

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189421/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Yang, Dingkan, Wei, Jinjie, Hu, Ming, Liu, Jiyao, Liu, Lihao, Chen, Zhaoyu, Li, Mingcheng, He, Junjun, Zhou, Wei, Liu, Yang and Zhang, Lihua 2026. Toward empathetic care: an LLM-based multi-intention recognition framework for mental health and complex medical queries. IEEE Transactions on Affective Computing 10.1109/taffc.2026.3728254

Publishers page: <https://doi.org/10.1109/taffc.2026.3728254>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



# Toward Empathetic Care: An LLM-based Multi-Intention Recognition Framework for Mental Health and Complex Medical Queries

Dingkang Yang<sup>1</sup>, Jinjie Wei<sup>2</sup>, Ming Hu<sup>3</sup>, Jiyao Liu<sup>4</sup>, Lihao Liu<sup>5</sup>, Zhaoyu Chen<sup>6</sup>, Mingcheng Li<sup>7</sup>, Junjun He<sup>8</sup>, Wei Zhou<sup>9</sup>, Yang Liu<sup>10</sup>, Lihua Zhang<sup>11</sup>, *Member, IEEE*

**Abstract**—Mental health has become a critical global concern, and many individuals now turn to conversational agents for support when facing emotional distress and complex health problems. In these high-stakes and emotionally sensitive settings, patients’ self-descriptions are often long, fragmented, and interweave medical facts with psychological needs, lifestyle concerns, and urgent safety questions. Large Language Model (LLM)–based assistants must therefore not only answer clinical questions, but also accurately uncover the multiple underlying intentions that drive a patient’s request, as a prerequisite for empathetic and trustworthy interaction. In this work, we propose MIRA, a multi-intention recognition and planning framework designed to understand complex patient intentions in mental health–related consultations. MIRA first introduces a two-stage intention recognition module that decomposes verbose queries into clinically aligned sub-queries using semantic matching with medical intent prototypes, including intents such as medical psychological support and lifestyle guidance. Next, a Dynamic Intent Orchestration (DIO) agent leverages clinical dependency modeling to plan execution sequences that respect medical best practices and safety considerations. Based on this plan, a multi-agent collaboration system with a Hierarchical Progressive Decision-making (HPD) agent synthesizes evidence across sequential diagnostic stages to generate context-aware advice that is suitable for emotionally sensitive scenarios. We evaluate MIRA on several real-world medical dialogue benchmarks. Both automated metrics and expert physician evaluations show that MIRA consistently outperforms existing different baselines in intent comprehension, clinical planning, and overall response quality.

**Index Terms**—Mental health, Complex medical queries, Large

This work was supported in part by the 2026 AI Breakthrough Initiative Research Project 2026JLJG0001GX, and in part by the Shanghai Municipal Science and Technology Major Project 2023SHZDZX02.

Dingkang Yang is with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University & Fysics AI, Shanghai, 200433, China, and also with the Multimedia Laboratory (MMLab), the Chinese University of Hong Kong, Hong Kong SAR, China. (E-mail: dkyang20@fudan.edu.cn).

Jinjie Wei, Zhaoyu Chen, Mingcheng Li and Lihua Zhang are with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai 200433, China. (E-mails: {zhaoyuchen20, lihuazhang}@fudan.edu.cn; {jjwei23, mingchengli21}@m.fudan.edu.cn)

Ming Hu, Lihao Liu, Junjun He are with the Shanghai Artificial Intelligence Laboratory, Shanghai 200032, China. (E-mails: ming.hu@monash.edu; ll610@cam.ac.uk; hejunjun@pjlab.org.cn)

Jiyao Liu is with the Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China. E-mail: liujy22@fudan.edu.cn.

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, United Kingdom. E-mail: zhouw26@cardiff.ac.uk.

Yang Liu is with the College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China. E-mail: yang\_liu@ieee.org.

Corresponding authors: Jinjie Wei, Yang Liu, and Lihua Zhang.

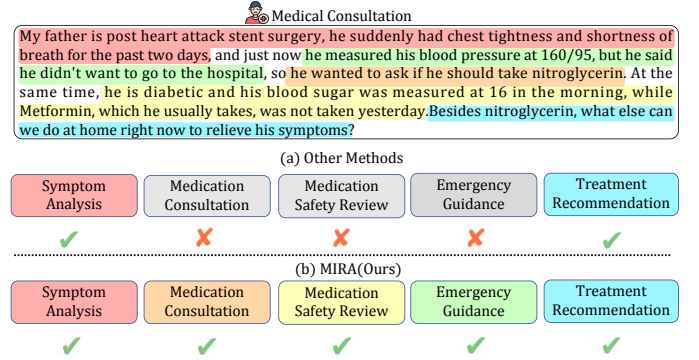


Fig. 1. Comparative analysis between MIRA and other Medical Multi-Agent Systems (such as MDAgents [15]) on real-world patient scenarios including patient intent extraction. Previous methods identify limited medical intentions, while MIRA extracts multiple intents for more comprehensive diagnosis.

language models, Agent collaboration, Intent comprehension

## I. INTRODUCTION

MENTAL health disorders impose a substantial burden on individuals and healthcare systems worldwide [1], [2]. Despite the growing demand, professional mental health services remain inaccessible for many, hampered by systemic workforce shortages, prohibitive waitlists, and the persistent barrier of social stigma [3]. As a result, many people describe their symptoms, emotions, and daily struggles in online platforms and conversational systems, hoping to receive timely guidance or reassurance [4], [5], [6], [7]. In these emotionally sensitive scenarios, patient utterances are typically verbose, mixing physical complaints, medication questions, psychological distress, and urgent safety concerns within a single message. Recent advances in Large Language Models (LLMs) have enabled more human-like interactions in healthcare and mental health applications [8], [9], [10]. However, most existing LLM-based assistants still treat each user query as a single, unified intent, overlooking the multiple intertwined needs that underlie real-world consultations [11], [12]. This limitation is particularly harmful in mental health interactions, where users may simultaneously seek symptom analysis, treatment information, lifestyle advice, and psychological support or crisis-related guidance [13], [14]. Accurately extracting these co-occurring intentions is a key prerequisite for building emotionally aware, trustworthy, and clinically safe AI agents.

Recently, Large Language Model (LLM)-based medical assistants [16], [17], [18], [19], [20] have emerged as a transformative solution, showing immense promise in improving diagnostic efficiency and automating widely accessible healthcare services. To empower these models with professional capabilities, previous research efforts have primarily focused on injecting healthcare-specific knowledge into general-purpose LLMs. This is typically achieved through strategies such as constructing specialized external knowledge bases for retrieval [21] or fine-tuning models on large-scale medical corpora [22]. While these methods effectively enhance the models' factual understanding of medical concepts, they often exhibit significant limitations when deployed in dynamic, real-world clinical environments where adaptability is key. To address the rigidity of single-model approaches, LLM-based multi-agent systems have attracted attention as a superior alternative. By simulating a collaborative environment, these systems utilize multiple agents to follow complex instructions and facilitate decision-making in realistic scenarios. In such frameworks [23], [24], distinct agents can assume specialized roles and collaborate to address diverse and shifting patient needs, theoretically leading to more accurate and comprehensive medical outcomes.

Despite this progress, current implementations face a critical bottleneck. As illustrated in Figure 1, existing multi-agent systems are predominantly designed for structured, single-turn tasks like medical question-answering (QA) [25]. Consequently, they struggle to navigate the intricacies of authentic, multi-turn medical dialogues. In actual clinical practice, patients rarely provide concise or structured information; their self-descriptions are frequently verbose, disorganized, and laden with hidden intentions or emotional nuances [26]. Therefore, the ability to accurately distill core medical needs from such unstructured input is not just an enhancement—it is a fundamental prerequisite for improving the reliability of medical decision-making.

To address the above challenges, we propose MIRA, an LLM-based multi-agent framework designed to better understand patients' complex intentions in healthcare and mental health-related consultations. MIRA is positioned as a core building block for empathetic AI assistants in emotionally sensitive settings, where users often intertwine clinical questions with psychological support needs and lifestyle concerns. First, a two-stage intention recognition module decomposes verbose, emotionally laden patient descriptions into clinically grounded sub-queries, aligning them with a taxonomy of medical and psychological-support intents and thereby enabling the model to identify co-occurring medical and mental-health-relevant goals. Next, a Dynamic Intent Orchestration (DIO) agent leverages clinical dependency modeling and safety-aware prioritization to coordinate the execution flow among multiple specialized agents, ensuring that high-risk or psychologically salient needs are addressed in an appropriate order. Finally, a multi-agent collaboration system employs intent-specific agents, coordinated by a Hierarchical Progressive Decision-making (HPD) agent that accumulates evidence across sequential stages to produce coherent, trustworthy advice suitable for complex and emotionally sensitive health interactions.

Beyond improving medical reasoning and task execution, MIRA also contributes to affective computing by enabling affect-aware intention understanding in emotionally sensitive healthcare interactions. Instead of treating emotions as isolated signals, MIRA models psychological-support needs as an integral component of patient intentions and coordinates them with medical concerns through safety-aware planning. Such intention-level affective understanding provides a foundation for building more empathetic and trustworthy AI assistants. The main contributions of this work are as follows:

- We introduce MIRA, an LLM-driven multi-agent framework that integrates multi-intention recognition and clinical planning to address complex patient queries in medical and mental health contexts. MIRA provides a foundation for building empathetic and trustworthy AI assistants in emotionally sensitive healthcare scenarios.
- We develop a clinically grounded intention understanding and planning mechanism, including a multi-intent taxonomy, a two-stage intention recognition module, and a dynamic orchestration strategy coordinated by a hierarchical progressive decision-making agent. These components enable safety-aware intent prioritization and evidence-grounded response generation for high-stakes medical and mental health interactions.
- We construct the Mental Multi-Intention (MMI) dataset from real-world medical consultation records and conduct comprehensive evaluations on multiple medical dialogue benchmarks. The proposed dataset and evaluation framework provide more realistic resources for studying complex multi-intent understanding and empathetic LLM-based healthcare assistants.

## II. RELATED WORK

### A. LLM-driven Multi-agent Healthcare Systems

LLMs such as ChatGPT and GPT-4 [27], [28] have demonstrated strong generalization across multidisciplinary tasks, including healthcare and early explorations in mental health support. However, while general-purpose LLMs [29] possess broad medical knowledge, they often lack the domain-specific alignment and safety required for deployment in high-stakes, emotionally sensitive scenarios. To address this gap, recent medical LLMs [16], [17], [20], [30], [16] (e.g., HuatuoGPT [31], [32], ZhongJing [18], [19], PediatricsGPT [17], Meditron [20], and MedicalGPT [30]) enhance clinical robustness through curated doctor-patient conversation corpora, expert feedback, and specialized training pipelines. These models primarily focus on improving factual accuracy and clinical reasoning, laying the foundation for safe AI assistance in healthcare.

Despite these advances, single-agent LLM systems still face challenges such as hallucinations, inconsistent reasoning, and limited contextual comprehension in complex clinical encounters, especially when patients mix physical complaints with psychological distress or lifestyle concerns. To alleviate these limitations, multi-agent architectures have emerged as a promising paradigm. Methods such as MDAgent [15] and MEDAgents [25] employ dynamic multi-expert discussion mechanisms, where specialized agents collaboratively handle

different aspects of the medical task. These systems improve decision robustness and interpretability, but they typically assume a single dominant task or intent per interaction and pay limited attention to the fine-grained, multi-intent nature of real-world medical and mental health-related consultations. Our work builds on this line of research by equipping an LLM-based multi-agent framework with explicit multi-intention modeling tailored to complex, emotionally sensitive patient queries.

### B. Multi-Intent Comprehension

In both general healthcare and mental health support, a single patient inquiry often implicitly contains multiple intertwined intents, such as symptom analysis, diagnosis, treatment recommendation, lifestyle counseling, and psychological support [33], [34]. Recent healthcare dialogue systems have also studied complex user intent detection. For example, CONCERNIE [35] investigates complex intent detection for smart healthcare cognitive bots and shows that user queries in medical dialogue often contain multiple correlated intents. Accurately recognizing these co-occurring intents and structuring them into clinically meaningful stages is crucial for safe and empathetic assistance, yet it remains a challenging problem. Traditional intent recognition methods evolved from rule-based systems to neural models based on sequence labeling and hierarchical classification, which work well for short, well-structured queries but struggle with verbose, narrative-style patient descriptions [36], [37], [38].

With the emergence of LLMs, research has increasingly leveraged prompt engineering and fine-tuning for intent classification in domain-specific settings. Medical LLMs and clinical agents have begun to incorporate intent-aware prompting and task decomposition, but many remain constrained by predefined workflows and rigid rule sets, limiting their adaptability when patients present multiple, partially overlapping needs [39], [40]. In particular, current LLM-based methods often treat psychological support or counseling as implicit sub-tasks rather than explicitly modeled intents, making it difficult to prioritize or adapt responses in emotionally sensitive interactions [41], [42], [43], [34]. Recent work has also explored intent-aware assistants for distress and mental-health support. Nazar et al. [44] introduce a RAG-based situation- and intent-aware assistant that infers emotionally charged user intents, adapts responses to the user state, and grounds support in curated knowledge sources. In contrast, our work explicitly models a clinically grounded taxonomy of medical and psychological-support intents and introduces a multi-intention recognition mechanism that decomposes verbose patient queries into aligned sub-intents, thereby providing a structured basis for downstream empathetic and context-aware response generation.

### C. Intent Orchestration and Task Planning

Beyond recognizing multiple intents, effectively orchestrating and prioritizing them is essential for clinical decision support and for building trustworthy AI assistants in mental health-related contexts. Traditional clinical workflow management systems [45] rely on static, rule-based scheduling, which lacks the flexibility to adapt to dynamic patient conditions or

evolving emotional states. AI-driven task planning has explored dependency modeling and priority-based scheduling [46], but many of these approaches are designed for generic planning problems and do not fully incorporate medical or safety-specific constraints. In healthcare, inappropriate task sequencing can lead to suboptimal or even harmful outcomes [47], such as prescribing treatment before sufficient diagnostic evaluation or failing to address urgent safety concerns early in the interaction.

Traditional clinical pathway and decision support systems [48], [49] often rely on rigid, linear protocols. Consequently, they struggle to handle the nuance of complex consultations where medical and psychological needs intersect. This rigidity is particularly risky in mental health settings, where the precise timing and framing of interventions are vital for trust and safety. To address this, we introduce a dynamic intent orchestration mechanism. This approach adaptively sequences tasks based on safety priorities and clinical dependencies, ensuring that the system handles composite patient needs with the flexibility of real-world clinical reasoning.

### D. Logical Reasoning in Mental Health

Logical reasoning is a core requirement for complex medical decision-making and for providing coherent, trustworthy guidance in emotionally charged consultations [50], [51], [52], [53]. Traditional methods such as Chain-of-Thought (CoT) [54] and its extensions (CoT-SC [55], ToT [56], GoT [57]) encourage LLMs to break down problems into intermediate reasoning steps, improving performance on many clinical QA and diagnostic tasks. However, these approaches typically operate on a single explicit task formulation and are less suited to settings where multiple, latent intents must be uncovered and reasoned over jointly.

Recent medical multi-agent systems, such as MDAgent [15] and MedAgent [25], leverage collaborative agents to enhance reasoning flexibility by distributing subtasks (e.g., diagnosis, treatment planning, evidence retrieval) across specialized roles. MedAIDE [39] further combines multi-modal data fusion with medical intent modeling to produce more context-sensitive outcomes. Nevertheless, these frameworks generally focus on improving clinical reasoning accuracy rather than explicitly modeling the multi-intent structure of patient queries or tailoring reasoning workflows to emotionally sensitive, health-related interactions. Our work complements these efforts by introducing a framework that not only enforces structured reasoning within and across agents but also explicitly identifies, decomposes, and prioritizes concurrent medical and psychological intents. By synthesizing multi-intent recognition with dynamic orchestration and hierarchical decision-making, MIRA enhances the efficiency and contextual relevance of AI-driven care.

## III. METHODOLOGY

### A. Overall Workflow of MIRA

As illustrated in Figure 2, the MIRA framework operates through a cohesive three-stage pipeline designed to bridge raw patient input with actionable clinical decisions. First, in the Intention Recognition stage, we deploy a specialized

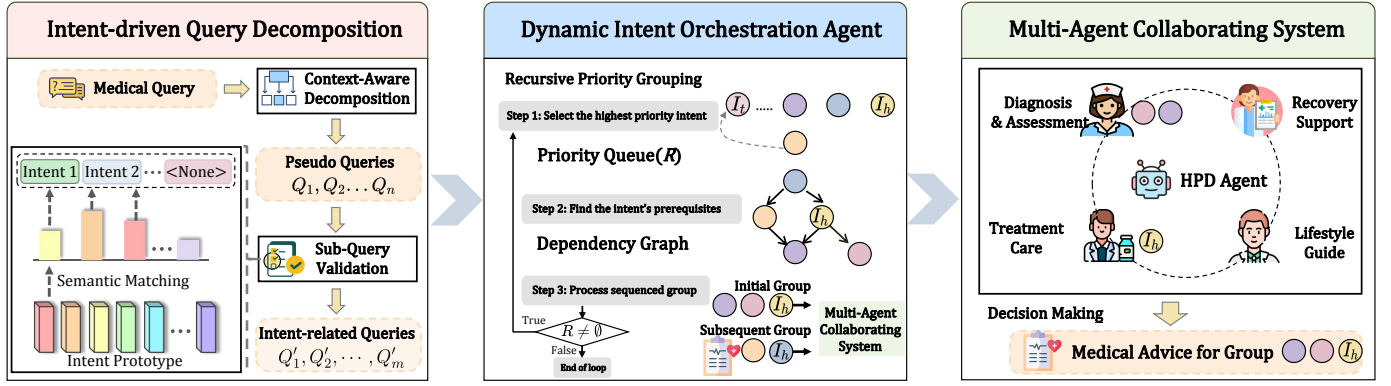


Fig. 2. Overview of MIRA framework. The framework comprises three core modules: (1) Intent-driven Query Decomposition (left): The system breaks down complex queries into sub-queries by performing intent recognition, context-aware decomposition, and sub-query validation. (2) Dynamic Intent Orchestration Agent (middle): This agent builds a dependency graph and calculates priority scores to reorder sub-query execution based on urgency, dependency, and contextual priority. (3) Multi-Agent Collaboration System (right): This module dynamically activates one or more of the four specialized agents based on recognized intents. A central HPD agent with CoT-based reasoning and Knowledge-Guided Few-shot Sampling (KGFS) integration consolidates their insights to generate context-aware medical advice. For the multi-agent collaboration system, intent-specific agents generate structured reports under HPD coordination, followed by evidence accumulation, safety checking, and final synthesis.

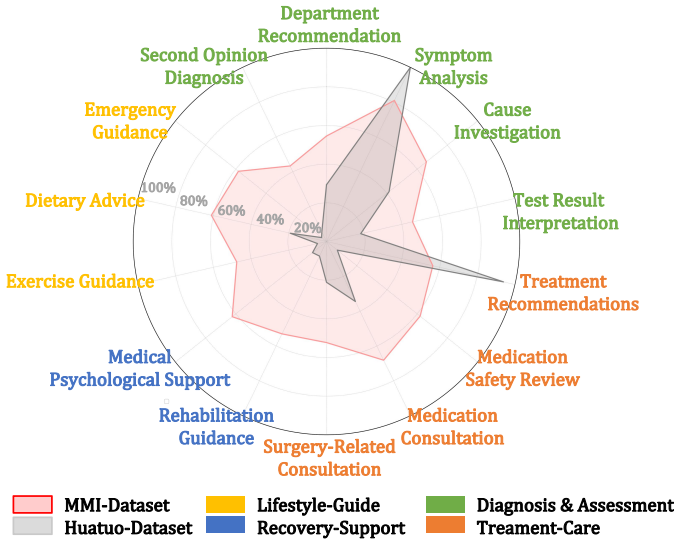


Fig. 3. Radar chart comparing the distribution of medical intent categories between the MMI and Huatuo Dataset. The values are normalized by dividing each data point by the maximum value across both datasets to ensure comparable scales from 0 to 1. The chart encompasses 14 medical intents with values plotted on a polar coordinate.

recognition module that ingests complex, unstructured medical queries from patients. This module is responsible for parsing the text to accurately distill the underlying, multifaceted intentions. Subsequently, in the Planning stage, we introduce the Dynamic Intent Orchestration (DIO) agent. Serving as the system's central conductor, the DIO agent dynamically sequences and coordinates the 14 fundamental medical intents (detailed in Figure 3). It ensures that the workflow is logically arranged based on the recognized needs. Finally, the process concludes with the Execution stage, which features a robust multi-agent collaboration system. Here, intent-specific agents address distinct steps of the clinical process, while the Hierarchical Progressive Decision (HPD) agent systematically executes these intentions to synthesize the final medical decision.

### B. Medical Intent Taxonomy

As shown in Figure 3, we organize the 14 medical care intents into four clinically motivated stages: (a) Diagnosis & Assessment, (b) Treatment-Care, (c) Recovery-Support, and (d) Lifestyle-Guide. This structure mirrors the natural workflow of real-world consultations. It begins with initial symptom exploration and risk assessment, then advances to therapeutic decision-making. Finally, it extends to long-term follow-up, covering both psychological support and lifestyle adjustments. To ensure validity, we derive this taxonomy from 500 clinical records. Three board-certified physicians then refine it to cover diverse medical and mental health needs.

The taxonomy is constructed through a clinical-workflow-guided and data-informed process. It is not purely derived from unsupervised clustering or directly mapped from ICD-10, because ICD-10 focuses on disease coding, whereas our taxonomy describes patient consultation intents. Specifically, we review 500 de-identified clinical consultation records to identify recurring patient needs, organize the candidate intents according to the general care workflow, and ask three board-certified physicians to refine the categories. The final 14-category taxonomy therefore reflects both empirical patterns in real consultations and clinical reasoning about patient care.

### C. Intent-driven Query Decomposition

In this module, we propose a robust two-stage approach to decompose complex, verbose patient narratives into clinically relevant sub-queries. This process ensures that each resulting sub-query addresses a distinct medical intent, facilitating more precise downstream processing.

**Stage 1: Pseudo Query Generation.** Given an input query  $Q$ , we first employ GPT-4.5 [58] to perform initial query decomposition. This stage generates a set of potential sub-queries  $\mathcal{P} = \{p_1, p_2, \dots, p_m\}$ , termed *Pseudo Queries*, by leveraging the LLM's capability to parse semantic structures and clinical contexts.

TABLE I  
INTENT DEPENDENCY TABLE FOR THE RPG ALGORITHM. ✓ INDICATES THAT THE INTENT IN THE ROW REQUIRES THE INTENT IN THE COLUMN AS A PREREQUISITE.

| Intent                              | SA | CI | TRI | D | TR | MSR | SRC | MC | DA | EG | MPS | RG | ExG | DR |
|-------------------------------------|----|----|-----|---|----|-----|-----|----|----|----|-----|----|-----|----|
| Symptom Analysis (SA)               | -  |    |     |   |    |     |     |    |    |    |     |    |     |    |
| Cause Investigation (CI)            | ✓  | -  |     |   |    |     |     |    |    |    |     |    |     |    |
| Test Result Interpretation (TRI)    | ✓  | ✓  | -   |   |    |     |     |    |    |    |     |    |     |    |
| Diagnosis (D)                       | ✓  | ✓  | ✓   | - |    |     |     |    |    |    |     |    |     |    |
| Treatment Recommendations (TR)      | ✓  | ✓  | ✓   | ✓ | -  |     |     |    |    |    |     |    |     |    |
| Medication Safety Review (MSR)      |    |    | ✓   | ✓ | ✓  | -   |     |    |    |    |     |    |     |    |
| Surgery-Related Consultation (SRC)  | ✓  | ✓  | ✓   | ✓ | ✓  |     | -   |    |    |    |     |    |     |    |
| Medication Consultation (MC)        | ✓  | ✓  | ✓   | ✓ | ✓  | ✓   |     | -  |    |    |     |    |     |    |
| Dietary Advice (DA)                 | ✓  | ✓  | ✓   | ✓ | ✓  |     |     |    | -  |    |     |    |     |    |
| Emergency Guidance (EG)             | ✓  |    |     |   |    |     |     |    |    | -  |     |    |     |    |
| Medical Psychological Support (MPS) | ✓  |    |     | ✓ |    |     |     |    |    |    | -   |    |     |    |
| Rehabilitation Guidance (RG)        | ✓  | ✓  | ✓   | ✓ | ✓  |     | ✓   |    |    |    |     | -  |     |    |
| Exercise Guidance (ExG)             | ✓  | ✓  | ✓   | ✓ | ✓  |     |     |    | ✓  |    |     |    | -   |    |
| Department Recommendation (DR)      | ✓  |    |     | ✓ |    |     |     |    |    |    |     |    |     | -  |

We govern this decomposition using a specific prompt design, denoted as  $P_{\text{decompose}}$ . This prompt explicitly instructs the model on how to structure the outputs. It requires each sub-query to be self-contained and linguistically concise. Furthermore, it ensures that the essential medical context of the original input is preserved:

$$P = \text{LLM}(Q, P_{\text{decompose}}). \quad (1)$$

The detailed prompt template used in this stage is presented in Table II. By strictly adhering to these constraints, we minimize ambiguity and ensure that multi-faceted patient descriptions are broken down into atomic, actionable medical questions.

**Stage 2: Semantic Matching.** To accurately categorize the intentions of the generated pseudo queries, we introduce a semantic matching mechanism. We employ a specialized BioBERT model [39], which has been pre-trained on large-scale biomedical corpora (including PubMed abstracts and PMC full-text articles) and fine-tuned for medical intent understanding.

For each pseudo query  $p_j \in \mathcal{P}$ , we compute its semantic similarity against our predefined set of intent prototypes  $\mathcal{I} = \{i_1, i_2, \dots, i_{14}, i_0\}$ , where  $i_0$  represents a None category for irrelevant queries. The matching score is calculated using cosine similarity based on the BioBERT embeddings:

$$s_{ji} = \text{sim}(\text{BioBERT}(p_j), \text{BioBERT}(i_i)). \quad (2)$$

Each pseudo query is deterministically assigned to the intent category that yields the maximum semantic similarity score, ensuring alignment with the most relevant clinical domain:

$$\text{intent}(p_j) = \arg \max_{k \in \{1, \dots, 15\}} s_{jk}. \quad (3)$$

The rigorous matching process effectively filters noise and maps unstructured patient language to the clinical taxonomy.

For each of the 14 medical intents, we construct 5 natural-language prototypes based on the intent definition, representative expressions from training records, and physician-reviewed boundary cases. This results in 70 clinical intent prototypes. Each prototype describes a typical patient need under the corresponding intent, such as asking about medication safety, symptom interpretation, emergency risk, or psychological support. We further define a None category with 8 prototypes,

TABLE II  
PROMPT TEMPLATE FOR PSEUDO QUERY GENERATION USED IN STAGE 1.

| System Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| You are a medical query analyzer. Your task is to decompose a complex patient query into multiple distinct sub-queries, where each sub-query addresses a single, specific medical intention.                                                                                                                                                                                                                                                                    |
| User Prompt Template                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Given the following patient query:<br>{patient_query}<br>Please decompose this query into multiple sub-queries. Each sub-query should:<br>1. Address a single, distinct medical intention<br>2. Be self-contained and understandable without external context<br>3. Preserve essential medical context from the original text<br>4. Use clear, concise language<br><br>Output format: Return a JSON list of sub-queries:<br>["sub-query 1", "sub-query 2", ...] |

covering pseudo-queries that are off-topic, non-actionable, duplicate, too vague, or not related to a stable medical or psychological-support intent. The prototype wording is fixed after physician review and is not optimized on the test set. The validation split is used only to calibrate matching thresholds.

Given a pseudo-query  $p$ , the score of intent  $i$  is computed as  $S_i(p) = \max_{r \in \mathcal{P}_i} \cos(e(p), e(r))$ , where  $\mathcal{P}_i$  denotes the prototype set of intent  $i$  and  $e(\cdot)$  denotes the BioBERT embedding. Let  $c$  be the best clinical intent,  $S_c$  be its score,  $S_2$  be the second-highest clinical score, and  $S_{\text{None}}$  be the score of the None category. The pseudo-query is assigned to  $c$  only when  $S_c \geq \tau_s$ ,  $S_c - S_2 \geq \tau_m$ , and  $S_c - S_{\text{None}} \geq \tau_n$ . Otherwise, it is marked as None or uncertain instead of being

**Algorithm 1** Recursive Priority Grouping

---

```

1: function RPG-ORCHESTRATION( $I, P, C, D$ )
2:    $R \leftarrow I$ 
3:   while  $R \neq \emptyset$  do
4:      $i^* \leftarrow \text{LLM-Priority}(R, P, C)$ 
5:      $T \leftarrow D[i^*] \cap R$ 
6:      $S \leftarrow \text{LLM-Sequence}(T \cup \{i^*\}, P, C)$ 
7:      $\text{ProcessGroup}(S)$ 
8:      $R \leftarrow R \setminus S$ 
9:   end while
10: end function

```

---

forced into a clinical intent. This rejection mechanism reduces false positives for ambiguous or off-topic fragments. For safety-sensitive intents, such as emergency guidance, medication safety review, and medical psychological support, low-confidence candidates are preserved with an uncertainty flag when their scores are close to the threshold, so that downstream agents can handle them conservatively.

*D. Dynamic Intent Orchestration Agent*

We propose a Dynamic Intent Orchestration (DIO) agent. It is an adaptive system for dynamically arranging 14 fundamental medical intents, including intents related to psychological support and lifestyle guidance. The DIO agent addresses a key challenge in clinical decision support systems and in mental-health-related interactions. It helps complex, multi-step medical workflows adapt to evolving patient conditions and emotional states while still following clinical best practices.

Without proper orchestration, healthcare delivery can suffer from what Ancker *et al.* [59] termed “*alert fatigue*”. Clinicians and AI assistants may become overwhelmed by non-prioritized information and may overlook critical interventions. They also face attention fragmentation [60], a cognitive burden that can lead to fatigue and medical errors. These risks are particularly serious in emotionally sensitive settings, such as mental health consultations, where safety-critical and psychologically salient intents must be recognized and addressed in a timely manner.

Diverging from static medical workflows, the proposed DIO agent uses large language models to evaluate execution sequences in a flexible way. It considers dynamic clinical factors and real-time medical context. It also incorporates patient-specific conditions, including those related to psychological distress and support needs, to improve clinical safety and efficiency. The agent employs a Recursive Priority Grouping (RPG) methodology, which we formalize in Algorithm 1.

**Correctness and Acyclicity Analysis.** We analyze the correctness of the proposed RPG algorithm. Let the intent dependency structure be denoted as a directed graph  $\mathcal{G} = (\mathcal{I}, \mathcal{E})$ , where  $\mathcal{I}$  is the set of medical intents and  $(i_j, i_k) \in \mathcal{E}$  indicates that intent  $i_j$  must be processed before intent  $i_k$ . The dependency table  $D$  is constructed from clinical precedence relations and is further reviewed by board-certified physicians. Before orchestration, we validate  $\mathcal{G}$  by topological sorting. If a cycle is detected, the dependency table is rejected and revised before execution. Therefore, Algorithm 1 operates only on a verified directed

**Algorithm 2** Hierarchical Progressive Decision-making

---

```

Require: Patient query  $Q$ , Ordered intent groups  $\{G_1, G_2, \dots, G_T\}$ 
Ensure: Final medical advice  $A_{\text{final}}$ 
1: Initialize historical evidence  $\mathcal{H} \leftarrow \emptyset$  and stage reports  $\mathcal{R} \leftarrow \emptyset$ 
2: for  $t = 1$  to  $T$  do
3:   Retrieve current intent group  $G_t = \{i_1, i_2, \dots, i_k\}$ 
4:   Initialize group analyses  $\mathcal{A}_t \leftarrow \emptyset$ 
5:   for each intent  $i_j \in G_t$  do
6:      $\triangleright$  Sequential Intent Processing
7:      $r_j \leftarrow \text{ProcessIntent}(Q, i_j, \mathcal{H})$ 
8:      $\mathcal{A}_t \leftarrow \mathcal{A}_t \cup \{r_j\}$ 
9:   end for
10:   $\triangleright$  Progressive Evidence Accumulation
11:   $R_t \leftarrow \text{SynthesizeGroup}(\mathcal{A}_t, Q, \mathcal{H})$ 
12:   $\mathcal{H} \leftarrow \mathcal{H} \cup \{R_t\}$ 
13:   $\mathcal{R} \leftarrow \mathcal{R} \cup \{R_t\}$ 
14: end for
15:  $A_{\text{final}} \leftarrow \text{GenerateFinalAdvice}(Q, \mathcal{R}, \mathcal{H})$ 
16: return  $A_{\text{final}}$ 

```

---

acyclic graph. The algorithm satisfies both termination and dependency preservation. At each iteration, the remaining set  $R$  contains all unprocessed intents. The algorithm selects a highest-priority intent  $i^* \in R$ , retrieves its unprocessed prerequisites  $T = D[i^*] \cap R$ , and constructs a sequenced group  $S \subseteq R$  that contains  $i^*$ . Since  $i^* \in S$ , the remaining set is strictly reduced after each iteration, i.e.,  $R \leftarrow R \setminus S$ . Thus, the algorithm terminates after at most  $|\mathcal{I}|$  iterations. In addition, since  $\mathcal{G}$  is acyclic, each group  $S$  admits a valid topological order. As a result, every intent is processed only after its prerequisite intents have either been included earlier in the same group or already been processed in previous groups.

The RPG algorithm follows a clinically intuitive strategy. It recursively identifies the highest-priority intent  $i^*$  from the remaining set of intents  $R$ . It then consults a dependency table  $D$  to determine prerequisite intents  $T$  that must be processed before  $i^*$ . This allows the system to group and execute intents in an order that respects clinical logic and safety considerations. In scenarios that involve mental health or emotional distress, this mechanism can place risk assessment and psychological support earlier in the workflow when needed. The dependency table  $D$  is structured as a directed acyclic graph (DAG). Each intent maps to its prerequisite intents:

$$D[i_k] = \{i_j \in \mathcal{I} \mid i_j \text{ must be processed before } i_k\}.$$

We construct this table in two phases. First, we extract fundamental prerequisite relationships from established clinical diagnostic protocols [61], [62]. Second, three board-certified physicians, specializing in internal medicine, emergency medicine, and clinical decision support, validate and refine these relationships. This process helps ensure that the dependency structure reflects real-world clinical reasoning in both general medical and mental health scenarios.

The design of  $D$  is dynamically extensible. When new medical or mental health intents are introduced, only the

dependency table needs to be updated by specifying  $\mathcal{D}[i_{\text{new}}]$ . The core MIRA framework does not require modification. The DAG structure automatically checks that no cycles are introduced, *i.e.*,  $i_{\text{new}} \notin \text{Ancestors}(\mathcal{D}[i_{\text{new}}])$ . This guarantees that the orchestration algorithm remains sound while supporting future expansion of medical and psychological-support intents. Complete dependency specifications and clinical rationale are provided in Table I.

### E. Multi-Agent Collaboration System

Our framework employs a Hierarchical Progressive Decision-making (HPD) agent designed to orchestrate the entire diagnostic reasoning process. Operating as the system's cognitive core, the HPD agent sequentially processes intent groups provided by the DIO agent. This approach emulates the systematic workflow of clinical diagnosis, where physicians meticulously gather prerequisite information before making high-stakes decisions.

Just as human doctors require a complete patient history, symptom analysis, and test results before recommending treatment plans, our HPD agent progressively builds medical reasoning by processing intents in their logical dependency order. This hierarchical mechanism ensures that each diagnostic step is grounded in previously established medical evidence, thereby minimizing the risk of premature or unsupported conclusions. The complete workflow of this decision-making process is formalized in Algorithm 2.

**Sequential Intent Processing.** As detailed in Algorithm 2, when the DIO agent dispatches an intent group  $\mathcal{G}_t = \{i_1, i_2, \dots, i_k\}$  at stage  $t$ , the HPD agent processes each intent strictly according to the predetermined execution order. For each intent  $i_j \in \mathcal{G}_t$ , the agent generates an intent-specific analysis  $r_j$ :

$$r_j = \text{HPD}(Q, i_j, \mathcal{H}_{<t}), \quad (4)$$

where  $Q$  represents the original patient query and  $\mathcal{H}_{<t}$  denotes the historical reasoning outputs accumulated from all previous stages. This formulation explicitly captures the dependency on prior diagnostic information. It ensures that each new inference is contextualized by previously established medical findings, preventing logical hallucinations.

**Progressive Evidence Accumulation.** Following the individual analysis of intents, the system enters the synthesis phase. The HPD agent aggregates all reasoning outputs within the current group to generate a stage-specific medical report  $\mathcal{R}_t$ :

$$\mathcal{R}_t = \text{Synthesize}(\{r_1, r_2, \dots, r_k\}). \quad (5)$$

This report serves a dual purpose: (1) it provides a cohesive intermediate diagnostic summary for the current stage, and (2) it is seamlessly integrated into the evidence base, such that  $\mathcal{H}_t = \mathcal{H}_{<t} \cup \mathcal{R}_t$ . This progressive accumulation is critical for handling complex medical cases. It guarantees that later diagnostic steps—such as treatment recommendations—are computationally conditioned on the verified findings of earlier stages, such as symptom analysis and lab test interpretations.

**Final Decision Synthesis.** Once all intent groups have been processed and the evidence base  $\mathcal{H}$  is fully populated, the agent executes the final decision-making step. By synthesizing the comprehensive stage reports  $\mathcal{R}$  and the accumulated history

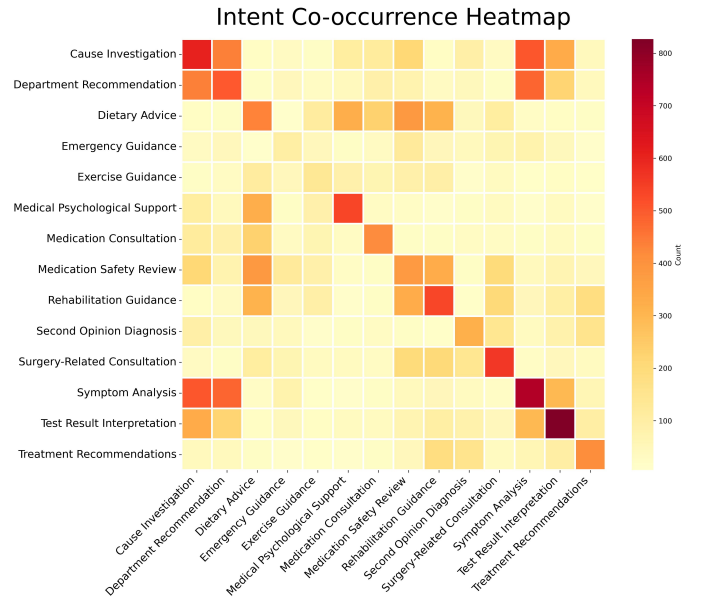


Fig. 4. Heatmap visualization of medical intent co-occurrence patterns based on the physician-validated 14-category intent taxonomy. The taxonomy was constructed through a clinical-workflow-guided and data-informed process rather than by direct ICD-10 mapping. The color intensity indicates how frequently two predefined intent categories appear together in patient queries.

$\mathcal{H}$ , the HPD agent formulates the final medical advice  $A_{\text{final}}$ . This ensures that the ultimate response is holistic, logically consistent, and clinically safe.

**Hallucination Mitigation and Safety Handling.** The HPD agent mainly reduces logical hallucinations by ensuring that each reasoning step is conditioned on the previously accumulated evidence. However, structured reasoning alone cannot fully eliminate factual medical hallucinations from the underlying LLM. Such hallucinations may include invented clinical findings, unsupported medication statements, inaccurate risk estimates, or overconfident diagnostic claims. To mitigate this risk, MIRA adopts a conservative evidence-anchored generation strategy. During final synthesis, the HPD agent is instructed to ground its advice in the original patient query, the recognized intents, and the accumulated stage reports. Medical claims that are not supported by the available context are removed or expressed with explicit uncertainty. The final response is also constrained to avoid definitive diagnosis, precise medication adjustment, or unsupported treatment recommendation without professional evaluation. For high-risk situations, such as emergency symptoms, medication safety concerns, or severe psychological distress, MIRA prioritizes safety-oriented guidance and recommends timely consultation with qualified healthcare professionals. Therefore, HPD reduces the risk of both inconsistent reasoning and unsupported medical advice, while the system remains a decision-support framework rather than a replacement for clinical judgment.

1) *Bounded Execution and Safety Control:* To avoid uncontrolled multi-agent reasoning loops, MIRA adopts a bounded execution strategy. Since the DIO agent operates on a finite set of recognized intents, the maximum number of intent groups is bounded by the number of intent categories. Each expert agent is executed once under normal conditions, with at most one

TABLE III  
PROMPT TEMPLATE USED FOR GENERATING REALISTIC, VERBOSE PATIENT  
QUERIES.

| System Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are simulating a patient describing their medical situation. Based on the provided medical record, generate a query that:</p> <ol style="list-style-type: none"> <li>1. Naturally embeds multiple medical intentions.</li> <li>2. Uses conversational, non-technical language.</li> <li>3. Includes relevant context and background.</li> <li>4. Reflects realistic patterns (<i>e.g.</i>, slight disorganization, tangential details).</li> <li>5. Maintains clinical accuracy despite verbosity.</li> </ol> |
| User Prompt Template                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <p>Generate a realistic patient query based on:</p> <p><b>Patient Background:</b> {patient_background}</p> <p><b>Chief Complaints:</b> {chief_complaints}</p> <p><b>Medical History:</b> {medical_history}</p> <p><b>Current Medications:</b> {current_medications}</p> <p><b>Recent Events:</b> {recent_events}</p> <p><b>Patient Concerns:</b> {patient_concerns}</p> <p><i>Task:</i> Generate a verbose query (150-300 words) incorporating these elements and expressing multiple medical intentions.</p>        |

refinement request when the generated report is incomplete or violates the required output format.

The HPD agent serves as the final safety gate before response generation. The system avoids unsupported diagnosis, unsafe medication modification, and overconfident medical claims. For high-risk cases involving emergency symptoms, medication safety concerns, or severe psychological distress, the final response prioritizes escalation guidance and professional medical consultation.

2) *Agent Roles and Execution Protocol:* The multi-agent collaboration system in MIRA follows a coordinator-mediated execution protocol rather than unrestricted peer-to-peer discussion. The HPD agent serves as the central coordinator, while four intent-specific agents are activated according to the intent groups generated by the DIO agent. The four specialized agents correspond to four clinical stages in the intent taxonomy. The Diagnosis and Assessment Agent handles symptom analysis, cause investigation, test result interpretation, department recommendation, and second opinion diagnosis. The Treatment-Care Agent focuses on treatment recommendations, medication consultation, medication safety review, and surgery-related consultation. The Recovery-Support Agent processes rehabilitation guidance and exercise guidance. The Lifestyle-Guide Agent addresses dietary advice, medical psychological support, and lifestyle-related concerns. Each

expert agent receives five types of information: the assigned intent, original patient query, previous stage evidence, agent-specific role instruction, and safety constraints. Instead of generating final responses independently, each agent produces a structured report containing key findings, uncertainties, risk factors, suggested next steps, and information required by the HPD agent.

The communication among agents is mediated by the HPD agent. Specifically, the DIO agent first determines the execution order of intent groups according to clinical dependencies and safety priorities. For each intent group, the HPD agent dispatches corresponding expert agents with the original query and accumulated evidence. After receiving the structured reports, the HPD agent performs consistency checking, evidence integration, and stage-level synthesis. The accumulated stage reports are then passed to subsequent reasoning stages until all intent groups are processed.

#### IV. EXPERIMENTS

In this section, we first introduce the dataset used in our experiments. We then describe the experimental setup and evaluation metrics. Then, we present the results from testing our framework on public benchmarks to evaluate its overall effectiveness. Finally, we conduct multiple ablation study to validate the accuracy of intent recognition and assess the necessity of each component in the framework.

##### A. Dataset Collection and Construction

To address the scarcity of resources for complex query understanding, we construct the Mental Multi-Intention (MMI) dataset. The corpus comprises 2,200 samples derived from authentic, de-identified medical records. We meticulously process these records to preserve clinical complexity while simulating realistic patient-physician interactions.

1) *The MMI Construction Pipeline:* Our process follows a systematic three-stage pipeline. This ensures the dataset reflects the multi-faceted nature of real-world consultations.

**Stage 1: Record Selection.** We do not select records randomly. Instead, we filter authentic medical cases based on strict criteria. First, records must possess *Clinical Completeness*, containing a full trajectory from diagnosis to treatment. Second, we enforce *Multi-Intent Richness*. Selected cases naturally involve multiple medical intentions, averaging 3 intents per sample. We ensure diversity across medical specialties and complexity levels.

**Stage 2: Verbose Query Generation.** Raw medical records differ significantly from natural patient language. To bridge this gap, we employ GPT-4.5 to simulate patient descriptions based on the authentic diagnostic trajectories. This step is crucial. It captures the verbose, often disorganized nature of real patient communication. The generation process utilizes the specific prompt design detailed in Table III. This strategy ensures the generated text remains clinically accurate while reflecting natural linguistic patterns and implicit intentions.

**Stage 3: Expert Annotation.** We validate all generated samples through a rigorous physician-involved annotation process. Initially, the system automatically proposes intent labels according to the predefined 14-category taxonomy. These

TABLE IV  
DISTRIBUTION OF MEDICAL INTENTS IN THE MMI DATASET.

| Intent Category                   | Freq. | %    |
|-----------------------------------|-------|------|
| <i>Diagnosis &amp; Assessment</i> |       |      |
| Symptom Analysis                  | 1847  | 83.9 |
| Cause Investigation               | 1456  | 66.2 |
| Test Result Interpretation        | 892   | 40.5 |
| Department Recommendation         | 734   | 33.4 |
| Second Opinion Diagnosis          | 445   | 20.2 |
| <i>Treatment-Care</i>             |       |      |
| Treatment Recommendations         | 1523  | 69.2 |
| Medication Consultation           | 1389  | 63.1 |
| Medication Safety Review          | 1067  | 48.5 |
| Surgery-Related Consultation      | 312   | 14.2 |
| <i>Recovery &amp; Lifestyle</i>   |       |      |
| Rehabilitation Guidance           | 678   | 30.8 |
| Medical Psychological Support     | 423   | 19.2 |
| Dietary Advice                    | 1234  | 56.1 |
| Exercise Guidance                 | 845   | 38.4 |
| Emergency Guidance                | 567   | 25.8 |

labels serve only as initial annotations and are independently reviewed by three board-certified physicians. Each physician examines whether the assigned intents accurately represent the underlying patient needs and can add, remove, or modify labels when necessary. For cases with inconsistent annotations, disagreements are resolved through discussion among the physicians to obtain the final consensus labels. To further evaluate annotation reliability, we randomly sample 300 cases and calculate the inter-annotator agreement before consensus resolution. The Fleiss'  $\kappa$  score reaches 0.82, demonstrating substantial agreement among expert annotators and supporting the reliability of the MMI annotation process.

2) *Hybrid Data Composition*: To complement the MMI dataset and ensure broad coverage, we incorporate an additional 2,200 samples from HuaTuo-26M [63]. We carefully select these samples to balance the representation of diverse consultation scenarios. Figure 3 illustrates the intent distribution across both sources, demonstrating their complementary nature. MMI excels in multi-intent complexity, while HuaTuo-26M provides breadth across general medical topics.

3) *Statistical Characteristics and Ethics*: Our analysis reveals distinct patterns in patient inquiries. As shown in Figure 4 of the intent co-occurrence heatmap, symptom analysis frequently appears alongside department recommendations and cause investigations. These combinations highlight the interconnected nature of real-world medical needs. Table IV details the specific distribution of intents within the MMI dataset, showing a strong prevalence of symptom analysis (83.9%) and treatment recommendations (69.2%).

Regarding ethics, we strictly adhere to privacy guidelines. We remove all personally identifiable information from the source records. Furthermore, the generated queries contain no traceable private data. Although all source records were de-identified, the MMI dataset may still reflect biases in the original clinical data, such as demographic imbalance, regional healthcare access,

TABLE V  
DECODING CONFIGURATIONS USED IN EXPERIMENTS.

| Component            | Temperature | Top-p | Max Tokens |
|----------------------|-------------|-------|------------|
| Baseline generation  | 0.2         | 0.95  | 2,048      |
| MIRA agents          | 0.2         | 0.95  | 2,048      |
| CoT-SC sampling      | 0.7         | 0.95  | 2,048      |
| MMI query generation | 0.7         | 0.95  | 512        |
| LLM evaluation       | 0.0         | 1.0   | 1,024      |

disease distribution, and hospital-specific practice patterns. We mitigated this issue by selecting cases across different specialties, complexity levels, and intent categories. Future work will further improve dataset representativeness through multi-center collaboration and demographic auditing under appropriate privacy and ethical constraints.

### B. Experiment Setup and Evaluation

1) *Compared Baselines*: We benchmark MIRA against a diverse set of state-of-the-art approaches. All experiments are conducted using GPT-4.5 [58] accessed through the official API released by OpenAI. The experiments were performed using the API version available in January 2025. The relevant decoding configurations are listed in Table V. We categorize these methods into two groups:

**Single-Agent Methods.** These approaches rely on a single model instance with varying prompting strategies.

- **Zero-shot**: Performs direct inference using only the task instruction. It lacks task-specific examples.
- **Few-shot**: Enhances performance by incorporating a limited set of in-context demonstrations to guide the model.
- **CoT** [54]: Extends Few-shot by explicitly generating intermediate reasoning steps before deriving the final answer.
- **CoT-SC** [55]: Improves robustness through self-consistency. It generates multiple reasoning chains and selects the final output via majority voting.
- **EnRe** [64]: Uses Ensemble Refinement. It aggregates outputs from diverse reasoning paths to produce a more reliable consensus.

**Multi-Agent Methods.** These approaches leverage collaboration between specialized agents.

- **MedAgents** [25]: Deploys five distinct expert agents. They engage in interactive discussions to solve medical multiple-choice questions.
- **MDAgents** [15]: Examined in three variants to test different collaboration depths: (a) **Base** (single agent), (b) **Collaboration** (three interacting agents), and (c) **Group** (structured group discussions).

2) *More Details*: For fair comparison, all methods receive the same patient query and the same output constraints, while only the reasoning or collaboration instructions differ according to each method. For MIRA, each agent follows a structured prompt format containing five components: agent role, target intent, original patient query, accumulated evidence from previous stages, and safety constraints. The intent-specific

TABLE VI  
COMPARISON OF MIRA AGAINST RULE-BASED LLM METHODS AND MULTI-AGENT SYSTEMS ON THE HUATUO-26M DATASET.

| Category     | Methods                     | Huatuo-26M Dataset   |                   |                  |             |           |
|--------------|-----------------------------|----------------------|-------------------|------------------|-------------|-----------|
|              |                             | Intent Comprehension | Clinical Planning | Response Quality | Average     | API calls |
| Single-Agent | Zero-shot                   | 3.80                 | 3.75              | 3.74             | 3.76        | 1.0       |
|              | Few-shot                    | 4.19                 | 3.64              | 4.03             | 3.95        | 1.0       |
|              | + CoT [54]                  | <u>4.12</u>          | 3.94              | 4.01             | 4.02        | 1.0       |
|              | + CoT-SC [55]               | 4.02                 | <u>4.11</u>       | 3.94             | 4.02        | 6.0       |
|              | EnRe [64]                   | 4.11                 | 3.73              | 4.02             | 3.95        | 12.0      |
| Multi-Agent  | MDAgents-Base [15]          | 3.77                 | 3.94              | 3.63             | 3.89        | 1.0       |
|              | MedAgents [25]              | 3.83                 | 3.99              | 3.85             | 3.78        | 19.3      |
|              | MDAgents-Collaboration [15] | 3.88                 | 4.02              | 3.81             | 3.89        | 9.8       |
|              | MDAgents-Group [15]         | 4.04                 | <u>4.11</u>       | <u>4.12</u>      | <u>4.09</u> | 42.2      |
|              | <b>MIRA (ours)</b>          | <b>4.76</b>          | <b>4.32</b>       | <b>4.62</b>      | <b>4.57</b> | 10.7      |

agents generate structured reports containing key findings, uncertainties, risk factors, suggested next steps, and information required by the HPD agent. The complete prompt templates for MIRA agents and all baseline methods are provided in the supplementary material.

3) *Evaluation Dimensions*: Clinical decision-making is multifaceted. We therefore evaluate performance across three distinct dimensions:

- **Intent Comprehension**: This metric goes beyond surface-level text matching. It measures the system’s capacity to identify implicit needs and contextual subtleties. High scores indicate the model successfully uncovers the hidden drivers behind a patient’s query.
- **Clinical Planning**: This assesses the strategic orchestration of care. It evaluates whether the model creates a coherent, prioritized reasoning pathway. A valid plan must align with medical best practices while strictly adhering to safety protocols.
- **Response Quality**: This evaluates the final output’s utility. We check for precision, clarity, and actionability. Responses must be grounded in evidence-based medicine and practical for real-world application.

4) *Scoring Protocol*: We adopt a hybrid evaluation strategy combining automated ensembles with human expertise.

**Automated LLM Ensemble**. Relying on a single LLM judge can introduce bias. To mitigate this, we employ an ensemble of three state-of-the-art models: GPT-4.5 [58], Gemini-2.0 Pro [65], and Claude Sonnet 4.5 [66]. Each model independently evaluates system outputs across the three dimensions using a strict 5-point Likert scale. We calculate the final score for each dimension by averaging the ratings from all three evaluators. Let  $\text{Score}_i$  denote the rating from the  $i$ -th model:

$$\text{Score}_{\text{final}} = \frac{1}{3} \sum_{i=1}^3 \text{Score}_i. \quad (6)$$

The above ensemble approach ensures a more objective and robust assessment of model performance.

**Human Expert Validation**. To ground our automated metrics in clinical reality, we complement the LLM scores with human

review. Board-certified medical experts assess a subset of responses, specifically focusing on clinical accuracy and safety. This dual-validation framework ensures our results remain clinically valid.

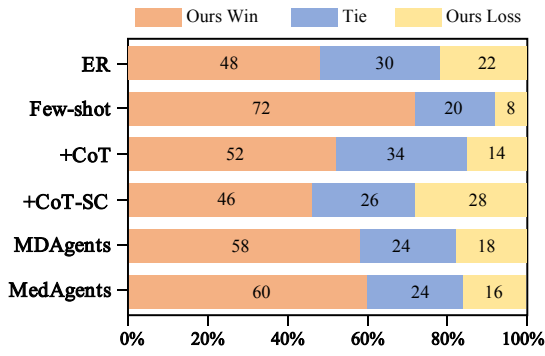
### C. Experimental Results

1) *Performance on HuaTuo-26M Benchmark*: In Table VI, the evaluation on the HuaTuo-26M dataset reflects performance patterns in standard medical consultation scenarios. In this environment, MIRA maintains a leadership position, outperforming the strongest baseline (*i.e.*, MDAgents-Group) by a margin of 8.4%. However, the performance gaps are notably narrower compared to complex settings. We attribute this to the relatively straightforward nature of HuaTuo-26M queries, which typically involve single-turn, single-intent interactions. Consequently, even single-agent strategies equipped with Chain-of-Thought (CoT) prompting achieve competitive scores in Response Quality, benefiting from the lower reasoning overhead. Despite this saturation in general QA metrics, MIRA still secures an 11.7% improvement in Clinical Planning. This suggests that even for general inquiries, our hierarchical intent modeling provides a more structured and safer diagnostic pathway than vanilla multi-agent discussions.

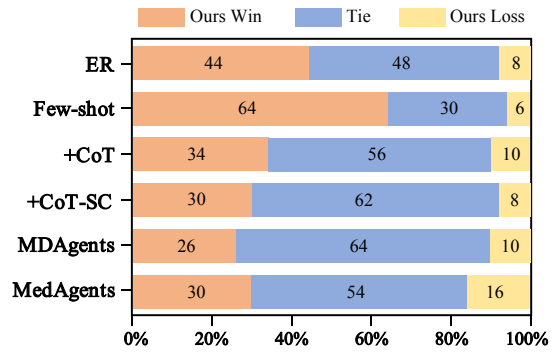
2) *Performance on MMI Dataset*: As shown in Table VII, the experimental landscape shifts significantly. The overall absolute scores for baseline methods drop perceptibly compared to HuaTuo-26M, highlighting the complexity penalty introduced by the intertwined, multi-intent narratives characteristic of our dataset. In this rigorous testing ground, single-agent methods struggle to maintain context, often hallucinating or missing secondary intents due to entanglement. Conversely, MIRA demonstrates its robustness, widening the performance gap to a substantial 12.1% absolute improvement over MDAgents-Group. While collaboration-based baselines achieve incremental gains through iterative debate, they lack the structural capacity to disentangle overlapping clinical needs. MIRA overcomes this through dynamic intent disentanglement and parallel reasoning, proving that complex clinical narratives require structured orchestration rather than simple model scaling.

TABLE VII  
COMPARISON OF MIRA AGAINST RULE-BASED LLM METHODS AND MULTI-AGENT SYSTEMS ON THE MMI DATASET.

| Category     | Methods                     | MMI Dataset          |                   |                  |             |           |
|--------------|-----------------------------|----------------------|-------------------|------------------|-------------|-----------|
|              |                             | Intent Comprehension | Clinical Planning | Response Quality | Average     | API calls |
| Single-Agent | Zero-shot                   | 4.02                 | 3.88              | 3.83             | 3.91        | 1.0       |
|              | Few-shot                    | 4.31                 | 3.91              | 4.05             | 4.09        | 1.0       |
|              | + CoT [54]                  | 4.39                 | 4.15              | 4.12             | 4.22        | 1.0       |
|              | + CoT-SC [55]               | 4.37                 | 4.32              | 4.10             | 4.26        | 6.0       |
|              | EnRe [64]                   | 4.26                 | 3.91              | 3.98             | 3.99        | 12.0      |
| Multi-Agent  | MDAgents-Base [15]          | 4.15                 | 4.26              | 3.85             | 4.09        | 1.0       |
|              | MedAgents [25]              | 4.19                 | 4.31              | 4.03             | 4.18        | 21.8      |
|              | MDAgents-Collaboration [15] | 4.27                 | 4.33              | 4.02             | 4.21        | 11.9      |
|              | MDAgents-Group [15]         | 4.43                 | 4.49              | 4.28             | 4.40        | 45.7      |
|              | <b>MIRA (ours)</b>          | <b>4.89</b>          | <b>4.75</b>       | <b>4.66</b>      | <b>4.77</b> | 17.3      |



(a) MMI Q&A Evaluation



(b) HuaTuo-26M Q&A Evaluation

Fig. 5. Comparison of MIRA against rule-based LLM methods and multi-agent systems via the doctor evaluation. MDAgents [15] employs its group setting.

TABLE VIII  
STATISTICAL SIGNIFICANCE ANALYSIS BETWEEN MIRA AND THE STRONGEST OVERALL BASELINE, MDAgENTS-GROUP.

| Dataset    | Metric         | Mean Diff. | 95% CI       | Corrected $p$ | Effect Size |
|------------|----------------|------------|--------------|---------------|-------------|
| MMI        | Intent Comp.   | 0.46       | [0.39, 0.53] | < 0.001       | 0.62        |
| MMI        | Clinical Plan. | 0.26       | [0.20, 0.32] | < 0.001       | 0.45        |
| MMI        | Response Qual. | 0.38       | [0.31, 0.45] | < 0.001       | 0.56        |
| MMI        | Average        | 0.37       | [0.32, 0.42] | < 0.001       | 0.60        |
| HuaTuo-26M | Intent Comp.   | 0.72       | [0.64, 0.80] | < 0.001       | 0.68        |
| HuaTuo-26M | Clinical Plan. | 0.21       | [0.15, 0.27] | < 0.001       | 0.39        |
| HuaTuo-26M | Response Qual. | 0.50       | [0.42, 0.58] | < 0.001       | 0.59        |
| HuaTuo-26M | Average        | 0.48       | [0.42, 0.54] | < 0.001       | 0.63        |

3) *Statistical Significance Analysis*: To further verify whether the observed improvements are statistically reliable, we conduct paired significance tests between MIRA and the strongest overall baseline, MDAgents-Group. Since all methods are evaluated on the same set of patient queries, the comparison is naturally paired at the sample level. We use the Wilcoxon signed-rank test as the primary test because the evaluation scores are based on a 5-point Likert scale. Holm-Bonferroni correction is applied to account for multiple comparisons. We also estimate 95% confidence intervals by bootstrap resampling with 1,000 iterations and report rank-biserial correlation as the

effect size. As shown in Table VIII, MIRA achieves statistically significant improvements over MDAgents-Group across all metrics on both datasets. On the MMI dataset, MIRA improves the average score by 0.37 points, with a 95% confidence interval of [0.32, 0.42] and a corrected  $p$ -value smaller than 0.001. On the HuaTuo-26M dataset, MIRA improves the average score by 0.48 points, with a 95% confidence interval of [0.42, 0.54] and a corrected  $p$ -value smaller than 0.001. These results indicate that the improvements of MIRA are not only reflected in mean scores but are also statistically reliable.

4) *Expert Doctor Evaluation*: To validate clinical utility beyond automated metrics, three board-certified physicians conducted a blinded pairwise comparison. They assessed 1,200 instances (*i.e.*, 100 randomly selected samples per dataset against six baselines) through majority voting. Figure 5 illustrates a striking divergence in model distinctiveness. On the MMI benchmark, MIRA achieves a decisive dominance, recording a 72% win rate against Few-shot baselines with only 8% losses. Even against sophisticated reasoning-enhanced methods (CoT, CoT-SC) and collaborative agents (MDAgents, MedAgents), MIRA maintains win rates between 46% and 60%, validating its superiority in handling complex cases. In contrast, the HuaTuo-26M benchmark yields high tie rates across all comparisons (56~64% for agent approaches). The

TABLE IX  
FOCUSED EVALUATION ON THE MEDICAL PSYCHOLOGICAL SUPPORT SUBSET. SCORES ARE OBTAINED THROUGH BLINDED PHYSICIAN EVALUATION USING A 5-POINT RUBRIC.

| Method              | Empathy | Emotional App. | Crisis Handling | Therapeutic Safety | Average |
|---------------------|---------|----------------|-----------------|--------------------|---------|
| GPT-4.5 Direct      | 3.84    | 3.79           | 3.61            | 3.70               | 3.74    |
| GPT-4.5 + CoT       | 4.03    | 3.98           | 3.82            | 3.88               | 3.93    |
| MedAgents [25]      | 4.00    | 3.95           | 3.86            | 3.91               | 3.93    |
| MDAgents-Group [15] | 4.12    | 4.08           | 3.95            | 4.01               | 4.04    |
| <b>MIRA (ours)</b>  | 4.51    | 4.47           | 4.38            | 4.42               | 4.45    |

TABLE X  
SAFETY-ORIENTED EVALUATION ON THE CRISIS-SENSITIVE SUBSET. RESULTS ARE BASED ON PHYSICIAN-ANNOTATED SAFETY FAILURE RATES. LOWER VALUES INDICATE SAFER RESPONSES.

| Method              | Missing Escalation | Unsafe Reassurance | Incomplete Risk Framing | Overall Failure Rate |
|---------------------|--------------------|--------------------|-------------------------|----------------------|
| GPT-4.5 Direct      | 18.6%              | 14.1%              | 22.4%                   | 18.4%                |
| GPT-4.5 + CoT       | 13.5%              | 10.3%              | 17.3%                   | 13.7%                |
| MedAgents [25]      | 12.8%              | 9.6%               | 16.7%                   | 13.0%                |
| MDAgents-Group [15] | 10.9%              | 8.3%               | 14.1%                   | 11.1%                |
| <b>MIRA (ours)</b>  | 5.8%               | 4.5%               | 8.3%                    | 6.2%                 |

TABLE XI  
PERFORMANCE OF DIFFERENT INTENT RECOGNITION METHODS. UNDERLINED VALUES INDICATE THE HIGHEST VALUES IN EACH METRIC, EXCLUDING OUR RESULTS.

| Methods            | Accuracy    | Precision   | Recall      | F1 Score    |
|--------------------|-------------|-------------|-------------|-------------|
| EnRe [64]          | 0.75        | 0.43        | 0.64        | 0.50        |
| MedAgent [25]      | 0.71        | 0.46        | 0.70        | 0.55        |
| Zero-Shot          | 0.75        | 0.44        | 0.55        | 0.47        |
| Few-Shot           | <u>0.79</u> | <u>0.61</u> | 0.72        | 0.64        |
| + Cot [54]         | 0.77        | 0.56        | 0.79        | 0.64        |
| + Cot-SC [55]      | 0.77        | 0.57        | <u>0.82</u> | <u>0.66</u> |
| MDAgent [15]       | 0.68        | 0.36        | 0.61        | 0.44        |
| + Collaboration    | 0.77        | 0.56        | 0.81        | 0.65        |
| + Group            | 0.76        | 0.53        | <u>0.82</u> | 0.64        |
| <b>MIRA (ours)</b> | <b>0.90</b> | <b>0.82</b> | <b>0.84</b> | <b>0.82</b> |

observed disparity between the 20~34% tie rate on MMI and the 48~64% rate on HuaTuo confirms that standard datasets often fail to differentiate advanced model capabilities. The MMI dataset, with its nuanced clinical queries, serves as a more effective discriminator for high-level medical reasoning.

5) *Evaluation on Psychological-support Scenarios:* Since MIRA targets emotionally sensitive healthcare interactions, we further evaluate its ability to preserve and respond to psychological-support needs. We construct a Medical Psychological Support (MPS) subset containing all 423 samples annotated with the corresponding intent. Unlike the general evaluation protocol, this analysis focuses on affective-care-related capabilities. Three board-certified physicians independently evaluate the generated responses under a blinded setting using a physician-designed 5-point rubric. The evaluation considers four dimensions: Empathy, Emotional Appropriateness, Crisis Handling, and Therapeutic Safety. These criteria measure whether the response acknowledges patient distress, appropriately adapts to emotional contexts, identifies potential

TABLE XII  
COMPARISON WITH REPRESENTATIVE BERT-BASED AND ERNIE-BASED INTENT RECOGNITION METHODS ON THE MMI INTENT DETECTION TASK.

| Method                     | Accuracy    | Precision   | Recall      | F1 Score    |
|----------------------------|-------------|-------------|-------------|-------------|
| BERT-MLC                   | 0.80        | 0.64        | 0.70        | 0.66        |
| Uncertainty Detection [36] | 0.78        | 0.67        | 0.65        | 0.65        |
| MISCA [37]                 | 0.83        | 0.71        | 0.75        | 0.72        |
| Conco-ERNIE [35]           | 0.85        | 0.74        | 0.76        | 0.75        |
| JPIS [38]                  | 0.82        | 0.70        | 0.73        | 0.71        |
| <b>MIRA (ours)</b>         | <b>0.90</b> | <b>0.82</b> | <b>0.84</b> | <b>0.82</b> |

risks, and avoids unsafe psychological guidance.

Table IX summarizes the results. MIRA achieves consistent improvements over existing baselines across all dimensions. These results indicate that explicit psychological-support intent modeling and safety-aware orchestration help MIRA generate responses that better address emotional needs while maintaining clinical safety.

Furthermore, to evaluate safety-critical interactions, we construct a crisis-sensitive subset containing 156 samples with Medical Psychological Support together with Emergency Guidance or explicit high-risk expressions. We measure three failure types: missing escalation advice, unsafe reassurance, and incomplete risk framing. Lower failure rates indicate safer responses. As shown in Table X, MIRA achieves the lowest overall failure rate among all methods. This improvement demonstrates that the proposed intent recognition and orchestration mechanisms help preserve high-risk psychological and medical signals embedded in complex patient narratives.

These results demonstrate that MIRA contributes to affective computing not through explicit emotion classification, but through affect-aware intention understanding and safety-sensitive response planning in healthcare conversations.

6) *Computational Efficiency Analysis:* Efficiency is as critical as accuracy in real-world deployment. Debate-based multi-agent frameworks, such as MedAgents (N=5) and MDAgents-

TABLE XIII

ADDITIONAL ZERO-SHOT GENERALIZATION RESULTS ON EXTERNAL DATA SOURCES AND REAL-RECORD HOLDOUT CASES. ALL METHODS USE GPT-4.5 AS THE BACKBONE. MIRA IS EVALUATED WITHOUT ADDITIONAL FINE-TUNING ON THESE DATASETS.

| Setting                   | Domain               | Lang.   | Size | Method         | Intent Comp. | Clinical Plan. | Average |
|---------------------------|----------------------|---------|------|----------------|--------------|----------------|---------|
| External medical dialogue | General medicine     | English | 500  | MDAgents-Group | 4.07         | 3.96           | 4.03    |
| External medical dialogue | General medicine     | English | 500  | MIRA           | 4.38         | 4.18           | 4.29    |
| Consumer health QA        | Patient consultation | English | 500  | MDAgents-Group | 4.10         | 3.98           | 4.06    |
| Consumer health QA        | Patient consultation | English | 500  | MIRA           | 4.35         | 4.14           | 4.25    |
| Mental health support     | Emotional support    | Chinese | 500  | MDAgents-Group | 3.94         | 3.89           | 3.97    |
| Mental health support     | Emotional support    | Chinese | 500  | MIRA           | 4.31         | 4.09           | 4.24    |
| Real-record holdout       | Mixed clinical cases | Chinese | 300  | MDAgents-Group | 4.18         | 4.07           | 4.15    |
| Real-record holdout       | Mixed clinical cases | Chinese | 300  | MIRA           | 4.54         | 4.32           | 4.43    |

Group, rely on iterative consensus mechanisms. The brute-force collaboration incurs high computational costs, averaging 21.8 and peaking at 45.7 API calls per query, respectively. MIRA adopts a precision-first approach. By structurally decomposing intents before execution, it reduces the need for redundant inter-agent dialogue. Consequently, MIRA requires only 17.3 average API calls while delivering superior diagnostic performance. The efficiency gain (*i.e.*, over  $2.5\times$  reduction compared to MDAgents-Group) confirms that solving multi-intent medical problems demands fine-grained, intent-driven processing rather than extensive, unstructured multi-agent discussions.

7) *Intent Recognition Comparison with Specialized Classifiers*: To further evaluate the effectiveness of MIRA in multi-intent recognition, we compare it with representative supervised intent recognition methods, including BERT-based and ERNIE-based approaches. These methods are adapted to the MMI taxonomy under the same train/test split. As shown in Table XII, MIRA achieves superior performance across all metrics. Compared with Conco-ERNIE, the strongest specialized intent recognition baseline, MIRA improves the F1 score from 0.75 to 0.82. This improvement demonstrates that explicit query decomposition and clinical prototype alignment are beneficial for handling verbose patient descriptions with intertwined medical needs.

#### D. Ablation Study

1) *Intent Detection Analysis*: Table XI confirms that MIRA establishes a new state-of-the-art across all intent detection metrics. Baseline methods exhibit distinct trade-offs. Few-shot prompting secures respectable accuracy and precision. Conversely, CoT-SC leads in recall, suggesting that explicit reasoning chains help uncover diverse, implicit intents. However, a critical bottleneck persists among these baselines: consistently low precision. They frequently generate redundant or hallucinated intents, failing to distinguish core medical needs from noise. While collaborative approaches like MDAgent-Collaboration improve recall through agent interaction, they still struggle to curb false positives. MIRA overcomes this limitation. By achieving significant gains in both precision and recall, our framework filters diagnostic noise while accurately capturing the full spectrum of true clinical intentions.

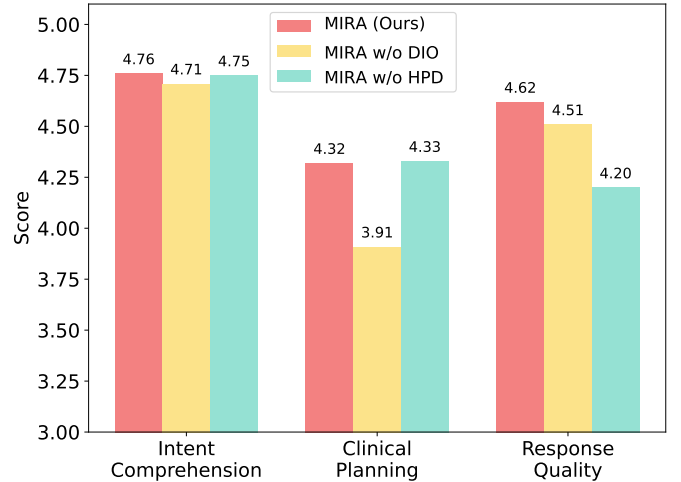


Fig. 6. Ablation results of the DIO agent and HPD agent on three clinical dimensions.

2) *Impact of Dynamic Intent Orchestration (DIO)*: To isolate the value of dynamic orchestration, we remove the DIO agent from the pipeline in Figure 6. The consequences are immediate. Performance degrades across all dimensions, with Clinical Planning suffering the most severe decline (*i.e.*, approximately 10%). This sharp drop validates our recursive priority grouping as a cornerstone of the architecture. In healthcare, the sequence of interventions is not arbitrary; it is a safety-critical factor. Without the DIO agent, the system fails to capture the complex interdependencies between tasks, potentially prioritizing treatments before necessary risk assessments. Furthermore, Response Quality decreases by 2.4%. This underscores that orchestration is not merely a scheduling task. Correct prioritization fundamentally shapes the coherence of the final advice, ensuring that recommendations follow a logical, medically sound progression.

3) *Role of Hierarchical Progressive Decision-making (HPD)*: From Figure 6, we further investigate the contribution of the HPD agent. Interestingly, removing this module has a negligible impact on Intent Comprehension and Clinical Planning. The system retains its ability to identify and schedule tasks. However, Response Quality plummets by a notable 9.09%. The specific degradation reveals the HPD agent's critical function:

TABLE XIV

ABLATION STUDY FOR ISOLATING THE CONTRIBUTION OF MIRA BEYOND THE BACKBONE LLM. ALL VARIANTS USE GPT-4.5 AS THE BACKBONE. “TIR” MEANS THE TWO-STAGE INTENTION RECOGNITION. “SM” MEANS THE SEMANTIC MATCHING.

| Variant        | Intent Comp. | Clinical Plan. | Response Qual. | Average |
|----------------|--------------|----------------|----------------|---------|
| GPT-4.5 Direct | 4.02         | 3.88           | 3.83           | 3.91    |
| GPT-4.5 + CoT  | 4.39         | 4.15           | 4.12           | 4.22    |
| MIRA w/o TIR   | 4.24         | 4.05           | 4.37           | 4.22    |
| MIRA w/o SM    | 4.41         | 4.18           | 4.45           | 4.35    |
| MIRA w/o DIO   | 4.71         | 3.91           | 4.51           | 4.38    |
| MIRA w/o HPD   | 4.75         | 4.33           | 4.20           | 4.43    |
| MIRA           | 4.76         | 4.32           | 4.62           | 4.57    |

context grounding. Without hierarchical evidence accumulation, the system essentially operates with amnesia between steps. It fails to carry forward historical medical references or prior test results into subsequent reasoning phases. Consequently, while the plan remains structured, the content becomes generic and lacks the precise, evidence-based reasoning required for high-quality medical consultation.

4) *Contribution of Framework Design Beyond Backbone LLM*: To investigate whether the improvements of MIRA mainly come from the proposed framework rather than the underlying LLM capability, we further compare MIRA with two LLM-only variants using the same GPT-4.5 backbone. Specifically, GPT-4.5 Direct directly generates medical advice from the patient query, while GPT-4.5+CoT introduces step-by-step reasoning without the proposed intent recognition and orchestration modules. As shown in Table XIV, MIRA consistently outperforms both LLM-only variants. This indicates that simply employing a powerful LLM or enhancing reasoning through prompting is insufficient for complex multi-intent medical consultations. The performance gains mainly come from the structured integration of intent decomposition, clinical dependency-aware orchestration, and hierarchical evidence accumulation.

#### E. Generalization Analysis

To evaluate the generalization ability of MIRA beyond the original evaluation datasets, we conduct additional zero-shot experiments on several external scenarios. No additional fine-tuning is performed, and the intent taxonomy, dependency structure, and agent framework remain unchanged. The evaluated settings include external medical dialogue, consumer health QA, mental health support conversations, and additional real-record holdout cases. As summarized in Table XIII, MIRA consistently outperforms the strongest multi-agent baseline across different domains. These results indicate that the advantage of MIRA is not limited to a specific dataset distribution. Instead, the proposed intent decomposition and dependency-aware orchestration provide a general framework for handling complex healthcare conversations with multiple intertwined needs.

#### F. Human Evaluation Reliability

To assess the reliability of physician evaluation, we further calculate inter-annotator agreement before majority voting in

TABLE XV

RELIABILITY ANALYSIS OF THE EVALUATION PROTOCOL.

| Evaluation item               | Metric                             | Value |
|-------------------------------|------------------------------------|-------|
| Physician pairwise evaluation | Number of response pairs           | 1,200 |
| Physician pairwise evaluation | Number of physicians per pair      | 3     |
| Physician pairwise evaluation | Average pairwise agreement         | 84.6% |
| Physician pairwise evaluation | Fleiss' $\kappa$                   | 0.73  |
| Physician pairwise evaluation | Krippendorff's $\alpha$            | 0.71  |
| LLM-human calibration         | Number of scored responses         | 300   |
| LLM-human calibration         | Spearman $\rho$                    | 0.79  |
| LLM-human calibration         | Kendall $\tau$                     | 0.63  |
| LLM-human calibration         | Quadratic weighted $\kappa$        | 0.71  |
| LLM-judge consistency         | Intraclass correlation coefficient | 0.81  |
| LLM-judge consistency         | Krippendorff's $\alpha$            | 0.78  |

Table XV. The average pairwise agreement among the three physicians is 84.6%, with Fleiss'  $\kappa$  of 0.73 and Krippendorff's  $\alpha$  of 0.71, indicating substantial agreement among expert annotators. These results demonstrate that the physician evaluation protocol provides reliable and consistent clinical judgments.

Furthermore, we conduct a calibration analysis between LLM-based evaluation and physician scores on 300 randomly sampled responses. The LLM-judge ensemble achieves a Spearman correlation of  $\rho = 0.79$ , Kendall correlation of  $\tau = 0.63$ , and quadratic weighted  $\kappa$  of 0.71 with physician ratings. These results indicate that the automated evaluation protocol is reasonably aligned with expert clinical judgment and can serve as a complementary evaluation approach when combined with human assessment.

## V. CONCLUSION

In this paper, we present MIRA, an LLM-based multi-intention recognition and planning framework for complex medical and mental health-related queries. Finally, we integrate these components into a multi-agent architecture coordinated by a Hierarchical Progressive Decision-making (HPD) agent, which accumulates evidence across stages and synthesizes coherent, trustworthy responses. Extensive experiments on several medical dialogue datasets show that MIRA improves intent comprehension, planning quality, and overall response effectiveness compared to other agent baselines.

Despite the promising results, MIRA has several limitations that motivate future research. First, the current evaluation mainly focuses on benchmark medical dialogue datasets and generated patient queries, while broader validation in multi-center and real-world healthcare environments remains necessary. Second, although the proposed intent taxonomy and dependency structure provide clinically meaningful guidance, extending MIRA to new healthcare domains may require additional domain-specific intents and expert knowledge. Third, the current framework primarily models affective needs through textual intention understanding. Future work will investigate multimodal affect recognition, longitudinal emotional state modeling, and personalized interaction strategies to develop more comprehensive affective healthcare assistants.

## REFERENCES

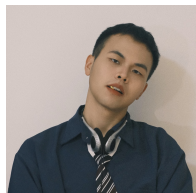
- [1] X. Wang, M. Ge, L. Li, D. Jin, K. He, and E. Cambria, "Global entity relationship enhancement network for multimodal sarcasm detection," *IEEE Trans. Affect. Comput.*, 2025.
- [2] M. Tlachac, M. V. Heinz, A. C. Bryan, A. LaPreay, G. L. Dimas, T. Zhao, N. C. Jacobson, and S. S. Ogden, "Datasets of smartphone modalities for depression assessment: A scoping review," *IEEE Trans. Affect. Comput.*, 2025.
- [3] Z. Lin, Y. Wang, Y. Zhou, F. Du, and Y. Yang, "Mlm-eoe: Automatic depression detection via sentimental annotation and multi-expert ensemble," *IEEE Trans. Affect. Comput.*, 2025.
- [4] H. Huang, T. Gong, K. He, J. Wu, E. Cambria, and M. Feng, "Robust multimodal sentiment analysis via double information bottleneck," *Inf. Fusion*, p. 103964, 2025.
- [5] Z. Gao, L. Huang, Q. Lin, B. Pu, K. He, M. Feng, and K. Li, "Domain-continual learning for multi-center anatomical detection via prompt-enhanced and densely-fused medsam," *Inf. Fusion*, p. 103614, 2025.
- [6] R. Chionga, G. S. Budhib, and E. Cambria, "Detecting signs of depression using social media texts through an ensemble of ensemble classifiers," *IEEE Trans. Affect. Comput.*, 2025.
- [7] J. Wu, X. Mei, R. Mao, K. He, and E. Cambria, "Takecare: A temporal-hierarchical framework with knowledge fusion for personalized clinical predictive modeling," *Inf. Fusion*, p. 103620, 2025.
- [8] Y. Zhang, C. Zou, Z. Lian, P. Tiwari, and J. Qin, "Sarcasmbench: Towards evaluating large language models on sarcasm understanding," *IEEE Trans. Affect. Comput.*, 2025.
- [9] A. N. Tak, J. Gratch, and K. R. Scherer, "Aware yet biased: Investigating emotional reasoning and appraisal bias in large language models," *IEEE Trans. Affect. Comput.*, 2025.
- [10] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-llm: Leveraging large language models for mental health prediction via online text data," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 8, no. 1, pp. 1–32, 2024.
- [11] S. Goyal, E. Rastogi, S. P. Rajagopal, D. Yuan, F. Zhao, J. Chintagunta, G. Naik, and J. Ward, "Healai: A healthcare llm for effective medical documentation," in *Proc. 17th ACM Int. Conf. Web Search Data Mining*, 2024, pp. 1167–1168.
- [12] T. Mirzaei, L. Amini, and P. Esmailzadeh, "Clinician voices on ethics of llm integration in healthcare: a thematic analysis of ethical concerns and implications," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 250, 2024.
- [13] H. He, D. Chen, A. Balakrishnan, and P. Liang, "Decoupling strategy and generation in negotiation dialogues," in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Brussels, Belgium: Assoc. Comput. Linguistics, 2018, pp. 2333–2343.
- [14] X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu, "Persuasion for good: Towards a personalized persuasive dialogue system for social good," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5635–5649.
- [15] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, H. Park *et al.*, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 79410–79452, 2024.
- [16] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, and Z. Wei, "Disc-medllm: Bridging general large language models and real-world medical consultation," *arXiv preprint arXiv:2308.14346*, 2023.
- [17] D. Yang, J. Wei, D. Xiao, S. Wang, T. Wu, G. Li, M. Li, S. Wang, J. Chen, Y. Jiang, Q. Xu, K. Li, P. Zhai, and L. Zhang, "Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 138 632–138 662.
- [18] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 17, 2024, pp. 19 368–19 376.
- [19] Y. Kang, Y. Chang, J. Fu, Y. Wang, H. Wang, and W. Zhang, "Cmlm-zhongjing: Large language model is good story listener," <https://github.com/pariskang/CMLM-ZhongJing>, 2023.
- [20] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami *et al.*, "Meditron-70b: Scaling medical pretraining for large language models," *arXiv preprint arXiv:2311.16079*, 2023.
- [21] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.
- [22] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," *arXiv preprint arXiv:2304.01097*, 2023.
- [23] Z. Fan, L. Wei, J. Tang, W. Chen, S. Wang, Z. Wei, and F. Huang, "AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*. Abu Dhabi, UAE: Assoc. Comput. Linguistics, 2025, pp. 10 183–10 213.
- [24] B. Li, T. Yan, Y. Pan, J. Luo, R. Ji, J. Ding, Z. Xu, S. Liu, H. Dong, Z. Lin, and Y. Wang, "Mmedagent: Learning to use medical tools with multi-modal agent," in *Findings Assoc. Comput. Linguistics: EMNLP 2024*. Miami, FL, USA: Assoc. Comput. Linguistics, 2024, pp. 8745–8760.
- [25] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, and M. Gerstein, "Medagents: Large language models as collaborators for zero-shot medical reasoning," in *Findings Assoc. Comput. Linguistics: ACL 2024*. Bangkok, Thailand: Assoc. Comput. Linguistics, 2024, pp. 599–621.
- [26] M. Tsulukidze, S. Grande, R. Thompson, K. Rudd, and G. Elwyn, "Patients covertly recording clinical encounters: Threat or opportunity? a qualitative analysis of online texts," *PLOS ONE*, vol. 10, p. e0125824, 05 2015.
- [27] OpenAI, "Introducing chatgpt," 2022, <https://openai.com/blog/chatgpt>.
- [28] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [29] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.
- [30] M. Xu, "Medicalgpt: Training medical gpt model," <https://github.com/shibing624/MedicalGPT>, 2023.
- [31] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, G. Chen, J. Li, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, and H. Li, "Huatuogpt, towards taming language model to be a doctor," in *Findings Assoc. Comput. Linguistics: EMNLP 2023*. Singapore: Assoc. Comput. Linguistics, 2023, pp. 10 859–10 885.
- [32] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong, J. Li, X. Wan, H. Li, and B. Wang, "Huatuogpt-ii, one-stage training for medical adaption of llms," in *Proc. Conf. Lang. Model. (COLM)*, 2024.
- [33] W. Hua, R. Fan, T. He, and M. Dong, "Picd-instruct: A generative instruction learning framework for few-shot multi-intent spoken language understanding," in *Findings Assoc. Comput. Linguistics: EMNLP 2025*, 2025, pp. 4942–4956.
- [34] M. Tageldien, B. Selim, and L. Sboui, "Large language models in intent-based networking: a comprehensive survey across the intent lifecycle," in *2025 Int. Telecommun. Conf. (ITC-Egypt)*. IEEE, 2025, pp. 810–817.
- [35] B. Zhang, Z. Tu, S. Hang, D. Chu, and X. Xu, "Conco-ernie: Complex user intent detect model for smart healthcare cognitive bot," *ACM Trans. Internet Technol.*, vol. 23, no. 1, pp. 1–24, 2023.
- [36] Y. Zhang and R. Y. Lau, "Uncertainty detection: A multi-view decision boundary approach against healthcare unknown intents," *Appl. Sci.*, vol. 15, no. 13, p. 7114, 2025.
- [37] T. Pham, C. Tran, and D. Q. Nguyen, "Misca: A joint model for multiple intent detection and slot filling with intent-slot co-attention," in *Findings Assoc. Comput. Linguistics: EMNLP 2023*, 2023, pp. 12 641–12 650.
- [38] T. Pham and D. Q. Nguyen, "Jpis: A joint model for profile-based intent detection and slot filling with slot-to-intent attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, 2024, pp. 10 446–10 450.
- [39] D. Yang, J. Wei, M. Li, J. Liu, L. Liu, M. Hu, J. He, Y. Ju, W. Zhou, Y. Liu *et al.*, "Medaide: information fusion and anatomy of medical intents via llm-based agent collaboration," *Inf. Fusion*, p. 103743, 2025.
- [40] J. Chen, D. Yang, Y. Jiang, M. Li, J. Wei, X. Hou, and L. Zhang, "Efficiency in focus: Layernorm as a catalyst for fine-tuning medical visual language models," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 3122–3130.
- [41] M. Geng, S. Wang, D. Dong, H. Wang, G. Li, Z. Jin, X. Mao, and X. Liao, "Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning," in *Proc. 46th IEEE/ACM Int. Conf. Softw. Eng.*, 2024, pp. 1–13.
- [42] J. Wang, F. Mo, W. Ma, P. Sun, M. Zhang, and J.-Y. Nie, "A user-centric multi-intent benchmark for evaluating large language models," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Miami, FL, USA: Assoc. Comput. Linguistics, 2024, pp. 3588–3612.

- [43] S. Yin, P. Huang, Y. Xu, H. Huang, and J. Chen, "Do large language model understand multi-intent spoken language?" *arXiv preprint arXiv:2403.04481*, 2024.
- [44] A. M. Nazar, B. Norman, H. Northway, A. Toutoungi, E. Zatkalik, G. Carlson, E. Sabado, H. Shawa, and M. Y. Selim, "Providers of relief in distress: Rag-based llms as situation and intent-aware assistants," *Front. Artif. Intell.*, vol. 9, p. 1712596, 2026.
- [45] R. Lenz and M. Reichert, "A practical approach to process support in health information systems," *J. Amer. Med. Inform. Assoc.*, vol. 14, no. 4, pp. 494–501, 2007.
- [46] D. S. Nau, T.-C. Au, O. Ilghami, U. Kuter, J. W. Murdock, D. Wu, and F. Yaman, "SHOP2: An HTN planning system," *J. Artif. Intell. Res.*, vol. 20, pp. 379–404, 2003.
- [47] D. F. Sittig, A. Wright, J. A. Osherooff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates, "Grand challenges in clinical decision support," *J. Biomed. Inform.*, vol. 69, pp. 340–351, 2015.
- [48] M. Peleg, S. Tu, J. Bury, P. Ciccarese, J. Fox, R. A. Greenes, R. Hall, P. D. Johnson, N. Jones, A. Kumar *et al.*, "Computer-interpretable clinical guidelines: a methodological review," *J. Biomed. Inform.*, vol. 46, no. 4, pp. 744–763, 2013.
- [49] R. T. Sutton, D. Pincok, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digit. Med.*, vol. 3, no. 1, p. 17, 2020.
- [50] J. Wu, X. Wu, Y. Zheng, and J. Yang, "Clinical pathway-aware large language models for reliable and transparent medical dialogue," *J. Biomed. Inform.*, p. 104942, 2025.
- [51] H. Zhou, F. Liu, J. Wu, W. Zhang, G. Huang, L. Clifton, D. Eyre, H. Luo, F. Liu, K. Branson *et al.*, "A collaborative large language model for drug analysis," *Nat. Biomed. Eng.*, pp. 1–12, 2025.
- [52] T. Yue, R. Mao, X. Shi, and E. Cambria, "Interarm: Interpretable affective reasoning model for multimodal sarcasm detection," *IEEE Trans. Affect. Comput.*, 2026.
- [53] G. Tu, R. Jing, X. Luo, E. Cambria, W. Li, and R. Xu, "Is multimodal conversational emotion recognition satisfactory? exploring the gaps in performance, generalization, and confidence," *Pattern Recognit.*, p. 113087, 2026.
- [54] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24 824–24 837, 2022.
- [55] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [56] S. Yao, D. Yu, J. Zhao, I. Shafraan, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 11 809–11 822, 2023.
- [57] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nycz, and T. Hoefler, "Graph of thoughts: Solving elaborate problems with large language models," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 16, p. 17682–17690, Mar. 2024. [Online]. Available: <http://dx.doi.org/10.1609/aaai.v38i16.29720>
- [58] OpenAI, "GPT-4.5 technical report," Available at <https://openai.com/research/gpt-4-5>, 2025.
- [59] J. S. Ancker, A. Edwards, S. Nosal, D. Hauser, E. Mauer, R. Kaushal, and W. the HITEC Investigators, "Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system," *BMC Med. Inform. Decis. Mak.*, vol. 17, pp. 1–9, 2017.
- [60] P. K. Wan, A. Satyabadi, L. Huang, H. Holtskog, and M. Nowostawski, "Reducing alert fatigue by sharing low-level alerts with patients and enhancing collaborative decision making using blockchain technology: scoping review and proposed framework (medalert)," *J. Med. Internet Res.*, vol. 22, no. 10, p. e22013, 2020.
- [61] J. P. Kassirer, "Diagnostic reasoning," *Ann. Intern. Med.*, vol. 110, no. 11, pp. 893–900, 1989.
- [62] R. L. Trowbridge, J. J. Rencic, and S. J. Durning, *Teaching Clinical Reasoning*. Philadelphia, PA: American College of Physicians, 2015.
- [63] X. Wang, J. Li, S. Chen, Y. Zhu, X. Wu, Z. Zhang, X. Xu, J. Chen, J. Fu, X. Wan, A. Gao, and B. Wang, "Huatuo-26m, a large-scale chinese medical qa dataset," in *Findings Assoc. Comput. Linguistics: NAACL 2025*. Albuquerque, NM, USA: Assoc. Comput. Linguistics, 2025, pp. 3828–3848.
- [64] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado,

Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

- [65] Google DeepMind, "Gemini 2.0: Our new AI model," Available at <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>, 2024.

- [66] Anthropic, "Introducing Claude Sonnet 4.5," Available at <https://www.anthropic.com/claude/sonnet>, 2025.



**Dingkang Yang** is an Assistant Professor with the College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China. He received the B.E. degree (with honors) in Communication Engineering from the joint training program of Yunnan University and the Chinese People's Armed Police (PAP), Kunming, China, in 2020, and the Ph.D. degree in Computer Science from Fudan University, Shanghai, China, in 2025. His research interests include multimodal learning, generative AI, and embodied AI. Dr. Yang has published multiple

papers as the first author at the reputable journals and top international conferences, such as IEEE TPAMI, TCSVT, NeurIPS, CVPR, ICCV, ECCV, and AAAI. He previously worked as a Principal Researcher at ByteDance, where his team developed a series of renowned multimodal foundational models, including SAIL-VL and SAIL-Embedding. He currently serves as CTO of Fysics AI, an artificial intelligence company dedicated to developing omnimodal interaction and physical real-world perception. He is also a postdoctoral fellow at MMLab of The Chinese University of Hong Kong.



**Jinjie Wei** is a Master of Computer Science candidate at Fudan University, advised by Prof. Lihua Zhang, and has conducted research at Shanghai AI Laboratory, Microsoft Research Asia, and UC Santa Cruz. His research interests focus on Large Language Models (LLMs) and Multimodal Learning, spanning systematic domain-specific model construction (e.g., PediatricsGPT), LLM-based multi-agent collaboration frameworks, hallucination detection and mitigation, and reinforcement learning in Vision-Language Models (VLMs). He has published papers

in leading AI conferences and journals such as NeurIPS, AAAI, CVPR, ACM MM, and MICCAI. His work has been recognized with distinctions including AAAI Orals and NeurIPS Poster presentations.



**Ming Hu** is a Ph.D. candidate at Monash University and a research intern at Shanghai AI Laboratory. His research focuses on multimodal foundation models and multimodal large language models (MLLMs), spanning vision-language pretraining, multimodal reasoning and reliability (e.g., hallucination mitigation and evaluation), video-language representation learning and benchmarking, and diffusion-based multimodal modeling, including Diffusion LLMs for unified multimodal understanding and generation. He has published over 30 papers in leading conferences

and journals such as CVPR, NeurIPS, ICCV, ECCV, MICCAI, ACL, AAAI, and Nature Medicine. His work has been recognized as a NeurIPS Spotlight, CVPR Oral, ICCV Highlight, and MICCAI Oral. He has co-authored several large-scale multimodal benchmarks and datasets for real-world multimodal learning and healthcare AI, serves as a student reviewer for major AI venues, and has organized challenges at APTOS 2025 and MICCAI 2025.



**Ji Yao Liu** is a Ph.D. student at the Institute of Science and Technology for Brain-inspired Intelligence, Fudan University. His research focuses on low-level vision, AI for healthcare and science, multimodal large language models, and agents. He was shortlisted for the 2025 MICCAI Best Paper and Young Scientist Awards and has published multiple papers at top-tier international conferences such as MICCAI and NeurIPS.

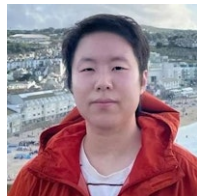


**Junjun He** is a young research scientist at Shanghai AI Laboratory, focusing on multimodal understanding, multimodal generation, unified multimodal understanding and generation, multi-agent system towards scientific discovery, and their applications in healthcare. His research has been cited over 6,000 times on Google Scholar, with an H-index of 34. He has won more than ten awards in international challenges, including six championships. He currently leads the General Medical AI (GMAI) team at Shanghai AI Laboratory. The GMAI team has developed and open-sourced multiple large-scale medical benchmarks and state-of-the-art models in the field of medical AI. Representative achievements include the large-scale 3D imaging pretraining model STU-Net, the large-scale medical segmentation foundation models SAM-Med2D and SAM-Med3D, the most comprehensive medical multimodal evaluation benchmarks OmniMedVQA and GMAI-MMBench, the general-purpose medical multimodal model GMAI-VL, the ultra-high-resolution pathology WSI multimodal model SlideChat, and large-scale surgical multimodal models OphClip and Ophora. His recent open-source project, Project Imaging-X, a large-scale survey and platform for medical imaging data, has attracted widespread attention. His paper was shortlisted for the 2025 MICCAI Best Paper and Young Scientist Awards.



Awards.

**Lihao Liu** is currently a Research Scientist at Amazon. He received his Ph.D. degree from the University of Cambridge in 2024, his M.Phil. degree from The Chinese University of Hong Kong, and his B.Eng. degree from Chongqing University. His research interests include LLM-based agents and healthcare AI. He focuses on leveraging agentic AI to advance the field of medical imaging and healthcare. He has previously interned at Shanghai AI Laboratory, Microsoft, and UCLA. His paper was shortlisted for the MICCAI 2025 Best Paper and Young Scientist



**Wei Zhou** is an Assistant Professor at Cardiff University, United Kingdom. His research interests include multimedia computing, perceptual image and video processing, computational vision, human-centric AI, and machine learning. He has published over 80 papers in leading journals and conferences including IEEE TIP, IEEE TCSVT, IEEE TMM, IEEE TMI, CVPR, ECCV, ACM MM, and MICCAI. He has industrial and research experience at the University of Waterloo (Canada), National Institute of Informatics (Japan), University of Science and Technology of China, Intel, Microsoft Research, and Alibaba Group. Prof. Zhou currently serves as Associate Editor for nine journals including IEEE TNNLS, ACM TOMM, Pattern Recognition, and Neurocomputing, and as Area Editor for Signal Processing: Image Communication. He chairs the Election Committee of the IEEE UK & Ireland Signal Processing Society Chapter and has served as Area Chair for major conferences including ACM MM, IEEE ICME, IEEE ICIP, and BMVC. His honors include the IEEE CASS VSPC Rising Star Honorable Mention, ACM SIGMM China Outstanding Doctoral Dissertation Award, CVPR CLIC Challenge Winner Award, and inclusion in Stanford University's World Top 2% Scientists list. He actively contributes to the research community by organizing special issues, tutorials, challenges, and workshops at venues such as IEEE JBHI, Pattern Recognition, and ICCV.



**Zhaoyu Chen** received his B.E. degree from Shandong University in 2020 and his Ph.D. degree from Fudan University in 2025. He is currently a Research Scientist at TikTok, with his primary research interests focusing on the security and privacy of Large Language Models and Multimodal Large Language Models (LLM/MLLM), including adversarial attack/defense and differential privacy training.segmentation.



**Yang Liu** is an Assistant Professor with the Tongji University, Shanghai, China. He received the B.S. degree (with honors) in Automation from Nanjing University, China, in 2020, and the Ph.D. degree in Computer Science from Fudan University, China, in 2025. From 2023 to 2025, he was a visiting scholar at the University of Toronto, Canada, under a joint Ph.D. program. His research focuses on embodied perception with applications in industry and intelligent transportation systems. Dr. Liu has served as Guest Editor for IEEE TCSS, Area Chair for BMVC 2025 and IEEE ICIP 2025, Workshop/Special Session Co-Chair for IEEE ICASSP, IEEE ICIP, and IEEE WF-IoT, and TPC member for top-tier conferences including CVPR, ICCV, ICLR, NeurIPS, AAAI, and ACM MM. He is an active reviewer for ACM CSUR, PR, INFU, and multiple IEEE Transactions including TPAMI, TIP, TH, TNNLS, TCSVT, TMM, and TIFS.



**Mingcheng Li** received the B.S. degree in Computer Science and Technology from Hefei University of Technology, Hefei, China, in 2021. He is currently pursuing the Ph.D. degree at the Academy for Engineering and Technology, Fudan University, Shanghai, China. His research interests include multimodal perception, large multimodal models, and causal inference.



**Lihua Zhang** received the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2000. He is currently a Professor at the Academy for Engineering and Technology, Fudan University, Shanghai, China. In recent years, he has participated in a number of national science and technology research and development projects as a project leader and sub-project. His current research interests are in artificial intelligence and its applications, including machine intuition, computer vision and intelligent perception, virtual reality and digital twinning, intelligent robotics and unmanned systems, intelligent computing and intelligent chips, intelligent healthcare, intelligent connected vehicles, etc.