/ 12.46	35.55/ 42.23	11.02/ 13.62	13.62/ 21.18	10.59/ 10.96
Baichuan4/ + MEDAIDE	15.89/ 26.10	7.06/ 13.92	27.69/ 44.55	6.92/ 13.05	14.86/ 21.35	6.59/ 11.61
LLama-3.1-8B/ + MEDAIDE	21.85/ 22.95	7.42/ 7.68	35.36/ 37.71	5.22/ 7.97	15.02/ 17.93	7.65/ 8.01
GPT-4o/ + MEDAIDE	19.27/ 26.83	9.28/ 11.54	30.72/ 40.83	8.64/ 8.71	16.22/ 18.12	8.41/ 9.89
Claude 3.7 Sonnet/ + MEDAIDE	20.14/ 27.52	9.87/ 12.05	32.25/ 42.16	9.12/ 9.83	16.84/ 18.87	8.92/ 10.43
DeepSeek-R1/ + MEDAIDE	20.92/ 28.27	10.35/ 12.63	33.78/ 43.52	9.75/ 9.86	17.51/ 19.58	9.43/ 10.94

Table 4 Comparison results on the Post-Diagnosis benchmark.

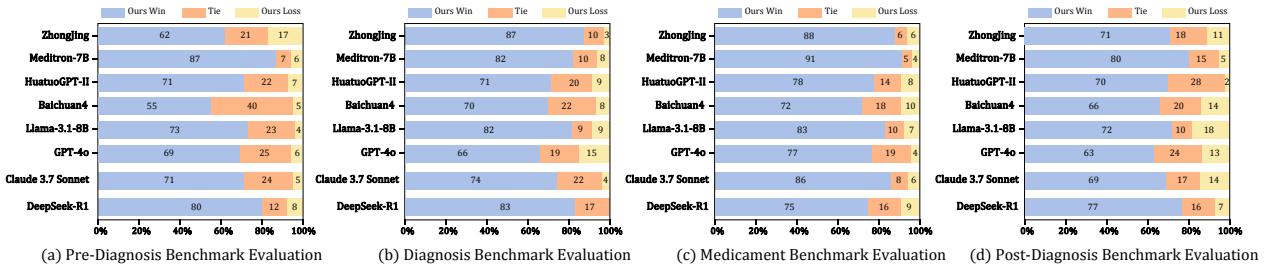


Figure 9 Response comparisons of MEDAIDE with other baselines via GPT-4o evaluation.

and specialized dosage recommendation capabilities. (iii) In the zero-shot reasoning pattern, Meditron-7B’s Meteor and BERT-Score remarkably enhance from 12.29% to 34.84% and 2.12% to 29.43% respectively, verifying that our decision-analysis module provides favourable factual evidence in drug contraindication understanding, even for smaller medical-specialized models.

Results on Post-Diagnosis Benchmark. (i) In the post-diagnosis applications from Table 4, we observe that most MEDAIDE-based models are superior to vanilla baselines by large margins. DeepSeek-R1 demonstrates exceptional improvements with a 28.8% increase in Meteor score, highlighting its superior capability in generating detailed post-treatment care instructions when augmented with our framework. (ii) Claude 3.7 Sonnet performs impressively in generating personalized rehabilitation recommendations, with a notable 7.8% improvement in BERT-Score and a 30.7% increase in Meteor score, suggesting enhanced capabilities in producing contextually appropriate recovery guidance tailored to individual patient conditions. (iii) Meanwhile, Baichuan4 shows substantial improvements in the ROUGE-L metric by 43.7%, revealing the advantages of the MEDAIDE-based version in processing dynamic medical data and producing comprehensive progress tracking reports with better continuity and coherence.

4.4 Automatic GPT-4o Evaluation

This study employs GPT-4o to evaluate how MEDAIDE enhances LLMs across reasoning and non-reasoning architectures. The framework differentially augments core capabilities: for reasoning models like DeepSeek-R1, it strengthens clinical reasoning pathways, particularly in complex diagnostic interpretation and longitudinal case analysis. For non-reasoning models, it improves structured knowledge retrieval in domains like medication management. Most significantly, MEDAIDE appears to create synergistic effects, not merely boosting performance metrics but changing how models process medical information, as evidenced by its distinct enhancement patterns across architectures. The persistent challenges in pharmacological tasks suggest these require specialized knowledge integration beyond general reasoning or retrieval improvements, pointing to a critical direction for future medical AI development.

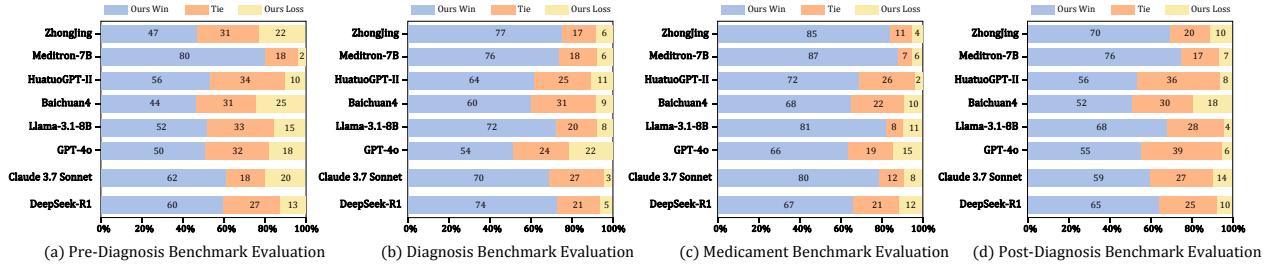


Figure 10 Response comparisons of MEDAIDE with other baselines via doctor evaluation.

Components	Pre-Diagnosis		Diagnosis		Medicament		Post-Diagnosis	
	ROUGE-L	GLEU	ROUGE-L	GLEU	ROUGE-L	GLEU	ROUGE-L	GLEU
Full MEDAIDE	17.64	7.31	23.67	12.69	52.71	37.09	18.12	9.89
w/o RIE	16.02	6.44	18.35	10.37	32.06	32.45	16.48	8.79
w/ GPT-4o Recognition	16.58	6.73	22.98	12.34	44.72	34.10	17.56	9.02
w/o Decision Analysis	17.22	7.22	23.14	12.54	48.72	35.25	18.04	9.56

Table 5 Ablation study results demonstrating component effectiveness across four medical benchmarks. w/” and w/o” are short for with and without, respectively.

4.5 Expert Doctor Evaluation

Expert evaluations play a key role in the practical applications of medical models. We invite 6 doctors (each paid \$300) to select the winners of responses generated by different models before and after the introduction of MEDAIDE by majority voting rule. The response content is holistically evaluated by considering factual accuracy, recommendation practicality, and humanistic care. (i) As Figure 10 shows, all the MEDAIDE-based models exhibit more win rates in different benchmark tests, indicating the effectiveness and applicability of the proposed framework. (ii) MEDAIDE not only improves the healthcare specialization of medical LLMs with different sizes, but also enhances the coping ability of general-purpose models when dealing with complicated medical tasks.

4.6 Ablation Study

In this section, we conduct systematic ablation studies to investigate the effects of various components, as shown in Table 5.

Necessity of RIE. Firstly, we remove the RIE to assess its impact on performance. (i) The observed significant declines in all metrics indicate that RIE is crucial for ensuring that the input information is accurately interpreted by downstream modules with respect to intents. (ii) Incorporating factual information as context enhances the model’s ability to accurately comprehend user queries, mitigating hallucinations and reducing ambiguous interpretations.

Importance of IPM. We replace the learning-based contextual encoder with the prompt-based GPT-4o to explore the impact of different intent recognition strategies. (i) Despite the improvement over the vanilla baseline, GPT-4o recognition is a suboptimal solution. We argue that the encoder trained with explicitly supervised information can align more purposefully to actual medical intents, producing more personalized and representative judgments than prompt engineering. (ii) Our default strategy has better flexibility, which can be dynamically optimized according to specific healthcare scenarios.

Effectiveness of RAC. Furthermore, we directly assemble the different outputs produced by the activated agents to serve as an alternative candidate. The results from the bottom of Table 5 show consistent performance drops of the model on different medical intent understandings, proving the effectiveness of our method. A plausible explanation is that the decision analysis module not only summarizes the different agent outputs in an organized manner, but also provides comprehensive and accurate conclusions based on medical guidelines and patient histories.

Benchmarks	MEDAIDE	w/o RIE	w/ GPT-4o R
Pre-Diagnosis	0.76	0.62	0.47
Diagnosis	0.83	0.61	0.49
Medicament	0.80	0.49	0.51
Post-Diagnosis	0.63	0.56	0.37
Intent Aggregation	0.86	0.60	0.60

Table 6 Comparison of F1 scores (%) across five benchmarks. “GPT-4o R” means the GPT-4o Recognition.

Benchmarks	MEDAIDE	MedAgents	MDAgents-Group
Pre-Diagnosis	15.93/17.64	14.33/15.26	15.74/17.21
Diagnosis	27.80/23.67	25.46/19.54	27.04/23.45
Medicament	45.13/52.17	28.69/36.07	43.37/49.58
Post-Diagnosis	26.83/18.12	25.45/16.69	25.94/18.11

Table 7 Comparison of BLEU-1/ROUGE-L scores (%) between MEDAIDE and other collaboration frameworks.

4.7 Intent Detection Analysis

To further observe the performance of intention detection, Table 6 compares the F1 scores computed between the intents recognized by MEDAIDE and ground truths under different strategies. We randomly sample 100 instances from each benchmark to compose a more challenging intention aggregation benchmark. The core observations are as follows. (i) The complete framework achieves the best results, reaching even an 86% performance on the aggregate benchmark, exhibiting omni-intention semantic understanding capabilities. (ii) Significant performance deteriorations occur at “w/o QR” and “w/ GPT-4o R”, implying the effectiveness of fine-grained decomposition of medical queries and the development of our customized intent recognition mechanism.

4.8 Collaboration Framework Comparison

Here, we compare the reproducible multi-agent framework MedAgents [34] and MDAgents [18]. MedAgents dynamically coordinates multidisciplinary medical experts across various specialties, with each expert system implemented based on GPT-4o architecture. These domain specialists collaboratively analyze cases to produce unified diagnostic reports. For optimal performance-efficiency balance, the system employs 6 active domain experts per case, while the pre-configured Group mode ensures top-tier clinical decision-making through optimized specialist combinations. (i) From Table 7, our framework outperforms MedAgents and MDAgents on all four benchmarks. For example, on the Medicament benchmark, MEDAIDE beats MedAgents by large margins with absolute gains of 16.44/16.70% on the BLEU-1/ROUGE-L scores. This strength comes from our specialized drug retrieval, which provides a refined context of drug information for the model. (ii) Also, the benefits of Pre-Diagnosis and Post-Diagnosis tasks reflect that the proposed MEDAIDE captures the key characteristics of diagnostic and nursing demands more efficiently.

4.9 Response Visualization Analysis

To intuitively compare the response quality of the different models, in Figure 11 we show the results generated by the MEDAIDE and SOTA methods under the same user query. HuatuoGPT-II provides extensive medical knowledge, but lacks targeted recommendations. Although the MedAgents framework gives detailed treatment solutions through multidisciplinary perspectives, it lacks coherent diagnostic reasoning. In contrast, our framework interprets patient outcomes in detail and provides personalized diagnoses and subsequent treatments. MEDAIDE shows favorable usefulness in the management of electrolyte imbalance and cardiovascular risk.

4.10 Effect of Task Decomposition Granularity

To systematically analyze how task decomposition granularity affects performance, we design a controlled experiment where the original 17 healthcare intents are dynamically allocated to agents in different stage configurations (ranging



User Query: I have experienced chest tightness and abdominal discomfort for 3 days. The auxiliary examination results are as follows. Complete Blood Count (CBC) shows an elevated White Blood Cell Count (WBC) of $11.83 \times 10^9/L$, and a Neutrophil Ratio of 92.0%. Electrolyte tests reveal that Sodium (Na) is 134.8 mmol/L and Chloride (Cl) is 95.2 mmol/L, both below normal levels. Lipid profile indicates that High-Density Lipoprotein Cholesterol (HDL-C) is 0.49 mmol/L and Low-Density Lipoprotein Cholesterol (LDL-C) is 1.53 mmol/L, both lower than usual. Chlamydia Pneumoniae Antibody (CPN-IgG) tested positive. Chest imaging results show an infection in the right lower lung, an exudative lesion in the right lung suggesting interstitial inflammation, mild thickening and adhesion of the right pleura, as well as aortic and coronary artery sclerosis. I would like to know what might be causing these symptoms?

MEDAIDE (Ours):

Query Rewriting

Please interpret the following examination results: elevated White Blood Cell Count (WBC) at $11.83 \times 10^9/L$ and Neutrophil Ratio at 92.0% in the CBC; decreased Sodium (Na) at 134.8 mmol/L and Chloride (Cl) at 95.2 mmol/L in the electrolyte panel; reduced HDL-C at 0.49 mmol/L and LDL-C at 1.53 mmol/L in the lipid profile; positive Chlamydia Pneumoniae Antibody (CPN-IgG). Chest imaging shows infection in the right lower lung, with CT indicating an exudative lesion, interstitial inflammation, mild right pleural thickening, and aortic and coronary artery sclerosis. What potential health risks do these findings suggest, and is further examination or treatment needed? Based on these results and the patient's symptoms of chest tightness and abdominal discomfort for 3 days, please analyze the possible causes.

Integrated Decision Analysis

Activated Diagnosis Agent

Test Interpretation Intent: The CBC results show elevated WBC at $11.83 \times 10^9/L$ and an increased neutrophil ratio of 92.0%, suggesting inflammation or bacterial infection, potentially related to a lung infection. Electrolyte levels indicate low Na at 134.8 mmol/L and low Cl at 95.2 mmol/L, pointing to mild hyponatremia or possible fluid imbalance, which could be linked to an infection or kidney function issues; The lipid profile reveals low HDL-C at 0.49 mmol/L and low LDL-C at 1.53 mmol/L, both of which could increase cardiovascular risk..... There is also mild thickening and adhesion of the right pleura, along with aortic and coronary artery sclerosis, which may suggest chronic inflammation or increased cardiovascular risk.

Etiology Detection Intent: Possible causes include acute infection, electrolyte imbalance, lipid abnormalities, and pleural changes. Elevated WBC and neutrophil ratio, along with chest imaging showing infection and exudative lesions in the right lung, suggest acute pneumonia, likely caused by Chlamydia pneumoniae or another pathogen. Low Na and Cl levels indicate electrolyte imbalance, potentially due to dehydration from infection, which could worsen symptoms. Reduced HDL-C and LDL-C, along with aortic and coronary artery sclerosis, suggest cardiovascular risk. Thickening and adhesion of the right pleura on CT indicate pleuritis or chronic inflammation, contributing to chest tightness.

Activated Post-Diagnosis Agent

I recommend that you first ensure adequate rest and avoid strenuous activities. In terms of diet, I suggest consuming foods rich in sodium and potassium, such as bananas, spinach, and dairy products, to help correct electrolyte imbalances. Additionally, drink warm water or electrolyte drinks daily to maintain electrolyte balance. If you feel anxious or uneasy, I encourage you to talk to family or friends and share your feelings.

HuatuoGPT-II:

Based on your symptoms and test results, several factors could be contributing to your discomfort, and it's essential to consult a healthcare professional for a thorough evaluation and tailored treatment plan. Your symptoms of chest tightness and abdominal discomfort, along with an elevated white blood cell count and neutrophil ratio, suggest a possible immune response or infection, such as pneumonia. Low sodium and chloride levels might indicate hydration issues or conditions like hyponatremia. It's important to evaluate your hydration status and overall health. Your low HDL-C and LDL-C levels require further assessment to manage any risk of cardiovascular diseases. Also, your positive Chlamydia pneumoniae antibody indicates past exposure, which needs clarification to determine its relevance to your current condition. Chest imaging shows an infection in the right lower lung with signs of interstitial inflammation and pleural changes, alongside evidence of aortic and coronary artery sclerosis. These findings should be monitored by specialists to ensure they are addressed properly. In conclusion, your varied symptoms and results point to multiple potential causes that need comprehensive evaluation by medical professionals to diagnose accurately and manage effectively.

MedAgent:

Cardiology: Emphasizes the need for cardiovascular risk management due to findings of aortic and coronary artery sclerosis and recommends initiating statin therapy or similar treatments to manage cholesterol levels and prevent plaque progression. **Gastroenterology:** Addresses concerns related to abdominal discomfort and electrolyte imbalances. Recommends management strategies for hydration and electrolyte supplementation, alongside monitoring for gastrointestinal complications. **Radiology:** Points out the critical nature of the lung findings and the importance of monitoring these through follow-up imaging to assess the resolution of infection and inflammation. **Surgery:** While currently recommending medical management, advises on the necessity of being prepared for potential surgical interventions if complications like abscess formation or severe pleural complications develop. **The agreed treatment plan** includes initiating an antibiotic regimen targeting Chlamydia pneumoniae, with adjustments based on clinical response. We'll also begin cardiovascular protective treatments for aortic and coronary sclerosis and manage electrolyte imbalances. Follow-up includes chest imaging within 4-6 weeks, regular cardiovascular assessments, and continuous gastroenterology monitoring. The surgery team will remain alert for potential surgical intervention if symptoms do not improve. Long-term, we recommend lifestyle and dietary adjustments, with regular multi-disciplinary reviews to adapt the treatment plan as needed.

Figure 11 Response visualization analysis on different models for the same user query.

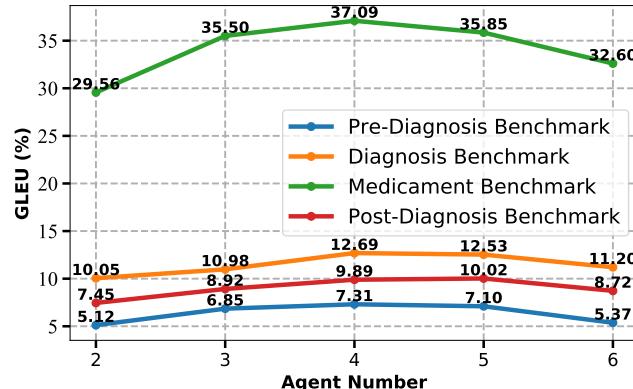


Figure 12 The performance of different agent groups, formed by varying numbers of agents, across the four medical benchmarks.

from 2 to 6 stages). Each agent is assigned to handle a specific stage, with intent coverage adjusted according to stage divisions (e.g., 2-stage: Pre-Diagnosis and Post-Diagnosis; 4-stage: Pre-Diagnosis, Diagnosis, Medication, and Post-Diagnosis). As shown in Figure 12, MEDAIDE achieves optimal performance with 4 decompositions, where intent-specific responsibilities are balanced without overlap. Progressive improvements were observed when increasing from 2 to 4 agents, confirming that finer task decomposition enhances multi-intent handling. However, further subdivision (5–6 stages) yields diminishing returns, as excessive granularity introduces coordination overhead.

5 Conclusion

This paper introduces MEDAIDE, a multi-agent framework for complex medical scenarios. Leveraging Regularization-guided information extraction module, Dynamic intent prototype matching module, and Rotation agent collaboration mechanism, MEDAIDE enhances the model's comprehension of medical intents and demonstrates effectiveness across sophisticated medical benchmarks.

Limitations. The MEDAIDE framework, while making significant strides in integrating large-scale medical agents with real clinical environments, has its limitations. Currently, the framework incorporates 26,684 drug samples and 506

genuine clinical case records. Despite this scope, it lacks comprehensive coverage, particularly in research on rare diseases and specialized medications. Furthermore, as the initial versions primarily emphasized linguistic processing, future research aims to extend into multimodal capabilities, particularly the integration of medical imaging data, to enhance the framework’s ability to process visual information in clinical diagnostics. The reliance on OpenAI’s API may pose potential operational challenges, suggesting that future studies should explore using more efficient open-source models as viable alternatives.

Ethics Consideration. We are acutely aware of the necessity for privacy and data protection. All data utilized has undergone thorough de-identification, with all sensitive information removed, and verified by a partnering medical institution. We invite doctors to perform only evaluations of model responses without involving any form of human subject research. All participants are compensated \$300 for their work, which strictly adheres to the minimum hourly rate for the region in which the work is performed. We strictly follow the license agreements of publicly available databases when utilizing healthcare-related data. For the constructed data, we have undergone an internal ethical review by the ethics review board of our partnering medical institutions and are licensed and approved.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (2021ZD0113503) and in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103).

References

- [1] Baichuan, 2023. <https://github.com/baichuan-inc/Baichuan-13B>.
- [2] Baichuan AI. Introducing baichuan4, 2024. <https://platform.baichuan-ai.com/homePage>.
- [3] Farman Ali, Shaker El-Sappagh, S M Riazul Islam, Dohyeun Kwak, Ahmad Ali, Muhammad Imran, and Kyung-Sup Kwak. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63:208–222, 2020.
- [4] Anthropic. Claude 3.7 sonnet. <https://www.anthropic.com/clause>, 2025. 2025.3.
- [5] Alan R Aronson and François-Michel Lang. An overview of metomap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
- [6] Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. Huatuogpt-ii, one-stage training for medical adaption of llms. *ArXiv*, abs/2311.09774, 2023. URL <https://api.semanticscholar.org/CorpusID:265221365>.
- [7] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.
- [8] Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17853–17861, 2024.
- [9] Edgar F Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.
- [10] Wei Du and Shifei Ding. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artificial Intelligence Review*, 54(5):3215–3238, 2021.
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [12] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024.
- [13] Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343, 2018.
- [14] Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Junqing He, Yuanhe Tian, Ping Yang, et al. Ziya2: Data-centric learning is all llms need. *arXiv preprint arXiv:2311.03301*, 2023.
- [15] Qiang Gao, Jie Zhang, Fan Zhou, et al. Multi-modal lifelog data fusion for improved human activity recognition: A hybrid approach. *Information Fusion*, 110:102468, 2024.
- [16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [18] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms in medical decision making. 2024.
- [19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019. URL <https://api.semanticscholar.org/CorpusID:59291975>.
- [20] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462*, 2025.
- [21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

- [22] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. Camel: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.
- [23] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, et al. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 112:102545, 2024.
- [24] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013/>.
- [25] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*, 20(2):2843–2855, 2023.
- [26] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7):1–38, 2024.
- [27] Ghulam Muhammad and M Shamim Hossain. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021.
- [28] Toan Nguyen-Mau, Anh-Cuong Le, Duc-Hong Pham, and Van-Nam Huynh. An information fusion based approach to context-based fine-tuning of gpt models. *Information Fusion*, 104:102202, 2024.
- [29] OpenAI. Introducing chatgpt, 2022. <https://openai.com/blog/chatgpt>.
- [30] OpenAI. Introducing o1, 2024. <https://openai.com/index/introducing-openai-o1-preview/>.
- [31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
- [32] Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 155–161. IEEE, 2024.
- [33] Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Vel’asquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102:102040, 2024.
- [34] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.
- [35] Jiacong Wang, Zijiang Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025.
- [36] Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. Medco: Medical education copilots based on a multi-agent framework. *arXiv preprint arXiv:2408.12496*, 2024.
- [37] Jinjie Wei, Jiyao Liu, Lihao Liu, Ming Hu, Junzhi Ning, Mingcheng Li, Weijie Yin, Junjun He, Xiao Liang, Chao Feng, et al. Learning, reasoning, refinement: A framework for kahneman’s dual-system intelligence in gui agents. *arXiv preprint arXiv:2506.17913*, 2025.
- [38] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [39] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, page 102888, 2024.
- [40] Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems*, 37:138632–138662, 2024.
- [41] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376, 2024.

- [42] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [43] Bo Zhang, Hui Ma, Jian Ding, Jian Wang, Bo Xu, and Hongfei Lin. Distilling implicit multimodal knowledge into large language models for zero-resource dialogue generation. *Information Fusion*, page 102985, 2025.
- [44] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [45] Xiaodong Zhang and Houfeng Wang. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999, 2016.