

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

CLINICAL PROCESS-AWARE MEDICAL DIALOGUE SYSTEMS

KAISHUAI XU

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Clinical Process-Aware Medical Dialogue Systems

Kaishuai Xu

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
August 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Kaishuai Xu

Abstract

Dialogue systems, driven by advances in Large Language Models (LLMs), enable human-like interactions across a wide range of user-facing tasks. In recent years, Medical Dialogue Systems (MDS) have gained increasing attention as demand for telemedicine has grown since the onset of the COVID-19 pandemic. These systems have expanded access to diagnostic and prescribing services via online consultations, helping to alleviate pressure on healthcare professionals. Although significant progress has been made, a substantial gap remains between current MDS and the standard of doctor-patient communication expected in clinical practice. Most existing MDS generally struggle to emulate aspects of the clinical process—such as structured clinical communication, comprehensive differential diagnosis, and prototypical clinical reasoning processes—in ways that align with clinical practice, so their behavior often diverges from clinical practice and falls short of clinician-like performance in terms of clinical accuracy, reliability, and safety. In the medical domain, a reliable dialogue system should model the clinical process and incorporate its core aspects into system design, which in turn improves safety and guideline concordance, and helps build user trust.

In this thesis, we present a clinical process-aware medical dialogue system that emulates clinical practice, aiming to deliver trustworthy, clinician-like interactions. Our goal is to advance clinical process-aware dialogue systems by proposing effective, novel approaches to this widely needed but underexplored area. Specifically, we investigate

three primary research questions (RQs): (1) How to model the critical information flow (e.g., medical entities and dialogue acts) in the medical dialogue and incorporate key entities and acts into response generation? (2) How to develop a comprehensive differential diagnosis and leverage it to improve the reliability and interpretability of system responses? (3) How to emulate clinicians’ internal reasoning processes and improve a system’s reasoning capabilities with external medical knowledge? To tackle these challenges, we advance the design of MDS from surface-level similarity to internal alignment with clinical practice. This thesis makes several contributions addressing these fundamental questions, presented in three parts.

In the first part, we address RQ (1) and identify that extracting critical information and capturing transitions in the information flow are essential for building MDS. We propose the Dual Flow Enhanced Medical (DFMed) dialogue generation framework. At each dialogue turn, we extract medical entities and doctor dialogue acts from the dialogue history and model their transitions via an entity-centric graph flow and a sequential act flow, respectively. We further introduce an interweaving component to strengthen interactions between the two flows. Experiments show that explicitly modeling the two flows helps predict key information for the next turn, thereby improving response accuracy.

The second part of the thesis focuses on RQ (2) and presents a comprehensive approach to differential diagnosis. Unlike prior approaches that rely on a generative model’s implicit diagnostic reasoning, we build a separate diagnosis module to improve diagnosis reliability and transparency. We propose IADDx, a framework that first generates a differential diagnosis and then uses the resulting candidate diseases to guide response generation. Inspired by research on diagnostic reasoning, we design a two-stage intuitive-then-analytical differential diagnosis method. In the intuitive stage, the system retrieves relevant patient cases and disease documents to construct a preliminary disease list; in the analytical stage, a multi-disease classifier refines this list. The refined candidate diseases are then used to enhance subsequent response

generation. Our experiments with IADDx demonstrate that a detailed diagnostic process enhances the response accuracy and interpretability.

In the third part, we investigate RQ (3) and develop methods that enable clinically aligned and knowledge-augmented reasoning processes. For clinically aligned reasoning, we propose EMULATION, a framework that emulates clinicians’ internal thought processes to generate responses aligned with clinician preferences. Three modules form our framework: an abductive reasoning module and a deductive reasoning module that jointly serve as a diagnostic-analysis component to construct the differential diagnosis, and a thought alignment module that reprioritizes candidate diseases and produces CoT-style thought processes and final responses. To learn the clinician preference, we also collect a synthetic thought process dataset with the help of an LLM. For knowledge-augmented reasoning, we adopt Thought-RAG, which outlines a thought process to identify implicit knowledge needs and retrieves relevant knowledge accordingly for generation. We propose a joint learning framework, RAR², that simultaneously improves thought process generation and knowledge-augmented generation. Specifically, we first curate a training dataset comprising preference pairs across different generation objectives, then apply Direct Preference Optimization (DPO) to enhance reasoning both with and without external knowledge. Extensive experiments on several biomedical question answering datasets demonstrate RAR²’s superiority over existing RAG baselines. Besides, further test-time scaling analysis on thought process generation validates the scalability.

In summary, this thesis advances clinical process-aware medical dialogue systems by integrating key aspects of clinical practice. We present models, system designs, and datasets that move systems from surface-level mimicry toward internal alignment with clinical practice. Experiments across medical dialogue and question-answering benchmarks show consistent gains over strong baselines, validating the effectiveness and scalability of our approach and pointing toward trustworthy, clinician-like interactions throughout the clinical encounter.

Publications Arising from the Thesis

1. **Kaishuai Xu**, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li, “Medical Dialogue Generation via Dual Flow Modeling”, in *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6771–6784, 2023.
2. **Kaishuai Xu**, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li, “Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment”, in *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6796–6814, 2024.
3. **Kaishuai Xu**, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li, “Medical Dialogue Generation via Intuitive-then-Analytical Differential Diagnosis”, arXiv preprint, arXiv:2401.06541, 2024.
4. **Kaishuai Xu**, Wenjun Hou, Yi Cheng, and Wenjie Li, “RAR²: Retrieval-Augmented Medical Reasoning via Thought-Driven Retrieval”, *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.
5. **Kaishuai Xu**, Tiezheng Yu, Wenjun Hou, Yi Cheng, Chak Tou Leong, Liangyou Li, Xin Jiang, Lifeng Shang, Qun Liu, and Wenjie Li, “Subtle Errors in Reasoning: Preference Learning via Error-injected Self-editing”, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.
6. **Kaishuai Xu**, Tiezheng Yu, Wenjun Hou, Yi Cheng, Liangyou Li, Xin Jiang, Lifeng Shang, Qun Liu, and Wenjie Li, “Learning to Align Multi-Faceted Eval-

- uation: A Unified and Robust Framework”, in *Findings of the Association for Computational Linguistics: ACL 2025*.
7. Yi Cheng, Wenge Liu, **Kaishuai Xu**, Wenjun Hou, Yi Ouyang, Chak Tou Leong, Xian Wu, and Yefeng Zheng, “AutoPal: Autonomous Adaptation to Users for Personal AI Companionship”, arXiv preprint, arXiv:2406.13960, 2024.
 8. Wenjun Hou, Yi Cheng*, **Kaishuai Xu***, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu, “RADAR: Enhancing Radiology Report Generation with Supplementary Knowledge Injection” (*: Equal Contribution), in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.
 9. Wenjun Hou, Yi Cheng*, **Kaishuai Xu***, Yan Hu, Wenjie Li, and Jiang Liu, “ICON: Improving Inter-Report Consistency in Radiology Report Generation via Lesion-aware Mixup Augmentation” (*: Equal Contribution), in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.
 10. Wenjun Hou, **Kaishuai Xu***, Yi Cheng*, Wenjie Li, and Jiang Liu, “ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning” (*: Equal Contribution), in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
 11. Wenjun Hou, Yi Cheng*, **Kaishuai Xu***, Wenjie Li, and Jiang Liu, “RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning” (*: Equal Contribution), in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Acknowledgments

The PhD journey is long and unforgettable for everyone, but especially so for me. If I could return to the moment I chose to pursue a PhD—and to switch from management science to NLP—I would probably still hesitate, and I certainly would not have imagined how dramatically the field would evolve. For someone who knew little about NLP at the start, moving from RNNs to Transformers and then to LLMs felt like leaping from the Neolithic era into the information age. I would thanks all people who have helped me along the way.

First and foremost, I would like to thank Professor Wenjie Li (Maggie). In 2020, when I was most confused and uncertain, she offered me an opportunity to learn NLP and to embark on a PhD. I am profoundly thankful for her guidance and support throughout my studies, especially for her encouragement to become an independent and self-directed researcher.

I am also grateful to my close collaborators, Wenjun Hou, Yi Cheng, Jian Wang, and Chak Tou Leong. In every discussion, they offered candid and constructive feedback that challenged my assumptions and helped me continually reflect on and improve my work.

My heartfelt thanks go to my girlfriend, Xiaojun Cai, for her companionship encouraging and comforting me when research was difficult, and celebrating with me when things went well.

Finally, I thank my parents. Whatever decisions I made, they always gave me their fullest support. Even when I chose a path that was not the easiest, they believed I could persevere and that things would work out.

To all of you: thank you.

Table of Contents

Abstract	i
Publications Arising from the Thesis	iv
Acknowledgments	vi
List of Figures	xiv
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.2 Research Problems	4
1.3 Research Overview and Contributions	6
1.4 Thesis Structure	10
2 Literature Review	12
2.1 Medical Large Language Models	12
2.1.1 Model Development	13

2.1.2	Comprehensive Evaluation	15
2.2	Medical Dialogue System	17
2.2.1	Medical Dialogue Datasets	17
2.2.2	Automatic Diagnosis Systems	18
2.2.3	Medical Dialogue Generation	20
2.3	LLM-based Medical Reasoning	22
2.3.1	Benchmarks and Evaluation	22
2.3.2	Knowledge-Augmented Reasoning	25
2.3.3	Agentic Reasoning	26
3	Modeling Medical Information Flow	29
3.1	Introduction	30
3.2	Preliminary	32
3.3	Method	33
3.3.1	Dual Flow Modeling	33
3.3.2	Flow-guided Response Generation	39
3.4	Experiments	41
3.4.1	Dataset	41
3.4.2	Baseline models	41
3.4.3	Evaluation Metrics	42
3.4.4	Implementation Details	42
3.5	Results and Analysis	43

3.5.1	Details of Dialogue Acts	43
3.5.2	Automatic Evaluation	47
3.5.3	Human Evaluation	48
3.5.4	Analysis of Dual Flow Modeling	49
3.5.5	Case Study	49
3.6	Chapter Summary	50
4	Developing Accurate Differential Diagnoses	52
4.1	Introduction	53
4.2	Preliminary	55
4.2.1	Problem Formulation	55
4.2.2	SOAP-based Content Structuring	56
4.2.3	Disease Knowledge Graph	57
4.3	Method	57
4.3.1	Intuitive-then-Analytic Differential Diagnosis	58
4.3.2	Diagnosis-guided Response Generation	63
4.4	Experiments	65
4.4.1	Datasets	65
4.4.2	Baseline Models	66
4.4.3	Evaluation Metrics	66
4.4.4	Implementation Details	67
4.4.5	Evaluation of Dialogue Generation	69

4.4.6	Analysis of Intuitive-then-Analytic Differential Diagnosis . . .	70
4.4.7	Case Study	71
4.5	Chapter Summary	72
5	Aligning Diagnostic Reasoning	74
5.1	Introduction	75
5.2	Preliminary	78
5.2.1	Problem Formulation	78
5.2.2	Diagnostic Reasoning	78
5.2.3	Disease Annotation	79
5.3	Method	81
5.3.1	Abductive Reasoner	81
5.3.2	Deductive Reasoner	83
5.3.3	Thought Alignment	84
5.4	Thought Process Corpus	85
5.5	Experiments	87
5.5.1	Datasets	87
5.5.2	Baseline methods	88
5.5.3	Implementation Details	90
5.5.4	Automatic Evaluation	91
5.5.5	Human Evaluation	92
5.5.6	Analysis of Diagnostic Reasoning Process	93

5.5.7	Analysis of Disease Alignment	95
5.6	Case Studies	96
5.7	Chapter Summary	96
6	Enhancing Medical Reasoning with External Knowledge	99
6.1	Introduction	100
6.2	Preliminary	103
6.2.1	Problem Formulation	103
6.2.2	Medical Corpus	103
6.3	Reasoning Before Retrieval	103
6.4	Method	105
6.4.1	Construction of Mixed Preference Pairs	107
6.4.2	Joint Learning	108
6.4.3	Test-Time Scaling	109
6.5	Experiments	110
6.5.1	Evaluation Datasets	110
6.5.2	Baseline Methods	111
6.5.3	Implementation Details	112
6.5.4	Main Results	113
6.5.5	Ablation Studies	114
6.5.6	Impact of Test-Time Scaling	115
6.6	Chapter Summary	117

7	Conclusions and Future Work	118
7.1	Summary of Contributions	119
7.1.1	Dual Flow Enhanced Medical Dialogue Generation	119
7.1.2	Differential Diagnosis Guided Medical Dialogue System	119
7.1.3	Diagnostic Reasoning Aligned Medical Dialogue System	120
7.2	Safety Considerations	120
7.3	Future Work	122

List of Figures

3.1	An example of a medical dialogue. <i>Diagnosis</i> and <i>Prescription</i> are short for <i>Make a diagnosis</i> and <i>Prescribe medications</i> , respectively. . .	31
3.2	The overall framework of DFMed.	34
3.3	The architecture of Dual Flow Modeling. It includes the dialogue act flow modeling in red and the medical entity flow modeling in blue. The one-hop sub-graph nodes of entities in dialogue are marked in yellow.	34
3.4	Detailed structure for flow interweaving. Each turn’s graph embedding attends to previous act embeddings, and each turn’s sequence embedding attends to previous entity embeddings.	37
3.5	The detail of Response Generation.	40
4.1	An example of differential diagnosis in a medical dialogue, which contains intuitive and analytic reasoning.	54
4.2	The prompt for extracting segments of the SOAP Notes.	56

4.3	Left: The architecture of Differential Diagnosis, which includes the Intuitive Association stage and the Analytic Refinement stage. The DDx Analyzer in Stage 2 predicts the affected systems and organs and refines the Preliminary Diagnosis to generate a Differential Diagnosis. Right: The structure of Response Generation. The generation model is fine-tuned and instructed to output dialogue acts (i.e., “inquire about symptoms”) and the corresponding responses.	58
5.1	An example of a medical dialogue. IBS is the abbreviation for irritable bowel syndrome.	76
5.2	Illustration of our Framework EMULATION. The symbols of a snowflake or a flame represent the module mainly operating in a prompt-based way or having undergone fine-tuning, respectively.	80
5.3	Human evaluation results of baseline methods. Red is for Win , yellow for Tie , and green for Lose	92
5.4	The top-5 diagnosis results with or without disease priority alignment.	94
5.5	The top-5 diagnosis results of EMULATION at different conversation turns.	95
5.6	Case studies on the MedDG dataset.	97
5.7	Case studies on the KaMed dataset.	98
6.1	Thought-RAG for medical reasoning. Our method jointly optimizes the generation of thought processes and the subsequent retrieval-augmented answer generation.	101
6.2	The performance comparison between Question-RAG and Thought-RAG on several medical question answering datasets.	104

6.3	Construction of mixed preference pairs. (1) We sample several thought processes and append each to the question to form a query for retrieval. Each thought is then annotated based on whether it leads to a correct answer in subsequent RAG. (2) We sample several answers and annotate each one according to its correctness.	106
6.4	Frameworks of test-time scaling. Parallel scaling concatenates multiple thought processes to retrieve documents, while iterative scaling uses the retrieved documents in sequence to enhance thought generation. .	109
6.5	The performance comparison with a different number of retrieved documents.	115
6.6	Accuracies with the Parallel Scaling strategy on the MedQA and Medbullets datasets.	116
6.7	Accuracies with the Iterative Scaling strategy on the MedQA and Medbullets datasets.	116

List of Tables

1.1	Overview of Research Work in this Thesis.	6
3.1	The proportion of each dialogue act in the MedDG and KaMed datasets.	44
3.2	Automatic evaluation results on MedDG. † denotes statistically significant differences ($p < 0.05$).	46
3.3	Automatic evaluation results on KaMed. † denotes statistically significant differences compared to all baselines ($p < 0.05$).	46
3.4	Human evaluation results on MedDG.	48
3.5	Results of Dual Flow Modeling with ablation. The best results are in bold	48
3.6	Case study. The responses are generated by different models, where key entities are underlined.	50
4.1	Automatic evaluation results on MedDG. † denotes statistically significant differences ($p = 0.05$).	65
4.2	Automatic evaluation results on KaMed. † denotes statistically significant differences ($p = 0.05$).	67
4.3	Human evaluation results on MedDG.	69
4.4	Ablation results on two datasets.	71

4.5	Case study. The responses are generated by different models, where diseases are underlined.	73
5.1	Statistics of MEDIATOR	86
5.2	An example of the generated thought process.	87
5.3	Human evaluation results of automatically generated thought processes. Values in the table represent average valid percentages.	88
5.4	Automatic evaluation results on two datasets. † denotes statistically significant differences ($p < 0.05$).	89
5.5	Automatic evaluation results based on GPT-4.	91
5.6	Ablation study on two datasets	93
6.1	The number of samples for each evaluation dataset	110
6.2	Comparisons of RAR ² with other medical large language models and tuning-free RAG methods. † denotes the in-domain dataset.	111
6.3	Comparisons of RAR ² with other RAG methods that involve fine-tuned components. † denotes the in-domain dataset.	112
6.4	Ablation study on several medical question answering datasets.	114

Chapter 1

Introduction

1.1 Background

Dialogue systems, enabled by advances in large language models (LLMs), have become a prominent area of research and application over the past decade [127, 152, 71, 89, 138, 139]. They operate in two primary settings, namely open-domain conversation [28, 60] and task-oriented services such as reservation booking [137] and personalized recommendations [43]. Against this backdrop, medical dialogue systems (MDS) have drawn increasing attention as telemedicine has expanded, particularly since the onset of the COVID-19 pandemic [71, 152]. By delivering online consultations, symptom triage, diagnostic support, and prescribing, MDS broaden access to timely and affordable care and help reduce workload for healthcare professionals [160]. In practice, MDS are deployed as patient-facing virtual doctors and clinician-facing copilots that integrate with clinical workflows and electronic health records [101, 100]. The domain imposes distinctive requirements for clinical validity and safety, transparency of diagnostic rationale, privacy and security, fairness across populations, and regulatory compliance [140, 138]. Thus, building clinically usable MDS remains a substantial challenge.

In recent years, research on MDS has broadened from a focus on automatic diag-

nostic systems to general medical dialogue generation [127, 141, 159, 133, 71, 152]. Automatic diagnostic systems aim to elicit relevant symptoms and history through multi-turn interactions and to produce an accurate diagnosis in as few turns as possible [127]. Medical dialogue generation focuses on building models that produce medically accurate, contextually appropriate, and clear responses to meet diverse patient needs [58, 101, 100]. With advances in pretrained language models and LLMs, MDS have shown promising potential as medical assistants and even virtual doctors through the incorporation of external medical knowledge, clinical pathways, and Chain-of-Thought (CoT) reasoning [139, 131, 61, 26, 153, 138, 33, 40].

Although substantial advances have been made, current MDS still fall short of clinical expectations for doctor-patient communication. The main challenge is to emulate key aspects of clinical processes so that the system’s behavior aligns with routine clinical practice in terms of clinical accuracy, reliability, and safety [100]. One of the most important aspects is ensuring the dialogue proceeds in a clear, clinically logical sequence [52]. According to the well-established Calgary-Cambridge framework for the medical interview, a consultation typically progresses through initiating the session, gathering information, conducting a physical examination where feasible, explaining and planning, and closing the session [52]. Qualified MDS is expected to follow a similar sequence and generate phase-appropriate responses [131]. Existing MDS commonly model and predict medical entities to maintain coherence across turns, leveraging entity relations in the dialogue context and external medical knowledge graphs [71]. However, entity prediction alone provides an incomplete signal for managing dialogue progression and fails to capture phase transitions. Effective progression management also requires additional signals, such as dialogue acts or explicit phase tracking [139].

Another aspect is maintaining an explicit differential diagnosis (DDx) that is updated at each turn. A DDx is a structured list of candidate diseases or conditions that could explain the patient’s symptoms and findings, often prioritized by likelihood and risk [77, 112]. In practice, clinicians maintain this list implicitly, revising it continuously

throughout the consultation [52]. The evolving set of candidate disease hypotheses can guide what information to elicit and how to formulate each response. It is therefore crucial to model this process and generate responses grounded in a specific disease hypothesis [140]. Unfortunately, most MDS generate responses by relying on the implicit diagnostic reasoning of generative models and therefore do not provide a clear, interpretable disease hypothesis [139, 71, 152, 100, 157]. Their outputs are also susceptible to redundant or irrelevant information in the dialogue history and may not consistently adhere to a specific disease hypothesis. Maintaining an explicit, turn-level DDx can help improve response reliability and even monitor dialogue progression.

A third aspect is to model prototypical clinical reasoning processes. Clinicians' reasoning is grounded in evidence-based medicine, a systematic approach to clinical decision making that integrates the best available evidence with clinical expertise and patient values [27]. In practice, clinicians synthesize medical knowledge with patient-specific information and draw on their clinical experience to reason comprehensively and rigorously [78]. Moreover, clinician preferences (stable tendencies in risk tolerance, prior beliefs, and decision thresholds) shape diagnostic reasoning beyond simple symptom-disease co-occurrence [80, 150]. For example, a safety-first clinician may prioritize ruling out a low-probability but high-severity condition before committing to a more likely diagnosis, even when the latter is strongly supported. Existing research has paid relatively little attention to knowledge-grounded diagnostic reasoning and often overlooked clinicians' internal reasoning [138]. Nevertheless, these two dimensions reflect prototypical clinical reasoning and, to some extent, differentiate novices from experts [27, 49]. Enhancing clinical reasoning across these two dimensions can further improve the quality of generated responses, particularly in complex patient presentations.

Overall, this thesis develops a clinical process-aware medical dialogue system that emulates doctor-patient communication. It formulates and addresses research problems of practical significance and academic value, contributing to the medical AI community.

The remainder of this chapter outlines the key research problems and presents a comprehensive overview of our approaches and findings.

1.2 Research Problems

The primary objective of this thesis is to explore novel and effective approaches to emulate aspects of the clinical process and align medical dialogue systems with clinical practice. To this end, we address the following three key research problems:

- Problem 1: How to model the critical information flow (e.g., medical entities and dialogue acts) in the medical dialogue and incorporate key entities and acts into response generation?
- Problem 2: How to develop a comprehensive differential diagnosis (DDx) and leverage it to improve the reliability and interpretability of system responses?
- Problem 3: How to emulate clinicians' internal reasoning processes and improve a system's reasoning capabilities with external medical knowledge?

The first problem concerns modeling critical information flow to improve medical dialogue understanding. Medical dialogues revolve around two essential dimensions: medical entities (e.g., symptoms, tests, medications) and dialogue acts (e.g., inquiry, confirmation, recommendation). The main challenge lies in modeling how these signals evolve across turns (i.e., the transition dynamics of medical entities and dialogue acts in the dialogue history). Because their temporal trajectories exhibit characteristic transition patterns in clinical practice, explicitly modeling these transitions can improve next-turn predictions of both the entities and the acts. Conditioning response generation on these predictions yields outputs that are more accurate and logically consistent with the ongoing clinical communication.

The second problem focuses on differential diagnosis (DDx) as a transparent basis for response generation. Building on the first problem, ensuring the reliability and interpretability of generated responses remains challenging. Instead of generating responses directly from the dialogue history, constructing a comprehensive DDx and conditioning response generation on the set of candidate diseases can alleviate concerns about reliability and interpretability. In particular, clinicians are more likely to trust such a system when they can understand and endorse how each generated response is grounded in a specific disease hypothesis. The main challenge in constructing a DDx is designing a framework that leverages the typical clinical diagnostic reasoning processes, namely intuitive and analytic reasoning.

The third problem dives deeper into the internal reasoning process of clinicians. Although the second problem offers a preliminary exploration of the diagnostic reasoning underpinning DDx construction, little attention has been paid to modeling clinicians' internal reasoning processes. Enabling MDS to emulate clinicians' reasoning processes remains a long-standing, under-addressed problem. The main challenges of this problem are twofold. On the one hand, clinicians' reasoning goes beyond simple recognition of symptom-disease co-occurrence and reflects complex, context-dependent clinician preferences. It is necessary to develop MDS that aligns with clinicians' internal reasoning processes. On the other hand, given the breadth and high-stakes nature of medicine, the system needs to ground its reasoning in external medical knowledge, such as guidelines, textbooks, curated knowledge graphs, and literature. The ability to reason over external knowledge, attribute sources, and keep references up to date is crucial for MDS to be clinically useful. Given that reasoning is a core capability of MDS, tackling this problem with innovative methods would materially advance the broader field.

Research Problem	Research Work	Publication Venue
Problem 1	Work 1: Medical Dialogue Generation via Dual Flow Modeling	Findings of ACL 2023 [139]
Problem 2	Work 2: Medical Dialogue Generation via Intuitive-then-Analytical Differential Diagnosis	arXiv [140]
Problem 3	Work 3: Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment	Findings of ACL 2024 [138]
	Work 4: RAR ² : Retrieval-Augmented Medical Reasoning via Thought-Driven Retrieval	Findings of EMNLP 2025

Table 1.1: Overview of Research Work in this Thesis.

1.3 Research Overview and Contributions

This thesis aims to emulate key aspects of clinical processes to align system behavior with routine clinical practice in terms of clinical accuracy, reliability, and safety. Despite recent work (see Chapter 2), the scope and methods of prior approaches remain insufficient to address our research problems. Motivated by these three research problems, this thesis comprises three parts, with an overview of the studies shown in Table 1.1. Specifically, we first explore the transition of critical information (i.e., medical entities and dialogue acts) in medical dialogues to improve dialogue understanding and response generation. Then, we focus on the design of a comprehensive differential diagnosis to improve the reliability and interpretability of system responses. Finally, we delve into the internal reasoning process of clinicians and improve a system’s reasoning capabilities with external medical knowledge. For problem 1, we propose a dual flow enhanced medical dialogue generation framework which models the transitions of

medical entities and dialogue acts via step-by-step interweaving (work 1). For problem 2, we explicitly models a differential diagnosis with intuitive-then-analytic processes and incorporates diagnosis to guide response generation (work 2). For problem 3, we first emulates clinicians’ diagnostic reasoning processes and aligns with clinician preferences (work 3), and then propose a joint learning framework that simultaneously improves reasoning-augmented retrieval and retrieval-augmented reasoning.

Work 1: Dual Flow Enhanced Medical Dialogue Generation

An efficient understanding of the medical dialogue history is crucial for response generation. To improve the dialogue understanding, we aim to capture the transitions of critical information in each turn, the medical entities and the doctor’s dialogue acts, by modeling two types of information flows. In Work 1, we propose a dual flow enhanced medical dialogue generation framework (DFMed). At each dialogue turn, it extracts the medical entities and the dialogue acts used in the dialogue history, and models their transitions with an entity-centric graph flow and a sequential act flow, respectively. Two sequential models are constructed to encode their transitions, with an interweaving component to enhance their interactions. The output representations are then used to predict the entities and the acts to be adopted in the following turn, which are employed to guide the response generation through gate control.

Contributions. In Work 1, we propose a dual flow enhanced medical dialogue generation framework, DFMed, that models a medical entity flow and a dialogue act flow to improve the relevant entity selection and dialogue act prediction. Besides, we design an interweaver to strengthen interactions of two flows. The selected entities and predicted acts are applied to guide response generation. Experiments validate the effectiveness of our DFMed on two datasets.

Work 2: Differential Diagnosis Guided Medical Dialogue System

In Work 1, we focus on the critical information flows (i.e., medical entities and dialogue acts) to improve dialogue understanding and use the predicted entities and acts to enhance response generation. However, the reliability and interpretability of the generated responses remain open challenges. To improve the reliability and interpretability, we produce a differential diagnosis and then utilize potential diseases to guide response generation. We draw inspiration from the diagnostic reasoning research and design a two-stage (i.e., intuitive-then-analytical) differential diagnosis method. In the intuitive stage, a preliminary disease list is retrieved based on patient cases and disease documents. Then in the analytical stage, the preliminary list is refined through a multi-disease classifier with multi-task learning. A disease knowledge graph is also incorporated to provide reliable external information for refinement. Finally, the cases and documents relevant to the refined potential diseases are utilized to enhance response generation.

Contributions. In Work 2, we propose a medical dialogue generation framework with a differential diagnosis, IADDx, which explicitly models the process of differential diagnosis through intuitive association followed by analytic refinement. Moreover, we devise a diagnosis-oriented graph to interpret the differential diagnosis. The diagnosis results are utilized to guide response generation. Experiments on two datasets demonstrate the efficacy of IADDx. Besides, we generate intermediate outcomes and graph-based diagnostic paths to help understand the diagnostic procedure.

Work 3 & 4: Diagnostic Reasoning Aligned Medical Dialogue System

Our previous studies try to model the surface-level information and rough emulation of differential diagnosis. Building on this foundation, we further investigate prototypical

clinical reasoning processes that approximate clinicians’ internal diagnostic reasoning. Emulate clinicians’ diagnostic reasoning process has been the long-standing research focus. Instead of the “output” of the reasoning process (e.g., the symptoms and diseases mentioned in clinicians’ utterances) as existing works, we explore the clinicians’ internal thought process and specifically model the clinician preference. Our Work 3 develops a medical dialogue system that can align with the internal diagnostic reasoning process of clinicians. We first model diagnostic analysis as iterative abductive-deductive reasoning that generates explanatory hypotheses for the patient’s condition and evaluates their adequacy against available evidence. This analysis provides a robust and comprehensive foundation for accurate diagnoses. Then, we align response generation with the clinician preference based on the analysis and dialogue context. The thought process of how clinicians reason and generate responses is a vital resource for learning their preferences. Extracting the thought process and modeling it together with responses can help learn the diagnostic reasoning process.

Given the breadth and high-stakes nature of medicine, Retrieval-Augmented Generation (RAG) has become a key approach to mitigating knowledge gaps and hallucinations by grounding outputs in external medical sources. However, RAG can struggle on complex medical questions that demand intensive reasoning, because surface-level cues often under-specify the underlying information needs. Our Work 4 constructs a thought process to reason through medical questions and identify their implicit knowledge requirements. This thought process is directly used to retrieve relevant information, which is subsequently incorporated into reasoning to derive the final answer. To optimize both thought and answer generation, we propose a joint learning framework, RAR², that simultaneously improves Reasoning-Augmented Retrieval and Retrieval-Augmented Reasoning. Specifically, we first construct a training dataset consisting of mixed preference pairs. One type is thought pairs, in which a sampled thought process is annotated based on the outcome of subsequent retrieval-augmented generation. The other type is answer pairs, in which a sampled answer is annotated

according to its correctness. We then apply Direct Preference Optimization (DPO) to fine-tune the LLM. These preference pairs enable supervised preference learning, allowing the model to identify relevant knowledge and reason effectively with external information.

Contributions. In Work 3, we propose a novel medical dialogue system framework, EMULATION, that emulates clinicians’ diagnostic reasoning processes to generate appropriate responses grounded in abductive and deductive diagnostic reasoning analysis and alignment with clinician preferences. Besides, a new diagnostic thought process corpus is presented and utilized to model the clinician preference. Experimental results demonstrate the effectiveness of EMULATION on two datasets.

In Work 4, we propose RAR², a novel joint learning framework designed to enhance LLM reasoning capabilities within the RAG framework, specifically targeting complex medical questions. Unlike existing RAG approaches, which rely primarily on direct query refinement, RAR² explicitly constructs and optimizes a thought-based reasoning process to uncover implicit knowledge requirements. We introduce a mixed preference dataset comprising thought pairs and answer pairs and leverage DPO for joint training. Extensive experiments on multiple biomedical question answering datasets demonstrate that RAR² outperforms state-of-the-art RAG baselines, both with and without fine-tuning. Moreover, our analysis of test-time scaling strategies validates the scalability and robustness of RAR², highlighting its potential to significantly improve medical reasoning and decision-making in clinical scenarios.

1.4 Thesis Structure

To provide a comprehensive overview, we structure the thesis as follows.

Chapter 1 provides background on medical dialogue systems, highlights limitations of existing approaches, and motivates the new problem settings addressed in this thesis.

It then articulates the three key research problems, outlines the overall research plan, and summarizes the thesis contributions.

Chapter 2 presents a comprehensive review of the literature on medical large language models, medical dialogue systems, and LLM-based medical reasoning.

Chapter 3 addresses challenges in medical dialogue understanding and introduces DFMed, a dual flow enhanced medical dialogue generation framework. DFMed models the transitions of medical entities and dialogue acts in each turn and improves generation by accurately predicting the entities and acts for the next turn.

Chapter 4 explicitly models the diagnostic process to establish a clear, interpretable basis for response generation and introduces IADDx, a framework that first generates a comprehensive differential diagnosis and then uses the resulting candidate diseases to guide response generation.

Chapter 5 investigates clinicians' internal reasoning processes and decision-making, and introduces EMULATION, a framework that first emulates diagnostic reasoning and then aligns response generation with clinician preferences.

Chapter 6 advances retrieval-augmented medical reasoning by constructing and optimizing a reasoning process that explicitly uncovers underlying knowledge requirements. To jointly optimize both knowledge-need identification and knowledge-aware answer generation, we propose RAR², a joint learning framework that simultaneously enhances reasoning-augmented retrieval and retrieval-augmented reasoning.

Chapter 7 synthesizes the proposed approaches, key findings, and overall contributions of the thesis, and outlines directions for future research.

Chapter 2

Literature Review

This chapter provides an in-depth survey of prior research that informs the work presented in this thesis. It primarily highlights advancements in medical large language models, medical dialogue systems, and medical reasoning. Through this review, we seek to extract key insights and establish a robust groundwork for the development of our own approach.

2.1 Medical Large Language Models

Medical large language models (Med-LLMs) are large language models adapted to the medical domain using domain-specific corpora, training objectives, and safety guardrails [105, 106, 122, 115, 121]. They are aligned to clinical language, knowledge, and workflows, and are increasingly applied to medical dialogue and triage, clinical documentation and summarization, literature- and guideline-grounded question answering, and decision support within care pathways [155, 16, 36]. Recent systems also integrate tool use, including retrieval over biomedical corpora and electronic health records (EHRs) as well as calculators and order sets [143, 53]. Some additionally accept multimodal inputs, such as imaging, physiological waveforms, or structured

laboratory data, to better reflect clinical practice [57, 154]. This section reviews model development, followed by comprehensive evaluation.

2.1.1 Model Development

Recent years have seen rapid progress in Med-LLMs, driven by both proprietary and open-source efforts [82, 105, 106, 122, 115, 121]. On the proprietary side, large general-purpose systems (e.g., GPT-4) and domain-specialized models (e.g., the Med-PaLM series) have achieved state-of-the-art results on medical benchmarks [105, 106]. On the open-source side, initiatives adapting LLaMA-family backbones have produced medical models fine-tuned on biomedical literature and clinical data [32, 128, 123, 115, 153, 15, 59, 90]. Collectively, these developments emphasize domain-specific pretraining, instruction tuning, multilingual capabilities, and emerging trends such as multimodality and agentic, tool-using workflows [148, 90, 57, 111, 143].

Among proprietary systems, the Med-PaLM line exemplifies this trajectory [105, 106, 117]. Med-PaLM was the first LLM to exceed the USMLE pass mark (60%) on exam-style questions, achieving 67% accuracy on the MedQA benchmark [105]. Its successor, Med-PaLM 2, reached 86.5% on MedQA—approaching expert-level performance [106]. Med-PaLM 2 combined advances in the base model with medical domain fine-tuning and an ensemble-refinement prompting strategy to strengthen reasoning. In blinded clinician evaluations, its long-form answers were often rated superior to physician-written responses across multiple quality axes [106].

Notably, there has been a surge in fine-tuning research on smaller open-source LLMs. For example, [32] introduces MedAlpaca, built on Alpaca and fine-tuned on approximately 160K diverse medical Q&A pairs, including exam-style questions and health advice, to improve instruction following. Several LLaMA-based systems inject domain knowledge in complementary ways: PMC-LLaMA continues pretraining on large biomedical corpora and is then instruction-tuned for medical Q&A, rationale genera-

tion, and dialogue [128]; BenTsao integrates the Chinese Medical Knowledge Graph (CMeKG), generating about 8K KG-derived instructions (diseases, drugs, symptoms) to ground factual knowledge [123]; and Clinical Camel encodes medical knowledge via synthetic Q&A dialogues distilled from clinical texts [115]. Another example is the HuatuoGPT series [153, 15, 12]. HuatuoGPT is a Chinese medical LLM trained by combining GPT-distilled supervision with real medical dialogues and then applying reinforcement learning from AI feedback, achieving state-of-the-art performance on Chinese medical consultation [153]. HuatuoGPT-II streamlines this approach into a one-stage, instruction-centric pipeline that unifies heterogeneous pretraining and fine-tuning sources with priority sampling, scales model and data sizes, emphasizes Traditional Chinese Medicine and contemporary licensing exams, and reports expert preference win rates over GPT-4 in Chinese medical evaluations [15]. Instead of domain adaptation, Baichuan-M1 is trained from scratch on a large training corpus to balance general-purpose and medical capabilities, and is positioned for clinical scenarios while remaining competitive on general benchmarks [121]. Some studies examine larger model sizes, including Me-LLaMA (13B and 70B, LLaMA-2 based) [134] and Med42-v2 (8B and 70B, Llama 3 based) [22].

Researchers are also expanding medical LLMs beyond text-only inputs to handle multimodal medical data, including images (e.g., radiographs, pathology slides) and physiological waveforms (electrocardiograms, ECGs) [87, 158, 20, 57, 66, 117, 79]. Med-PaLM Multimodal is a research prototype that accepts clinical images alongside text [117]. Other works have combined vision models with LLMs. For example, LLaVA-Med [57] and Med-Flamingo [79] integrate medical image understanding into dialogue generation. A notable example is HealthGPT, a medical large vision-language model (Med-LVLM) that integrates image understanding and generation within an autoregressive LLM framework [66]. It introduces Heterogeneous Low-Rank Adaptation (H-LoRA) to inject visual knowledge into the language model and is trained on VL-Health, a new multimodal dataset of paired medical images and text.

Another research frontier is agentic Med-LLMs, where models that can act as clinical agents or collaborators [111]. Instead of a single static model, this involves multiple specialized LLM-based agents or tool-using agents to solve complex medical problems. MedAgents is one framework that deploys multiple LLM agents in a role-playing collaborative discussion to perform medical reasoning in a zero-shot manner [111]. Other work presents hierarchical agent frameworks that incorporate external tools [143].

2.1.2 Comprehensive Evaluation

As Med-LLMs achieve high exam scores, the focus has shifted toward rigorous evaluation of clinical competence and deployment safety [45, 105, 16]. Such evaluation is challenging because models must not only recall facts but also reason, maintain well-calibrated uncertainty, avoid harmful errors, and handle sensitive medical content ethically [118, 96, 84, 2].

Early evaluations of Med-LLMs relied heavily on exam-style QA benchmarks, including MedQA [45], MedMCQA [83], and PubMedQA [47]. However, as prior work notes, passing such exams is necessary but not sufficient for clinical readiness [6]. In response, the community has developed comprehensive benchmarks that better approximate real-world clinical tasks [106, 68, 129]. MultiMedQA broadens evaluation by adding the HealthSearchQA dataset of consumer health queries and a human evaluation rubric for long-form answers [106]. MedBench targets Chinese Med-LLMs, assembling 300,901 questions across 43 clinical specialties and providing a standardized cloud-based platform for multi-faceted assessment of medical knowledge, reasoning, and language [68]. MedS-Bench focuses on advanced clinical tasks beyond QA and spans eleven task families, such as clinical report summarization, treatment recommendation, diagnostic reasoning, and clinical named-entity recognition [129]. In its initial evaluation, MedS-Bench assessed nine leading models, including GPT-4, Llama3-based models, and

Med42-v2. Results show that most models struggle with these complex tasks, even under few-shot prompting. For example, models that perform well on multiple-choice questions often fail to generate coherent care plans or summarize patient histories. These findings highlight a gap between exam performance and practical utility. Furthermore, the HELM framework (Holistic Evaluation of Language Models) defines a taxonomy of 121 clinical tasks across five categories (clinical decision support, note generation, patient communication, research assistance, and administration), developed in consultation with clinicians [6]. The Med-HELM initiative emphasizes multi-metric evaluation, assessing not only accuracy but also bias, calibration, and computational efficiency for each task.

Additionally, the generation of factually incorrect or fabricated content is a well-recognized failure mode of LLMs and is particularly perilous in clinical settings [96, 84]. To address this, recent evaluations increasingly focus on hallucination detection and factuality assessment. MedHalt is a benchmarking initiative that evaluates LLMs’ propensity to hallucinate in medical Q&A, with an emphasis on qualitative comparison of model outputs [84]. MedHallu presents the first dedicated benchmark for medical hallucination detection, comprising 10,000 Q&A pairs derived from PubMedQA, each annotated as correct or hallucinated. It also stratifies hallucinations by difficulty into “easy” (blatant errors) and “hard” (subtle, plausible inaccuracies) [1]. Evaluations of state-of-the-art models on MedHallu revealed poor performance at detecting the “hard” hallucinations. Incorporating domain-specific knowledge and permitting a “not sure” response improved detection precision. These results indicate that enabling calibrated uncertainty and selective refusal can reduce hallucination rates by discouraging overconfident, unsupported answers. Beyond automated benchmarks, clinician review of model outputs for errors remains an essential component of evaluation.

2.2 Medical Dialogue System

Medical dialogue systems are designed to conduct comprehensive medical consultations, including diagnosis, treatment planning, and health advice [100]. With the rise of telemedicine and virtual health assistants, research in this area has accelerated in recent years [152, 160]. This review surveys key developments in three areas: (1) dataset creation for medical dialogue, (2) automatic diagnostic dialogue systems, and (3) medical dialogue generation.

2.2.1 Medical Dialogue Datasets

A fundamental challenge in developing medical dialogue systems is the paucity of large, high-quality conversational data, owing to privacy constraints and the clinical expertise required to curate such dialogues [152]. Early work drew on online health forums to synthesize medical dialogues [127]. For example, a dataset compiled from a Chinese medical forum includes patient self-reports and doctor-patient conversations, with symptom mentions extracted to support dialogue-based diagnosis [127]. However, these early datasets were relatively small and often emphasized single-turn question-answer pairs rather than multi-turn dialogues.

Recently, substantial multi-turn medical dialogue datasets have become available in both English and Chinese [152, 71, 160, 58]. A landmark contribution is the MedDialog dataset which contains around 1.1 million Chinese medical dialogues and around 0.3 million English medical dialogues [152]. These dialogs cover a broad range of hundred-plus medical specialties and diseases, making MedDialog the largest open medical dialogue corpus to date. Each consultation typically includes the patient’s initial description of their condition (symptoms, history, etc.) followed by multiple doctor-patient exchanges, and often a final diagnosis/ treatment suggestion. The

data were scraped from online telemedicine platforms (e.g. Haodf¹ in Chinese, and iCliniq² and HealthCareMagic³ in English). MedDialog has been pivotal in training and benchmarking dialogue models, enabling researchers to pretrain large transformer-based models on medical conversations. Another important dataset is MedDG, an entity-annotated, Chinese medical dialogue corpus comprising approximately 17,000 multi-turn dialogues collected from an online health community [71]. Each utterance in MedDG is labeled with five medical entity types, i.e., disease, symptom, attribute, test, and medication. This fine-grained, entity-annotated dataset marks a shift toward entity-centric dialogue modeling and provides a high-quality benchmark for evaluating medical dialogue systems on entity use and accuracy [58]. In response to the COVID-19 pandemic, the CovidDialog dataset was introduced in [160]. It contains approximately 1,000 dialogues focused on COVID-19 symptoms, testing, and treatment, and as a narrow-domain resource it underscores the need for transfer learning.

2.2.2 Automatic Diagnosis Systems

One important class of medical dialogue systems focuses on automatic medical diagnosis, namely engaging a patient in conversation to elicit symptoms and ultimately predict the underlying disease [127, 13, 112, 133, 67, 141, 159]. This task is typically formulated as a sequential decision-making problem: at each turn, the system chooses which question to ask or which symptom to probe and, when sufficiently confident, makes a diagnosis. Unlike open-domain conversation, automatic diagnosis systems are goal-oriented, and success is measured by making an accurate diagnosis while asking as few unnecessary questions as possible. Researchers have explored multiple approaches for such systems, including reinforcement learning (RL) to optimize the question-selection policy, knowledge-graph-based reasoning over medical entities, and

¹<https://www.haodf.com/>

²<https://www.icliniq.com/>

³<https://www.healthcaremagic.com/>

hybrid methods that combine these techniques [13, 112, 133, 67, 141, 159].

The work presented in [127] was among the first to implement an end-to-end automatic diagnostic dialogue system. The system used a task-oriented dialogue policy learned via reinforcement learning, specifically a Deep Q-Network (DQN), and employed template-based doctor responses. The reward function encouraged accurate diagnoses, and the system was able to elicit important symptoms that were not present in the patient’s self-report, thereby significantly improving diagnostic accuracy relative to using the initial description alone. This demonstrated the feasibility of diagnosis through dialogue and established an early baseline. Building on this framework, subsequent work by [67] enhanced the symptom-based diagnostic model by introducing a symptom-association graph and a global attention mechanism. Instead of treating symptoms as independent features, the authors constructed a graph in which nodes represent symptoms and edges encode co-occurrence or clinical relationships (e.g., “cough” and “fever”). During the dialogue, global attention over this graph enabled the system to consider related symptoms that had not yet been mentioned. Besides, to directly incorporate a knowledge graph into the dialogue policy, an end-to-end knowledge-routed relational dialogue system (KR-DS) was proposed in [141]. KR-DS employs a custom deep reinforcement learning agent, KR-DQN, with two branches: one that refines the dialogue state using relational knowledge (capturing dependencies among symptoms and symptom-disease pairs) and one that performs knowledge-guided action selection.

As the number of medical entities grows, RL-based diagnostic systems face a key challenge: a large action space comprising hundreds of symptoms to query plus a stop-and-diagnose action. This has motivated research into hierarchical policies and other action-space reduction strategies. [159] address this complexity by proposing a two-level hierarchical policy for diagnostic dialogue. In their framework, a high-level “master” policy first selects a symptom category, and a low-level, specialized symptom-checker policy then queries specific symptoms within that category. This

mirrors a doctor’s approach of first narrowing to an organ system (e.g., respiratory or gastrointestinal) and then drilling down, thereby enabling the system to handle larger disease sets and longer symptom lists while mitigating combinatorial explosion. Another trend has been to combine reinforcement learning with supervised learning. In [151], an RL agent handles symptom inquiry while a separate classifier performs disease diagnosis, and the two modules are aligned during training. Such hybrid approaches improve stability and accuracy by assigning clear roles to the disease classifier and the RL policy. An alternative to RL is modeling the dialogue and diagnosis processes in a generative probabilistic framework. [58] proposes a semi-supervised variational reasoning model that unifies dialogue generation with latent reasoning about patient state and doctor’s actions.

Despite these advances, automatic diagnostic dialogue systems face significant challenges before they can approach real clinical utility. The main challenge is that these systems rely on fixed response templates, which yield a poor user experience and prevent them from conveying nuanced diagnoses or multiple recommendations simultaneously [127]. With the advancement of pretrained language models and LLMs, research on medical dialogue systems has shifted toward dialogue generation, which enables more flexible and diverse responses [71, 61, 5, 153, 15].

2.2.3 Medical Dialogue Generation

Beyond diagnosis, medical dialogue generation focuses on enabling models to conduct coherent, informative conversations in clinical settings [71, 100]. The goal is to produce responses that are medically accurate, contextually appropriate, and clear, akin to those of a clinician or health advisor. In recent years, methods have evolved from sequence-to-sequence neural models with limited medical knowledge to large-scale transformer-based language models and knowledge-enhanced generation [152, 131].

Early medical dialogue generation models adopted architectures from open-domain

chatbots, such as RNN-based sequence-to-sequence (seq2seq) and Transformer models, and were trained on medical dialogue corpora [71]. For example, using the Chinese MedDialog dataset, researchers compared three architectures, namely Transformer, BERT-GPT, and GPT, with and without pretraining on general dialogue data. Common automatic metrics (BLEU, perplexity, and Distinct-n) were used to benchmark these models [152]. However, pure sequence models often struggled with factual accuracy and could produce plausible-sounding but medically incorrect statements when data for a condition were sparse. They also lacked mechanisms to ensure that critical medical knowledge appeared in their outputs beyond what was present in the training data.

To address knowledge gaps, a line of work augments generation with external medical knowledge and explicit fact memory [131, 33, 119, 114]. A representative example is MKMed, which integrates a medical knowledge base into a generative dialogue model via memory networks [131]. The system first identifies medical entities in the dialogue history, then uses them to retrieve relevant triples from a knowledge graph (e.g., (Disease, Symptom, X) or (Drug, Indication, Y)). These facts are written to a memory module and retrieved during response generation. Similarly, MDKG uses a graph-based generator with a graph neural network (GNN) over a medical knowledge graph to guide response generation. Another facet of knowledge grounding is entity-centric evaluation [81]. Experiments on MedDG show that even advanced models often fail to mention all relevant symptoms and may hallucinate symptoms not mentioned by the patient [71]. Retrieval-based models, which select physician responses from the dataset, perform better on factual correctness, suggesting that open-ended generation still lags. This has motivated hybrid retrieval-generation approaches, in which a generative model is constrained or informed by retrieved candidate responses or reference passages to keep outputs consistent with the evidence.

The emergence of LLMs such as GPT-4, and in particular the open-source LLaMA, has substantially influenced medical dialogue generation [100, 61, 136, 153, 15]. Rather

than training models from scratch on medical data, recent work focuses on adapting general LLMs to the medical domain. These billion-parameter models possess broad world knowledge and strong language fluency, but they require domain adaptation and safety alignment for clinical use. ChatDoctor [61] fine-tuned a 7-billion-parameter LLaMA model on roughly 100,000 medical dialogues to enable medical consultation. Fine-tuning markedly improved the model’s ability to understand patient questions and generate appropriate medical advice. ChatDoctor also incorporates a self-directed retrieval capability, allowing the model to autonomously query external medical knowledge sources and integrate retrieved facts into its responses. This framework reflects a broader trend in medical dialogue generation: fine-tuning on domain-specific dialogues and augmenting with tools to keep information up to date and accurate. In the Chinese research community, similar efforts have produced models like DISC-MedLLM [5], DoctorGLM [136], and HuatuoGPT series [153, 15].

2.3 LLM-based Medical Reasoning

As a core element of clinical decision-making [54], medical reasoning entails combining domain knowledge, patient data, and situational context to produce accurate diagnoses and effective management strategies [124]. Recent research has focused on enhancing LLMs with transparent and verifiable reasoning abilities, using both training-time and test-time strategies [12, 41, 149, 124]. In what follows, we review key aspects of medical reasoning research, including benchmarks and evaluation, knowledge-augmented reasoning, and agentic reasoning.

2.3.1 Benchmarks and Evaluation

LLMs for medical reasoning are typically evaluated on specialized benchmarks that span a range of clinical reasoning tasks [63, 161]. Medical QA benchmarks have

been crucial for standardizing evaluation, enabling fair comparisons across models, and tracking progress over time [45, 83, 47, 10]. For example, MedQA comprises multiple-choice questions from the United States Medical Licensing Examination [45]; MedMCQA is a large collection of multi-subject multiple-choice questions from medical school curricula [83]; and PubMedQA targets research-level questions, requiring models to read abstracts and answer in yes/no/maybe categories [47]. MultiMedQA [106] aggregates diverse tasks, including consumer health queries from the web (HealthSearchQA and LiveQA) and medication-focused questions (MedicationQA), thereby broadening coverage beyond professional exam items. This diverse benchmark allows testing LLMs on both structured exam-style problems and open-ended questions that require detailed explanations. JAMA Clinical Challenge and Medbullets are two recent evaluation datasets from challenging clinical cases [10]. Several proprietary models demonstrate strong performance across these benchmarks [105]. GPT-3.5 scored 60% on USMLE-style questions, while GPT-4 reached 86%, surpassing the passing threshold by a wide margin. Domain-tuned models like Med-PaLM were the first to exceed the USMLE pass mark [105]. Med-PaLM’s successor, Med-PaLM 2, further pushed performance to 86.5% on MedQA and achieved state-of-the-art results on MedMCQA, PubMedQA, and other benchmarks [106]. These results represent a leapfrog improvement over prior biomedical models such as BioBERT [55] and BioGPT [74], underscoring the impact of LLM scale and fine-tuning on knowledge-oriented tasks.

Beyond accuracy, new scoring protocols emphasize clinical utility and safety [7, 161, 106]. Recognizing that answering medical questions is not just about correctness but also reasoning quality, researchers have developed multi-axis evaluation rubrics. [106] introduces a detailed human evaluation where clinicians rate answers on axes such as medical factuality, reasoning rigor, clarity, empathy, and potential for harm. In their evaluations, physician raters compared LLM-generated answers to those written by other physicians. This human-centric scoring supplements automated metrics,

highlighting that evaluation of medical reasoning requires alignment with clinical norms.

Researchers are also expanding evaluation to complex, multi-step diagnostic tasks [95, 77]. Standard QA benchmarks may not capture the iterative reasoning process that clinicians use like taking history, ordering tests, and refining differentials. To bridge this gap, recent work evaluates LLMs on case-based and dialogical settings [95, 118]. For example, [95] converts NEJM Case Records (difficult real-world cases with an unknown diagnosis) into an LLM challenge. GPT-4 is prompted to generate a differential diagnosis for each case after being given the case narrative. Similarly, Google’s AMIE system was tested on 149 Objective Structured Clinical Examination (OSCE)-style patient cases via text dialogue [118]. In these scenarios, the model must ask questions and respond interactively, mimicking a consultation. Evaluation metrics include not only final diagnostic accuracy (e.g. whether the true diagnosis is in the model’s differential) but also process-oriented scores, such as quality of history-taking, appropriateness of recommended tests, and logical consistency of reasoning. Such large-scale evaluations have revealed that current top models approach clinician-level diagnostic accuracy on many cases [77], but may still falter on rare diseases or require careful prompt design for optimal results [93]. This highlights that while LLMs excel on broad knowledge and common cases, specialized diagnostic support systems or domain fine-tuning still hold an edge in certain niche settings.

To ensure fair and comprehensive evaluation, researchers are addressing known limitations by introducing adversarial test sets and real-world evaluations [106]. For example, Med-PaLM 2 was tested on adversarial questions, including items probing health-equity concerns and clinical edge cases, to identify failure modes such as misleading or biased outputs [106]. Similarly, some evaluations include bedside consultations (real clinical questions posed by physicians during patient care) to assess whether LLMs can assist in practice [106, 118]. These multifaceted evaluations (combining automated metrics, expert human judgment, and scenario-based testing) provide a

stronger foundation for measuring progress in LLM-based medical reasoning.

2.3.2 Knowledge-Augmented Reasoning

One major line of work for improving medical reasoning in LLMs is knowledge augmentation, in which a model’s internal knowledge is supplemented with external medical information [113, 130, 14]. Despite pretraining on massive corpora, LLMs can hallucinate facts or omit domain-specific details, particularly in medicine, where up-to-date and verifiable knowledge is essential [132]. Knowledge-augmented methods address this by retrieving relevant facts from authoritative sources (e.g., biomedical literature and clinical databases) and grounding the model’s reasoning in that evidence. A common approach is retrieval-augmented generation (RAG): the LLM queries an external corpus (e.g., PubMed articles or clinical guidelines) and conditions its answers on the retrieved documents [44]. For example, [44] proposes Self-BioRAG, in which a biomedical LLM decides when to fetch documents from a domain-specific corpus and integrates the excerpts into its reasoning steps. Self-BioRAG identifies salient cues in the question, retrieves relevant documents when necessary, and synthesizes answers from retrieved evidence and its prior knowledge, mirroring how a clinician consults textbooks or the literature for challenging cases.

Beyond plain text retrieval, researchers are integrating structured knowledge bases to bolster LLM reasoning [146, 147, 3, 75, 132]. Medical ontologies and knowledge graphs (such as UMLS or custom graphs of diseases, symptoms, and treatments) can provide a network of verified relationships that the LLM can draw upon. [132] propose MedGraphRAG, which builds a hierarchical graph of medical facts from literature and reference dictionaries, and uses it to retrieve not just raw text but a set of interconnected facts relevant to the query. By linking disorders to their definitions, symptoms, tests, etc., the graph retrieval provides a form of semantic memory for the LLM, improving reliability. This emphasis on evidence grounding markedly increases

the trustworthiness of the model in practical applications, since clinicians can verify and audit the LLM’s output against primary sources. Besides, a key advantage of knowledge augmentation is the reduction of hallucinations and reinforcement of factual accuracy [102]. LLMs paired with a medical literature database tend to answer clinical questions with cited evidence, in contrast to standalone LLMs that might fabricate references or rely on incorrect memory [3, 75]. In summary, augmenting LLMs with external medical knowledge, whether through literature retrieval, specialized databases, or domain-specific knowledge graphs, has been shown to improve factuality and reasoning fidelity. By grounding answers in evidence, these approaches increase both the accuracy and the transparency of medical reasoning.

2.3.3 Agentic Reasoning

A recent development in LLM-based medical reasoning is the shift from single-model inference to multi-agent, role-based collaboration [99, 69, 88, 19, 111]. Rather than relying on one model to complete the entire reasoning process, these approaches coordinate several LLM agents with complementary roles that communicate and critique one another, similar to a multidisciplinary team (MDT) of clinicians. By simulating such consultations, multi-agent systems can achieve deeper and more reliable diagnostic reasoning than a single model.

One notable system is the Multi-Agent Conversation (MAC) framework proposed by [19]. In this setup, four “doctor” LLM agents and one “supervisor” agent collaboratively analyze a rare-disease case. An “admin” agent first presents the patient’s history and findings; the doctor agents then hold a joint discussion, asking clarifying questions and proposing candidate diagnoses and tests, much like clinicians during rounds. The supervisor monitors the discussion, challenges reasoning when needed, and guides the team to a consensus diagnosis and plan. Using GPT-4 as the base model for each agent, MAC was evaluated on 302 difficult cases and outperformed a single GPT-4

model on both initial diagnosis and follow-up queries. These results suggest that simulated peer collaboration can enhance diagnostic reasoning, as agents effectively peer-review one another’s ideas and reduce errors.

Other studies have explored role-play dynamics [88]. A common pattern is a solver-critic pairing: one LLM acts as the primary solver (e.g., generating a differential diagnosis or treatment plan), while a second LLM serves as a critic or verifier that reviews the solver’s output for errors, omissions, or clinical safety concerns and then prompts revisions. This iterative critique process can catch reasoning flaws that a single-pass model may miss. For example, [88] presents a multi-agent framework, Tree-of-Reasoning (ToR), for complex diagnoses. Specialized agents (e.g., one focused on laboratory results, one on imaging, one on symptoms) each propose a partial diagnosis together with an evidence-supported reasoning tree. A cross-validation step then allows the agents to review and correct one another’s proposals, reconciling conflicts to reach a consistent final diagnosis. This design improves interpretability by explicitly recording each step in a tree of evidence and ensures convergence across agent perspectives before output.

Another form of role specialization incorporates a retriever agent into the loop. In a solver-retriever-critic triad, the solver requests supporting data, the retriever locates relevant medical literature or guidelines, and the critic checks whether the solver’s reasoning aligns with the retrieved evidence. This pattern appears implicitly in some retrieval-augmented systems; for example, Self-BioRAG includes a component that decides whether to retrieve and what to retrieve, functioning as an internal retrieval agent [44]. More explicitly, MedAgents proposes an adaptive collaboration in which a “recruiter” agent assesses case complexity and summons either a single generalist or a team of specialist agents as needed [111]. For moderate cases, it may form a small team, whereas for highly complex scenarios it can convene multiple teams in stages, mirroring iterative consultations across specialties. The agents exchange information and produce written reports that a final decision-making agent integrates into the

answer. This adaptive multi-agent workflow models real-world referral pathways and interdisciplinary coordination in health care.

Chapter 3

Modeling Medical Information Flow

For medical dialogue systems (MDS), accurately understanding the dialogue history is crucial, as patients often struggle to describe their symptoms precisely and tend to provide redundant information that may not be relevant for diagnosis [71, 78]. Identifying critical information within lengthy dialogue and predicting relevant content can significantly enhance dialogue comprehension [71, 65, 70]. In this chapter, we further investigate the transitions of critical information and seek to model the flow of information in medical dialogues. Building on this intuition, we propose a **Dual Flow** enhanced **Medical** dialogue generation framework (DFMed). It extracts medical entities and dialogue acts from the dialogue history and models their transitions using an entity-centric graph flow and a sequential act flow, respectively. Furthermore, an interweaving component is built to enhance the interaction between the two flows. Experimental results demonstrate the superiority of DFMed over previous frameworks and highlight the effectiveness of incorporating dual flows to improve medical dialogue generation.

3.1 Introduction

Medical dialogue systems (MDS) have drawn considerable research attention with the increasing demand for telemedicine, especially after the outbreak of the COVID-19 pandemic [152, 71, 160, 34, 133, 144], as they can provide much more people with in-time and affordable access to medical services such as health consultation, diagnosis, and prescription.

For MDS, an efficient understanding of the dialogue history is challenging, as patients usually cannot describe their symptoms precisely and tend to convey lots of redundant information unnecessary for diagnosis [71, 78]. To extract the critical information in the lengthy dialogue, previous research focused on identifying the important medical entities mentioned in the context, such as diseases, medicine, and symptoms [71, 65, 70].

In our work, we argue that capturing the transitions of the medical entities and the doctor’s dialogue acts in each turn (as depicted in Figure 3.1) is also essential for the construction of MDS, which was largely overlooked by previous studies. In medical dialogues, the flows of medical entities and dialogue acts both follow particular transition patterns. For the medical entity flow, entities to appear in the following utterance are usually closely related to the ones mentioned recently. As in Figure 3.1, the entities mentioned in adjacent dialogue rounds are logically related, being neighboring nodes in the medical knowledge. For the dialogue act flow, though variations are allowed to some extent, it usually needs to follow the medical consultation framework suggested in [104]. Modeling these two types of transitions would be helpful in dialogue understanding as they effectively capture how the dialogue history flows. Moreover, learning their transition patterns would also enhance the prediction of the dialogue acts and the medical entities to be adopted in the future turn.

Based on the above intuition, we propose a **Dual Flow** enhanced **Medical** (DFMED) dialogue generation framework. At each dialogue turn, it extracts the medical entities and the dialogue acts used in the dialogue history, and models their transitions with

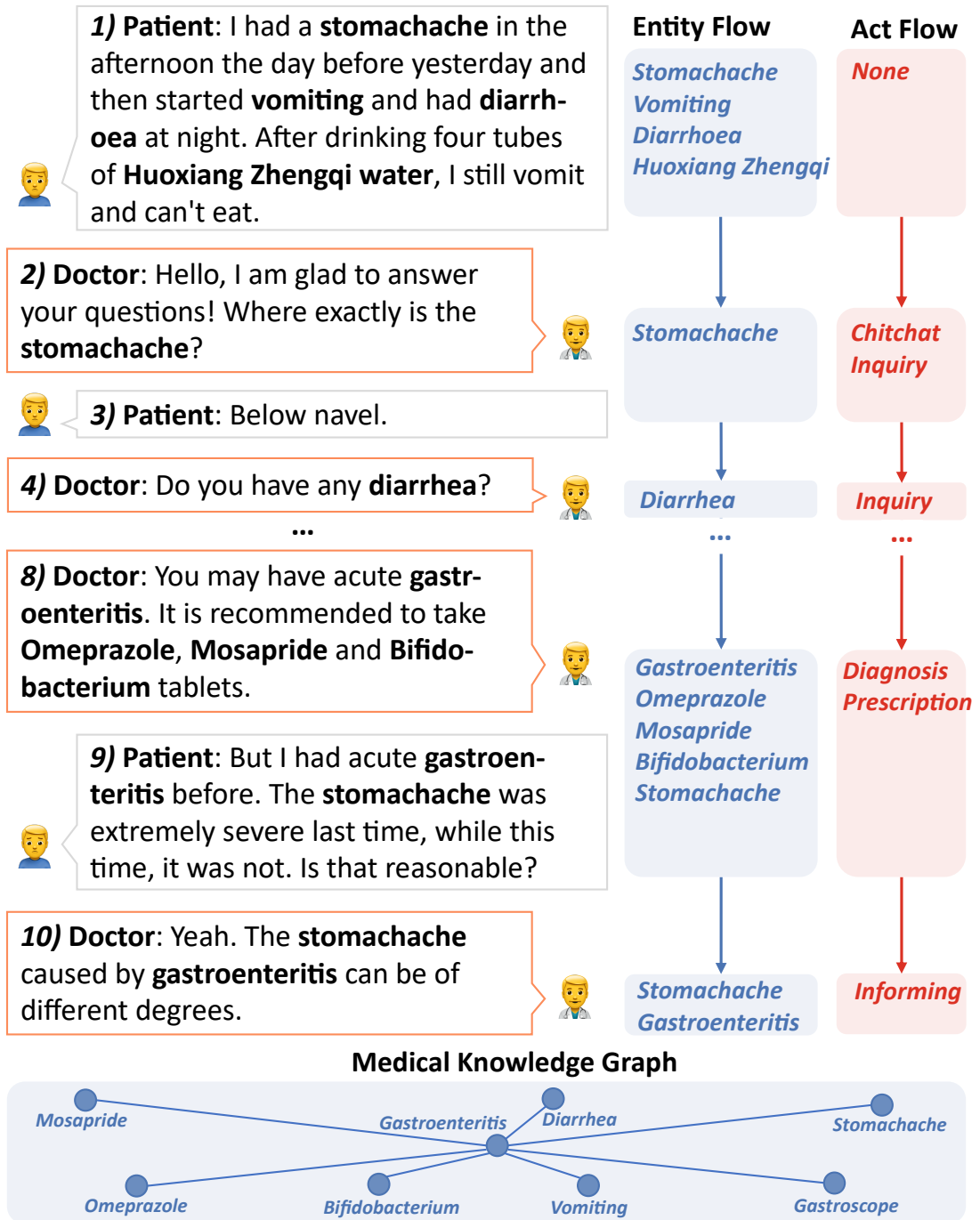


Figure 3.1: An example of a medical dialogue. *Diagnosis* and *Prescription* are short for *Make a diagnosis* and *Prescribe medications*, respectively.

an entity-centric graph flow and a sequential act flow, respectively. Two sequential models are constructed to encode their transitions, with an interweaving component to enhance their interactions. The output representations are then used to predict the entities and the acts to be adopted in the following turn, which are employed to guide the response generation through gate control.

Our main contributions are summarized as follows:

- We propose a novel MDS framework, DFMED, which models the transitions of medical entities and dialogue acts via step-by-step interweaving.
- We summarize the dialogue acts in the medical consultation scenario grounded on the medical documentation standards, SOAP notes [8], including *make a diagnosis*, *prescribe medications*, etc.
- Experimental results show the superiority of DFMED over the previous frameworks and demonstrate the effectiveness of introducing the medical entity and dialogue act flows.

3.2 Preliminary

Problem Formulation. We define a medical dialogue as $U = \{(P_k, D_k)\}_{k=1}^T$, where P and D represent utterances from patients and doctors. Each dialogue turn t is associated with a set of medical entities $E = \{e_i\}$. Doctor turns additionally carry multi-label dialogue act annotations $A = \{a_j\}$. Given the dialogue history $U_t = \{P_1, D_1, \dots, P_t\}$, the system is supposed to generate the t -th doctor utterance D_t .

Dialogue Acts. We summarize several common dialogue acts implied in medical dialogues. One type is medical-related dialogue acts. We design acts that represent a function in a medical documentation standard, the SOAP note [8], and occur

throughout the dialogue. For example, *State a required medical test* and *Prescribe medications* in the “Plan” function of the SOAP note are included in our designed acts. The other type is general open-domain dialogue acts. We choose some acts introduced by [156] and further refine them, such as merging acts that behave as social obligation management into *Chitchat*. Eventually, we obtain 7 dialogue acts for flow modeling, and later experiments demonstrate the effectiveness of guidance for response generation.

3.3 Method

The proposed method learns two flows (i.e., a medical entity flow and a dialogue act flow) to model the propagation of medical entities and medical-related interactions and guide response generation with the corresponding hints. As shown in Figure 3.2, the overall framework contains two modules. The Dual Flow Modeling module learns the entity and act transitions and infers probable entities and acts. The Response Generation module outputs a response under the guidance of the selected entities and predicted acts.

3.3.1 Dual Flow Modeling

Figure 3.3 and Figure 3.4 left displays the architecture of the Dual Flow Modeling module. To extract flow transitions, we encode the medical entities and the dialogue acts in a sequential way. Besides, at each turn, entity and act embeddings are mutually integrated through an interweaving component, named Interweaver. The context states S^c produced by a context encoder also play a role in the integration. After encoding the whole dialogue history, we obtain the entity state S_t^e and act state S_t^a . The entity states are adopted to select the relevant entities. The act states are sent to a multi-label act predictor to estimate the next acts.

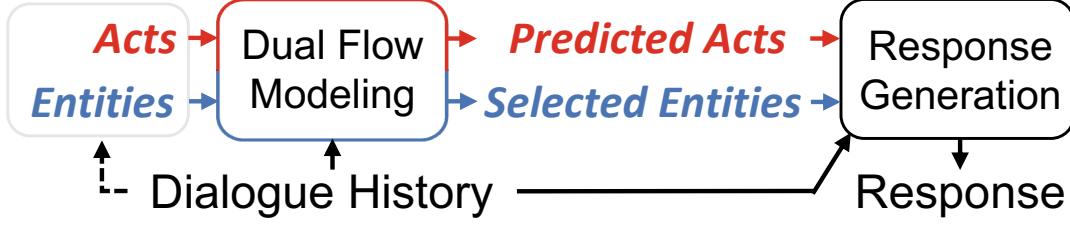


Figure 3.2: The overall framework of DFMed.

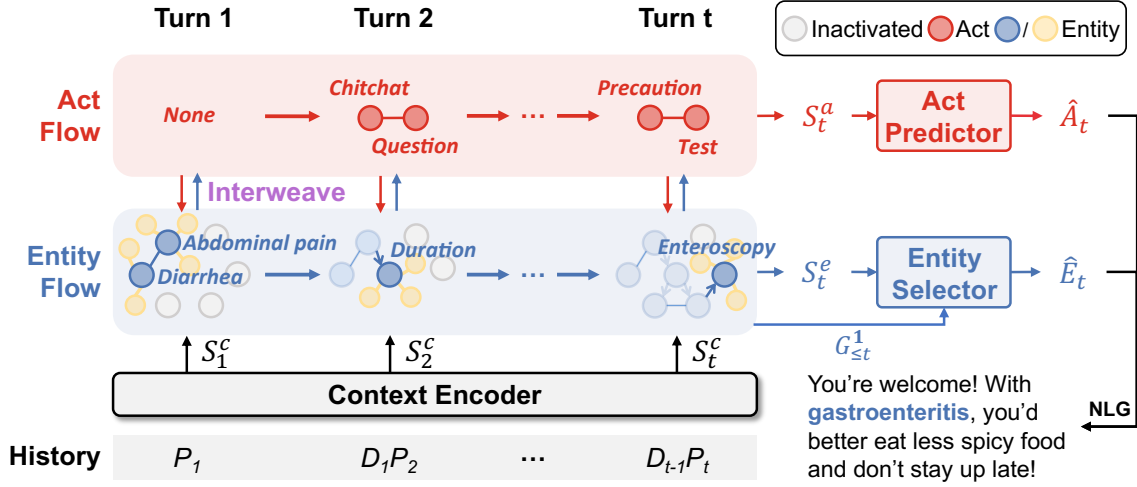


Figure 3.3: The architecture of Dual Flow Modeling. It includes the dialogue act flow modeling in red and the medical entity flow modeling in blue. The one-hop sub-graph nodes of entities in dialogue are marked in yellow.

We apply BERT [25] as the context encoder to encode the dialogue history. All utterances are concatenated with a unique role token (i.e., “[P]” or “[D]”) to separate the patient and doctor utterances. We compute the context states with different history ranges $U_k = \{P_1, D_1, \dots, P_k\} (k \in [1, t])$ for supporting the interweaving at each turn. The mean embedding of tokens in history U_k is applied as the k -th turn context state $S_k^c \in \mathbb{R}^d$, where d is the dimension of the state.

Entity Flow. To learn the sequential transition of medical entities, we create the Entity Flow that encodes an entity graph transition and selects the most relevant entities. Specifically, the flow consists of an entity graph for each turn. Medical entities of each dialogue turn and entities from an external knowledge graph with a one-hop connection to the entities in dialogue are included in the graph of the k -th turn, denoted as G_k . Since we assume entities hop along the links on the graph, the ones in the sub-graph partially provide future transition hints.

We define graphs that include entities until the t -th turn as $G_{\leq t}$. For all entities in graph $G_{\leq t}$, we use the same encoder as above to get entity embeddings. Instead of randomly initializing an embedding, BERT-based encoding keeps token-level semantics. Then, the average token embedding of each entity is employed as the raw embedding, denoted as $h^{e_0} \in \mathbb{R}^d$. Finally, Graph Attention Network (GAT) [120] is implemented to merge neighboring information for each entity:

$$\alpha_{ij}^k = \frac{\exp(\sigma_1(a^\top [W^k h_i^{e_0} \| W^k h_j^{e_0}]))}{\sum_{\mu \in \mathcal{N}_i} \exp(\sigma_1(a^\top [W^k h_i^{e_0} \| W^k h_\mu^{e_0}]))}, \quad (3.1)$$

$$h_i^e = \left[\sigma_2 \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k h_j^{e_0} \right) \right]_{k=1}^K, \quad (3.2)$$

where $h_i^e \in \mathbb{R}^d$ is the updated entity embedding, K is the number of heads, $a \in \mathbb{R}^{2d}$ is a trainable weight, $W^k \in \mathbb{R}^{d_h \times d}$ is the linear transformation matrix for the k -th head, σ_1 and σ_2 are the activation function, and $\mathcal{N}_i$ represents neighboring entities that connect to entity i with one hop. The updated embedding is used to compute each turn’s overall graph embedding via mean pooling:

$$\bar{h}_k^e = \frac{1}{|G_k|} \sum_{i \in G_k} h_i^e, k \in [1, t], \quad (3.3)$$

where $\bar{h}_k^e \in \mathbb{R}^d$ is the entity graph embedding, t represents the turn of the target response. Following [116], we employ a GRU to model the entity graph transition throughout the dialogue. The GRU takes all the previous graph embeddings

$\{\bar{h}_1^e, \bar{h}_2^e, \dots, \bar{h}_t^e\}$ as input and produces the entity state as follows:

$$S_t^e = \text{GRU}_{\text{Entity}}(S_{t-1}^e, \bar{h}_t^e), \quad (3.4)$$

where $S_t^e \in \mathbb{R}^d$ is the final hidden state of the GRU. We denote S_t^e as the entity state that implies clues for the entity transition in the previous context.

Then, we apply the entity state S_t^e to calculate relevant scores for candidate entities. These entities are from the sub-graph of entities in a historical dialogue context. The score is defined as follows:

$$\text{Score} = \langle S_t^e, h_i^e \rangle, i \in G_{\leq t}^1, \quad (3.5)$$

where $\langle, \rangle$ represents a similarity function, h_i^e is the entity embedding, $G_{\leq t}^1$ is the one-hop sub-graphs for entities until the t -th turn. We select top- k relevant entities $\hat{E}_t$ to guide response generation.

Act Flow. To learn the medical-related interactions of the doctor, we design the Act Flow that encodes act sequences and predicts the next acts. Specifically, the flow is composed of the act sequence of each turn. We first randomly initialize the trainable embedding of each dialogue act, denoted as $h^a \in \mathbb{R}^d$. The act sequence of the k -th turn can be defined as $A_k = \{h_1^a, h_2^a, \dots, h_{n_a^k}^a\}$, where n_a^k represents the number of dialogue acts for each turn. Then, we compute the act sequence embedding $\bar{h}_k^a \in \mathbb{R}^d, k \in [1, t]$ through mean pooling.

Similar to the entity flow, we employ a GRU to model dialogue act transition. With all sequence embeddings as input, the final hidden state of the GRU is calculated and denoted as the act state:

$$S_t^a = \text{GRU}_{\text{Act}}(S_{t-1}^a, \bar{h}_t^a), \quad (3.6)$$

where $S_t^a \in \mathbb{R}^d$ is the act state of the t -th turn. Then, the multi-act probability of the t -th turn is computed based on the act state S_t^a with a sigmoid and linear

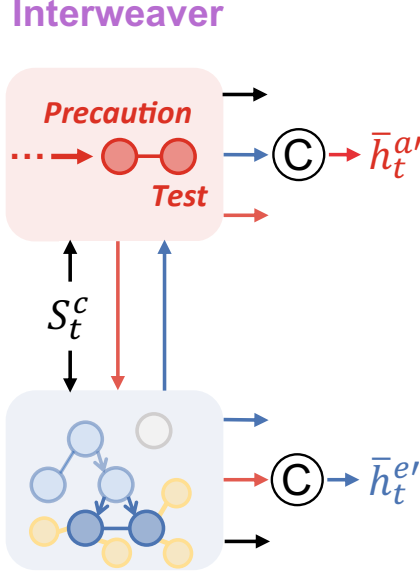


Figure 3.4: Detailed structure for flow interweaving. Each turn’s graph embedding attends to previous act embeddings, and each turn’s sequence embedding attends to previous entity embeddings.

transformation layer:

$$\text{Prob} = \text{sigmoid}(W_a S_t^a + b_a), \quad (3.7)$$

where $W_a \in \mathbb{R}^{n_a \times d}$ and $b_a \in \mathbb{R}^{n_a}$ are model parameters, and n_a denotes the number of candidate dialogue acts. The predicted dialogue acts $\hat{A}_t$ are obtained through an appropriate threshold.

Flow Interweaving. To achieve the integration of these two flows, we present an Interweaver to extend the entity graph embedding and act sequence embedding, as shown in Figure 3.4. This component incorporates the dialogue context into entity/act states and integrates entity/act sequential information from each other.

For the entity flow, we first fuse the historical context into the entity graph embedding.

The context-aware graph embedding of each turn is computed via cross-attention:

$$\alpha_{ki} = \text{softmax}\left(\frac{Q_k^c \top K_i^e}{\sqrt{d}}\right), \quad (3.8)$$

$$\bar{h}_k^{ec} = \sum_{i \in G_k} \alpha_{ki} V_i^e, k \in [1, t], \quad (3.9)$$

where $\bar{h}_k^{ec} \in \mathbb{R}^d$ is the context-aware graph embedding at the k -th turn, K_i^e and V_i^e are linear projected vectors based on the entity embedding h_i^e , and Q_k^c is based on the context state S_k^c . We denote this operation on S_k^c and $[h_i^e]_{i \in G_k}$ as $\text{CA}(S_k^c, [h_i^e]_{i \in G_k})$. Then, the act transition is fused into the graph embedding via CA as follows:

$$\bar{h}_k^{ea} = \text{CA}(\bar{h}_k^e, [h_j^a]_{j \in A_{\leq k}}), k \in [1, t], \quad (3.10)$$

where $\bar{h}_k^{ea} \in \mathbb{R}^d$ is the act-aware graph embedding, and $A_{\leq k}$ represents act sequences until the k -th turn. The final entity graph embedding is defined as the concatenation of three embeddings:

$$\bar{h}^{e'} = [\bar{h}^e; \bar{h}^{ec}; \bar{h}^{ea}], \quad (3.11)$$

where $\bar{h}_k^{e'} \in \mathbb{R}^{3d}$ is the extended graph embedding that incorporates historical dialogue context and act transition pattern.

For the act flow, we apply the CA operation to compute the context-aware and entity-aware act sequence embedding following the same way. The context-aware sequence embedding $\bar{h}_k^{ac}$ is based on S_k^c and $[h_j^a]_{j \in A_k}$, and the entity-aware sequence embedding $\bar{h}_k^{ae}$ is based on $\bar{h}_k^a$ and $[h_i^e]_{i \in G_{\leq k}}$. These two embeddings incorporate the historical dialogue context and entity transition pattern. Then, the final sequence embedding $\bar{h}^{a'} \in \mathbb{R}^{3d}$ is concatenated as follows:

$$\bar{h}^{a'} = [\bar{h}^a; \bar{h}^{ac}; \bar{h}^{ae}]. \quad (3.12)$$

We send the extended embeddings $\bar{h}^{e'}$ and $\bar{h}^{a'}$ instead of the pure embeddings to the above GRUs, allowing two flow models to acquire context information and lead each other.

Training Objective. The training of the dual flow modeling module includes two tasks, medical entity ranking and multi-act classification. The first one follows a contrastive learning [31] way with a negative log likelihood loss:

$$\mathcal{L}_e = - \sum_t \sum_{t^+} \log \frac{e^{\langle S_t^e, h_{t^+}^e \rangle}}{\sum_{i^- \in G_{\leq t}^1} e^{\langle S_t^e, h_{i^-}^e \rangle}}, \quad (3.13)$$

where $h_{t^+}^e$ is the embedding of a target entity mentioned in the t -th response, $\langle, \rangle$ is the dot product operation that calculates relevant scores (see Eq. 3.5), and $G_{\leq t}^1$ denotes the one-hop sub-graphs for the dialogue history until the t -th turn. We randomly select several entities instead of the whole sub-graph as negative entities. The second one is defined as a multi-label classification with a binary cross-entropy loss:

$$\mathcal{L}_a = \sum_t \sum_j^{n_a} \text{BCE}(\hat{y}_{t,j}^a, y_{t,j}^a), \quad (3.14)$$

where $\hat{y}_{t,j}^a$ is the probability of the dialogue act j for the t -th response (see Eq. 3.7), and $y_{t,j}^a$ is the ground-truth act label. The overall training objective of the dual flow modeling module can be calculated as:

$$\mathcal{L}_F = \lambda_e \mathcal{L}_e + \lambda_a \mathcal{L}_a, \quad (3.15)$$

where λ_e and λ_a are weights for each task.

3.3.2 Flow-guided Response Generation

After training the dual flow modeling module, we exploit the top- k relevant entities $\hat{E}^t$ (see Sec. 3.3.1) and predicted dialogue acts $\hat{A}^t$ (see Sec. 3.3.1) to guide response generation. We first encode the entity/act and dialogue history separately, allowing a relatively complete dialogue context. Then, these two types of information is merged into the decoder via a fusion component.

Act-Entity Fusion. As shown in Figure 3.5, the dialogue acts and entities are concatenated into one token sequence. We assign a unique token to each dialogue

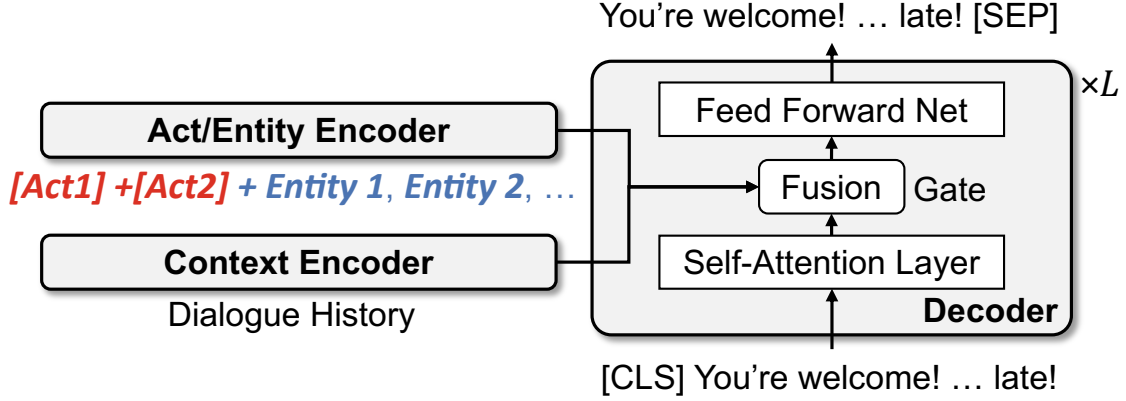


Figure 3.5: The detail of Response Generation.

act. Given the entity/act sequence and dialogue history sequence as input, the corresponding encoder produces final hidden states, H^{ea} and H^c . We fuse two types of information through a gate mechanism in the decoder:

$$h_w^{l,ea} = \text{CA}(h_w^l, H^{ea}), \quad (3.16)$$

$$h_w^{l,c} = \text{CA}(h_w^l, H^c), \quad (3.17)$$

$$g_w^l = \text{sigmoid}(W^l h_w^{l,c}), \quad (3.18)$$

$$h_w^{l'} = \text{FFN}(g_w^l h_w^{l,c} + (1 - g_w^l) h_w^{l,ea}), \quad (3.19)$$

where $h_w^{l'}$ is the final hidden state at the w -th token after fusion, h_w^l denotes the output of the l -th self-attention layer, FFN is the feed forward network, and W^l is the trainable parameter. We adopt the final hidden state $h_w^{L'}$ of the decoder to compute the probability distribution of the next token:

$$p(D_{t,w+1}) = \text{softmax}(W_d h_w^{L'} + b_d), \quad (3.20)$$

where W_d and b_d are linear mapping parameters.

Training and Inference. When training the response generation module, we use the selected top- k entities from the dual flow modeling module and ground truth

dialogue acts as encoder input. The training objective is defined as follows:

$$\mathcal{L}_G = - \sum_t^T \sum_{w=1}^w \log p(D_{t,w+1}), \quad (3.21)$$

where $p(D_{t,w+1})$ denotes the probability of the next token in the t -th response. Then in inference, we apply the top- k entities and predicted dialogue acts to generate the next response.

3.4 Experiments

3.4.1 Dataset

Our experiments are conducted on two medical dialogue datasets, MedDG [71] and KaMed [58]. The MedDG dataset contains 17K dialogues, focusing on 12 diseases in the gastroenterology department. The medical entities mentioned in the dialogues are annotated in MedDG. We split the dataset into 14862/1999/999 dialogues as the training/validation/test sets, following its original division. The KaMed dataset contains over 63K dialogues, covering diverse diseases in about 100 hospital departments. We filter out some dialogues with privacy concerns in KaMed and obtain 29,159/1,532/1,539 dialogues as the training/validation/test sets. The final data accounts for around 51% of the total.

3.4.2 Baseline models

We compare DFMED with five baseline models. (1) **Seq2Seq** [110] is an RNN-based sequence to sequence model with an attention mechanism. (2) **HRED** [98] is a hierarchical RNN that models a dialogue as a token sequence and an utterance sequence. (3) **GPT-2** [91] is a transformer decoder-based language model. (4) **BART** [56] is a transformer-based encoder-decoder model. (5) **VRBot** [58] is a medical

dialogue generation model with patient entity tracking and doctor entity learning.

For the experiments on the MedDG dataset, we extend Seq2Seq, HRED, GPT-2, and BART with entity hints, following [71]. Specifically, the extracted medical entities are appended at the end of the input sequence. In the following, these models with entity modeling are displayed with the **-Entity** suffix (e.g., Seq2seq-Entity).

3.4.3 Evaluation Metrics

Automatic Evaluation. We evaluate modules in the proposed framework separately. For the dual flow modeling module, the top-20 recall rate (R@20) of target entities and the weighted F1 score (Weighted-F1) of different dialogue acts considering the act imbalance are adopted. For the response generation module, BLEU [85] and ROUGE [64] scores in different n-grams (i.e., B-1, B-2, B-4, R-1, and R-2) are adopted to assess the response quality. Besides, we also measure the precision, recall, and F1 of entities (i.e., E-P, E-R, and E-F1) in responses to demonstrate the reliability following [71].

Human Evaluation. We chose 100 cases at random and invited three annotators to evaluate them manually. Results of the proposed framework are compared with different baseline models. We use three metrics to evaluate all of the generated responses based on past research [71, 58]: *sentence fluency* (FLU), *knowledge accuracy* (KC), and *entire quality* (EQ). On a 5-point Likert scale, from 1 (worst) to 5 (best), three annotators are asked to score these responses.

3.4.4 Implementation Details

For all baselines, we implement the open-source algorithm following [71] and [58]. We use the MedBERT¹ pretrained in the medical domain as the backbone of the

¹<https://github.com/trueto/medbert>

dual flow modeling module. To extract the medical entities in the dialogue history, we refer to the medical knowledge graph CMeKG² and extract the text spans that match the string of nodes in the graph. Entities with a one-hop connection to the historical entities are target ones for entity flow modeling. Besides, following [144], to identify the dialogue acts, we apply an open-source pseudo-labeling algorithm³ to automatically label each utterance.

We train the dual flow modeling module with the AdamW [72] optimizer. The learning rate and batch size are 4e-5 and 12 with 1000 warmup steps. The best loss weights λ_e and λ_a are 1 and 0.05 through grid searching. After training ten epochs, checkpoints with the highest average F1 for act prediction and the highest recall rate for top-20 entity selection on the validation set are selected. We select the threshold for each dialogue act, which achieves the best F1 score for the corresponding act on all validation samples. Then, we use Chinese pre-trained BART_{base}⁴ model with a six-layer encoder and a six-layer decoder for the response generation module. The entity/act encoder and context encoder share the same encoder. We adopt the AdamW optimizer and set the learning rate to 3e-5 with 2000 warmup steps. The model is trained for ten epochs with a batch size of 4. We implement all experiments on a single RTX 3090 GPU.

3.5 Results and Analysis

3.5.1 Details of Dialogue Acts

In this section, we will describe the details of our summarized dialogue acts. The seven dialogue acts can be divided into two types: (i) medical-related functions and

²<http://cmekg.pcl.ac.cn/>

³<https://github.com/yanguojun123/Medical-Dialogue>

⁴<https://huggingface.co/fnlp/bart-base-chinese>

Dialogue Acts	MedDG	KaMed
Inquire	25.49%	20.95%
Make a diagnosis	6.72%	8.64%
Prescribe medications	10.12%	13.74%
State a required medical test	4.25%	8.29%
Provide daily precautions	7.51%	5.29%
Inform	29.91%	30.04%
Chitchat	15.98%	13.04%

Table 3.1: The proportion of each dialogue act in the MedDG and KaMed datasets.

(ii) general open-domain dialogue acts. The specific meaning of each act is interpreted as follows:

Medical-related functions. (1) **Inquire.** The doctor asks questions about the history of the present illness (e.g., the location and duration of one symptom), previous surgery and medical conditions, current medications, allergies, etc. It corresponds to the “Subjective” function of the SOAP note [8]. (2) **Make a diagnosis.** The doctor makes a differential diagnosis based on historical dialogue context. It corresponds to the “Assessment” function of the SOAP note. (3) **Prescribe medications.** The doctor provides medication names and instructions. (4) **State a required medical test.** The doctor explains which tests are required and why each one was chosen to resolve diagnostic ambiguities; besides, what the next step would be if the results are positive or negative. (5) **Provide daily precautions.** The doctor explains the things that need to be paid attention to every day. Acts (3), (4), and (5) correspond to the “Plan” function of the SOAP note.

General open-domain dialogue acts. (6) **Inform.** The doctor tells the patient some information that he assumes to be correct. (7) **Chitchat.** The doctor expresses welcome, goodbye, apology, and thanks to the patient.

The proportion of each act in the MedDG and KaMed datasets is shown in Table 3.1. Examples for each dialogue act are listed as follows:

1. **Inquire:** “Hello, do you usually have diarrhea?”, “Have you taken any medicine before?”;
2. **Make a diagnosis:** “You may have gastroenteritis”, “You may have allergic rhinitis, which is easy to get sick this season.”;
3. **Prescribe medications:** “Please take Amoxicillin capsule 1.0g 2 times a day and Clarithromycin tablet 0.5g 2 times a day. Both are taken after meals.”;
4. **State a required medical test:** “If you often feel sick to your stomach, you can do a gastroscopy.”, “If your condition does not improve, I suggest you do a gastroscopy and Helicobacter pylori detection.”;
5. **Provide daily precautions:** “Please drink plenty of water, eat more fruits and vegetables. And try to have a morning poop.”;
6. **Inform:** “Migraine is a primary headache whose etiology and pathogenesis are not fully understood.”, “It can be effective in five days, and individual differences are relatively large.”;
7. **Chitchat:** “You’re welcome, and I wish you a speedy recovery!”, “Thank you so much”, “Hello!”;

	Methods	B-1	B-2	B-4	R-1	R-2	E-P	E-R	E-F1
w/o Pre-training	Seq2Seq	28.55	22.85	15.45	25.61	11.24	16.79	10.44	12.88
	Seq2Seq-Entity	29.13	23.22	15.66	25.79	11.42	23.79	15.89	19.06
	HRED	31.61	25.22	17.05	24.17	9.79	15.56	10.12	12.26
	HRED-Entity	32.84	26.12	17.63	24.26	9.76	21.75	15.33	17.98
	VRBot	29.69	23.90	16.34	24.69	11.23	18.67	9.72	12.78
w/ Pre-trained LM	GPT-2	35.27	28.19	19.16	28.74	13.61	18.29	14.45	16.14
	GPT-2-Entity	34.56	27.56	18.71	28.78	13.62	21.27	17.10	18.96
	BART	34.94	27.99	19.06	29.03	14.40	19.97	14.29	16.66
	BART-Entity	34.14	27.19	18.42	28.52	13.67	23.49	16.90	19.66
	DFMED	42.56[†]	33.34[†]	22.53[†]	29.31	14.21	22.48	22.84[†]	22.66[†]
	w/o Act Flow	36.79	29.18	19.81	29.45	14.26	22.73	21.70	22.20
	w/o Entity Flow	42.14	32.83	21.95	29.26	13.73	16.86	22.06	19.11
	w/o Interweaving	42.35	33.02	22.19	29.02	14.11	22.14	20.62	21.34

Table 3.2: Automatic evaluation results on MedDG. [†] denotes statistically significant differences ($p < 0.05$).

Methods	B-1	B-2	B-4	R-1	R-2
Seq2Seq	23.52	18.56	12.13	23.56	8.67
HRED	26.75	21.08	13.91	22.93	7.80
VRBot	30.04	23.76	16.36	18.71	7.28
GPT-2	33.76	26.58	17.82	26.80	10.56
BART	33.62	26.43	17.64	27.91	11.43
DFMED	40.20[†]	30.97[†]	20.76[†]	28.28[†]	11.54[†]
w/o Act Flow	35.47	28.11	18.78	27.97	11.45
w/o Entity Flow	39.14	29.92	19.73	27.17	10.47
w/o Interweaving	39.34	30.45	20.38	28.03	11.39

Table 3.3: Automatic evaluation results on KaMed. [†] denotes statistically significant differences compared to all baselines ($p < 0.05$).

3.5.2 Automatic Evaluation

The overall comparison of DFMED and other baseline models on the MedDG dataset is illustrated in Table 3.2, and the KaMed dataset is in Table 3.3. The observations from the comparison are as follows:

- (1) Our proposed framework DFMED outperforms these baseline models in most metrics. Specifically, on the MedDG dataset, it is 8.42%, 6.15%, 4.11%, and 0.79% higher than the best baseline model, BART-Entity, on B-1, B-2, B-4, and R-1. On the KaMed dataset, DFMED outperforms BART by 6.58%, 4.54%, 3.12%, and 0.37% on B-1, B-2, B-4, and R-1, indicating better similarity to the content of ground truth responses. Besides, on the MedDG dataset, DFMED exceeds BART-Entity by 3.00% on E-F1, meaning that it can generate responses with more accurate entity mentions. The above increases are because DFMED learns the transition of a medical entity flow and a dialogue act flow and predicts the entities and acts to guide response generation. Moreover, interweaving two flows enhances the prediction of entities and acts.
- (2) DFMED combines medical entity cues with dialogue act cues to steer the conversation. Compared to GPT-2 and BART, which have a performance drop on BLEU with the incorporation of medical entities (i.e., **-Entity**), DFMED shows increases in all these metrics. The main reason is that the force incorporation of entities may reduce response fluency. In DFMED, the dialogue act guidance influences attention to entities in the encoder and implicitly prevents the enforcement. The gate mechanism in the decoder controls the proportion of information from entities and acts. Besides, dialogue acts also provide essential content for responses without entity hints. This improvement is significant as about half of the responses in datasets do not match an entity in CMeKG via automatic string matching.

Methods	FLU	KC	EQ
BART	3.82	1.86	3.06
BART-Entity	3.85	2.03	3.35
DFMED	4.00	2.14	3.61
Gold	4.12	3.97	4.35

Table 3.4: Human evaluation results on MedDG.

Methods	MedDG			KaMed		
	Weighted-F1	R@20	B-4	Weighted-F1	R@20	B-4
DFMED	62.83	56.53	22.53	55.81	52.23	20.76
w/o Flow Modeling	62.09	54.74	22.11	55.16	51.31	20.21
w/o Interweaving	62.13	54.98	22.19	55.17	51.44	20.38
w/o Entity attends to Act	62.37	55.75	22.34	55.43	51.84	20.62
w/o Act attends to Entity	62.43	55.91	22.40	55.62	52.10	20.67

Table 3.5: Results of Dual Flow Modeling with ablation. The best results are in **bold**.

3.5.3 Human Evaluation

We select methods with high accuracy on the MedDG dataset to conduct a human evaluation, as shown in Table 3.4. Our framework displays an overall better response quality. Especially on the EQ, DFMED performs significantly better than baselines due to the incorporation of dialogue acts. The Cohen’s Kappa coefficient equals 0.53 and indicates a moderate agreement [23].

3.5.4 Analysis of Dual Flow Modeling

To further explore the effectiveness of our method, we investigate the following variants of DFMED: (1) **w/o Flow Modeling**, where we use a context state produced by mean pooling of all hidden states of dialogue context tokens to rank entities and predict dialogue acts. (2) **w/o Interweaving**, which removes the interweaving between entities and acts. (3) **w/o Entity attends to Act**, which removes the interweaving from act sequences. (4) **w/o Act attends to Entity**, which removes the interweaving from sequential entity graphs.

Table 3.5 shows the ablation study results. We observe drops in all metrics with the ablation variants, indicating the effectiveness of our proposed module. Specifically, the result of *w/o Flow Modeling* significantly drops on the entity recall rate and slightly drops on the overall act prediction F1 compared to the full model. It demonstrates that flow modeling can be necessary for learning the transition of entities and acts in dialogues. Besides, comparison among variants of the interweaver illustrates that incorporating sequential entity graphs and act sequences assists in the transition. Notably, the act sequence is more conducive to the transition of the entity and act flows. It may be because the entity graph introduce noise entities from patient utterances. These entities are more variable and can be deviated from the main content.

3.5.5 Case Study

A case generated by the above methods is illustrated in Table 3.6. Compared to baseline models, DFMED can produce responses more consistent with the dialogue context reflected by medical entities and dialogue acts. We can observe that the dual flow modeling module correctly predicts all acts and selects several medical entities, although two medications (e.g., “mosapride”) are different from ground truth mentions. Then, the beneficial guidance from medical entities and acts is employed to generate

Patient: Hello, doctor! I feel a little stomachache and suffocated. The stomachache is intermittent, sometimes it hurts, sometimes it doesn't. When I am not active, the pain is not obvious, but when I move, bloating makes me sick. I still have diarrhea. :(

Doctor: Hi, is it colic? Does stomachache have anything to do with eating?

Patient: No! It hurts before and after eating.

...

Patient: I've never had a gastroscopy.

Gold Response: You may have gastritis or stomach cramps and need to do a gastroscopy or a barium meal first! If diagnosed with the above diseases, you can take omeprazole, Kangfuxin liquid, and belladonna tablets following specific instructions.

BART: I suggest you try some omeprazole and hydrotalcite chewable tablets.

BART-Entity: It is recommended to have a gastroscopy to see if there is any history of gastritis or gastric ulcer.

Selected Entities: gastritis, gastroscopy, omeprazole,...;

Predicted Acts: [diagnosis], [prescription], [test].

Ours (DFMED): [You may have gastritis or gastroesophageal reflux.] [It is recommended to take omeprazole, mosapride, hydrotalcite.] [If the symptoms are not relieved, it is recommended to do a gastroscopy.]

Table 3.6: Case study. The responses are generated by different models, where key entities are underlined.

the next response. The response generation module fuses the entities and acts and produces the response containing these two hints.

3.6 Chapter Summary

In this chapter, we propose a dual flow enhanced medical dialogue generation framework, DFMED, that models a medical entity flow and a dialogue act flow to improve

the relevant entity selection and dialogue act prediction. Besides, we design an interweaver to strengthen interactions of two flows. The selected entities and predicted acts are applied to guide response generation. Experiments validate the effectiveness of our DFMED on two datasets.

Chapter 4

Developing Accurate Differential Diagnoses

In medical dialogues, diagnosis is a crucial process as the results of diagnosis establish an essential foundation for subsequent consultations [76, 104]. One of the most significant challenges in medical dialogue generation is producing responses grounded in a meticulous medical diagnosis, allowing doctors to understand the underlying diagnostic rationale. To address this challenge, we propose the IADDx framework, which generates a differential diagnosis using an intuitive-then-analytic approach and then leverages the inferred diseases to guide response generation. We also incorporate a disease knowledge graph to provide reliable external medical knowledge and improve the transparency of the diagnosis. Experimental results on two medical dialogue datasets demonstrate the effectiveness and interpretability of the IADDx framework.

4.1 Introduction

Medical dialogue systems (MDS) endeavor to provide diverse medical services such as diagnosis, treatment plans, and health consultations [144, 139, 18]. These systems have garnered increasing research attention in recent years due to their potential to assist doctors in diagnosing and prescribing [101, 71, 152, 160, 35, 141].

In medical dialogues, diagnosis is a crucial process as the results of diagnosis establish an essential foundation for subsequent consultations [76, 104]. However, previous studies on medical dialogue generation using pre-trained language models neglected to explicitly model the diagnostic process [139, 157, 58, 70, 65, 71]. One significant issue with these methods is that although responses benefiting from pre-trained models may appear coherent, they usually lack an interpretation grounded in meticulous medical diagnosis. It is challenging for doctors or patients to accept responses without a clear and interpretable diagnostic basis [52]. Particularly, the premise for doctors to refer to the system’s response is that they understand the diagnosis rationale behind the response and agree with it.

In practice, doctors typically employ both intuitive and analytic reasoning during the dialogue to construct a differential diagnosis, i.e., a set of potential diseases supported by available clinical findings, and generate responses based on the diagnosis [24, 112, 104]. Specifically, intuitive reasoning forms a rough disease list through a quick review of extensive clinical knowledge, while analytic reasoning cautiously verifies some diseases via a systematic analysis of body systems, organs, and symptoms. As an example shown in Figure 4.1, the doctor first utilizes two reasoning processes to arrive at a differential diagnosis of cholecystitis and gastritis. Then, he organizes a response to review the medical test results and provide clarification regarding the potential diagnoses. Prior studies often overlook the importance of differential diagnosis [58, 157, 141, 127, 48]. In our work, we argue that modeling a differential diagnosis with intuitive and analytic reasoning is crucial, and generating responses

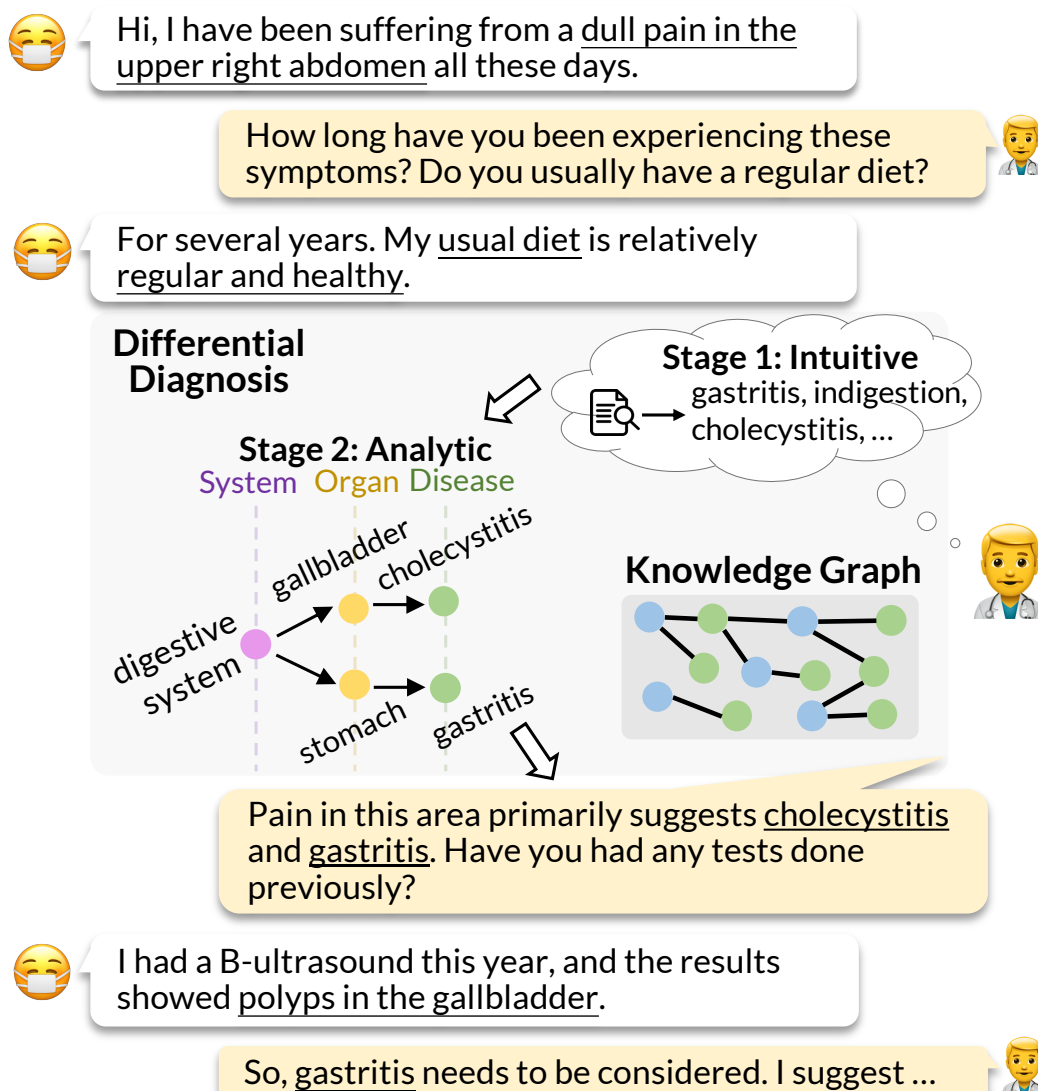


Figure 4.1: An example of differential diagnosis in a medical dialogue, which contains intuitive and analytic reasoning.

conditioned on potential diseases can improve their reliability and accuracy.

To address the above issues, we propose a medical dialogue generation framework with **Intuitive-then-Analytic Differential Diagnosis (IADDx)**, which first produces a differential diagnosis and then utilizes potential diseases to guide response generation.

We draw inspiration from the diagnostic reasoning research [24] and design a two-stage (i.e., intuitive-then-analytical) differential diagnosis method. In the intuitive stage, a preliminary disease list is retrieved based on patient cases and disease documents. Then in the analytical stage, the preliminary list is refined through a multi-disease classifier with multi-task learning. A disease knowledge graph is also incorporated to provide reliable external information for refinement. Finally, the cases and documents relevant to the refined potential diseases are utilized to enhance response generation.

Our main contributions are summarized as follows:

1. We propose a medical dialogue generation framework, IADDx, which explicitly models a differential diagnosis with intuitive-then-analytic processes and incorporates diagnosis to guide response generation.
2. We build a multi-disease classifier with multi-task learning of affected systems and organs. A disease knowledge graph composed of systems, organs, diseases, and symptoms is adopted to interpret the diagnostic process.
3. Experimental results on two medical dialogue datasets show the effectiveness and interpretability of our IADDx.

4.2 Preliminary

4.2.1 Problem Formulation

A medical dialogue is denoted as $U = \{(U_k^P, U_k^D)\}_{k=1}^T$, where utterances from patients and doctors are represented by U^P and U^D respectively. Each dialogue is annotated with several possible diseases $D = \{d_i\}_{i=1}^{n_d}$, and each doctor utterance is annotated with multiple dialogue acts $A = \{a_i\}_{i=1}^{n_a}$. Given the historical dialogue sequence $U_t = \{U_1^P, U_1^D, \dots, U_t^P\}$, the objective is to produce the t -th doctor utterance U_t^D .

Prompt for SOAP Note extraction

Your task is to precisely identify and extract phrases that correspond to different parts of the SOAP Notes.

The SOAP Notes are structured as follows:

'S' stands for 'subjective' patient information, which includes symptoms, conditions, injury, poisoning, previous diagnoses, patient's age, sex, history, current medications, allergies, and personal views or feelings.

'O' stands for 'objective' data gathered from the patient encounter.

'A' stands for 'assessment' by doctors, i.e., differential diagnosis, which includes possible diseases or problems.

'P' stands for 'plan' by doctors, i.e., what to do next, which details the need for additional testing, medication, other treatments (like surgery, physical therapy, or massage), and precautions (like eat, drink, exercise, etc.).

Please note the following guidelines:

1. Your focus should be on the phrase affirmed by the patients (e.g., 'doctor: Any other discomforts? Do you experience acid reflux or heartburn? patient: Sometimes I feel heartburn and a bloated stomach.' should be translated to 'Subjective: heartburn, bloated stomach').

2. All numbers should be represented in Arabic numerals (for example, 'fifteen years old' should be translated to '15 years old').

3. Phrases should be semantically clear (e.g., 'Had' should be translated to 'Had a gastroscopy examination.' by concatenating phrases from the previous utterance).

4. If no phrase is extracted for one part of the notes, 'null' should be the output (e.g., 'Subjective: null').

5. Phrases denoting negative or unknown should also be extracted (e.g., 'No, cannot, I don't know').

6. 'Assessment' and 'Plan' are generally extracted from 'doctor: XXX', while 'Subjective' and 'Objective' are generally extracted from 'patient: XXX'.

Some examples including conversation history and extracted phrases are as follows:

""

doctor: XX

patient: XX

""

The phrases are as follows:

Subjective: XX, XX

Objective: null

Assessment: null

Plan: null

Then, please extract the phrases in the following conversation:

Please STRICTLY follow the format of the output.

The new conversation history is as follows:

""

doctor: XX

patient: XX

""

The phrases are as follows:

Figure 4.2: The prompt for extracting segments of the SOAP Notes.

4.2.2 SOAP-based Content Structuring

In each round of a medical dialogue (U^D, U^P) , we first extract medical-related segments in accordance with the SOAP Notes [51], i.e., **S**ubjective personal reports $\{S_i^S\}$, **O**bjective quantifiable data $\{S_i^O\}$, an **A**ssessment (or diagnosis results) $\{S_i^A\}$, and subsequent **P**lans $\{S_i^P\}$ for patient care. SOAP notes serve as a documentation method used by medical specialists to structure patient information. For example, a subjective segment can be “stomach cramps” and an objective segment can be “body temperature is 37.8 degrees Celsius”. The segment extraction helps filter out irrelevant details, thereby improving the accuracy of subsequent retrieval and diagnosis. We utilize a

proprietary Large Language Model (LLM), GPT-4o [82], to extract medical-related segments. The input prompt for GPT-4o contains SOAP descriptions, some extraction examples, and the t -th doctor and patient utterances. The prompt template of the extraction is shown in Figure 4.2. We use approximately 10 human-annotated segment extraction examples to guide GPT-4o in performing precise few-shot extraction.

4.2.3 Disease Knowledge Graph

We employ a disease knowledge graph to assist in differential diagnosis, which is inspired by the *problem-specific framework* [108]. This framework helps analyze medical conditions to guide the formulation of a differential diagnosis. We construct the knowledge graph based on a disease corpus from an online medical encyclopedia website¹, named “xiaohe”. The corpus includes knowledge documents covering around 10,000 diseases, which have been edited and reviewed by medical specialists. For documents of each disease, we use an open-source relation extraction model² to obtain triplets of heads, relations, and tails. The knowledge graph we construct encompasses entities such as symptoms, pathogens, medications, and medical examinations. These entities provide the necessary external knowledge for the differential diagnosis.

4.3 Method

Our proposed IADDx framework comprises two main components, as illustrated in Figure 4.3. First, the Intuitive-then-Analytic Differential Diagnosis component (Sec. 4.3.1) conducts a differential diagnosis using both intuitive and analytic reasoning, generating a list of potential diseases. Subsequently, the Diagnosis-guided Response Generation component (Sec. 4.3.2) draws on the list of potential diseases, relevant

¹<https://www.xiaohe.cn/medical/home>

²https://github.com/king-yyf/CMeKG_tools

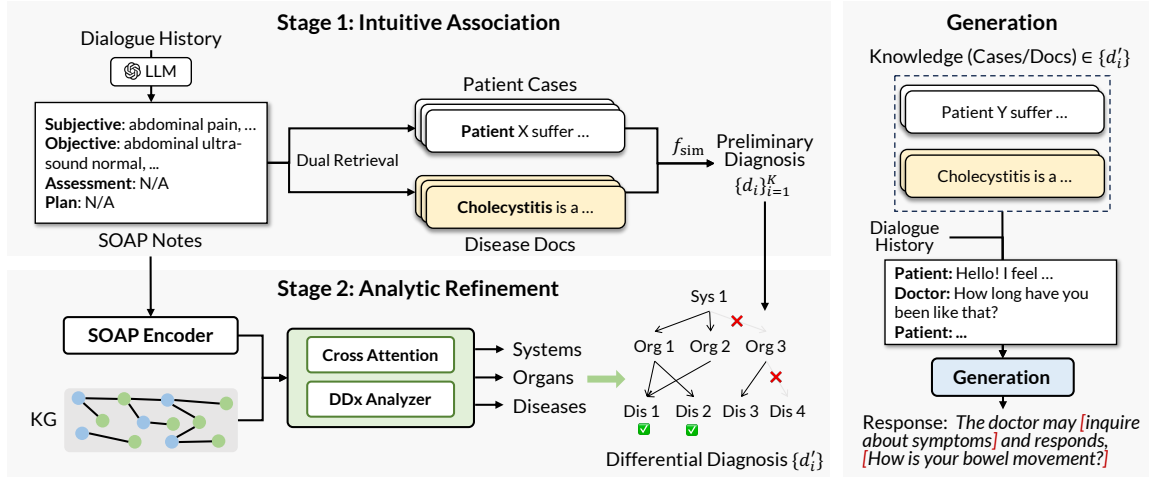


Figure 4.3: **Left:** The architecture of Differential Diagnosis, which includes the Intuitive Association stage and the Analytic Refinement stage. The DDx Analyzer in Stage 2 predicts the affected systems and organs and refines the Preliminary Diagnosis to generate a Differential Diagnosis. **Right:** The structure of Response Generation. The generation model is fine-tuned and instructed to output dialogue acts (i.e., “inquire about symptoms”) and the corresponding responses.

patient cases, and knowledge documents to generate dialogue acts and the appropriate response.

4.3.1 Intuitive-then-Analytic Differential Diagnosis

To enhance the interpretability of generated responses and make them easier for doctors and patients to understand, we constructed a differential diagnosis component prior to generating the responses. Compared with previous studies, an additional diagnostic component helps the response focus on candidate diseases. We model the differential diagnosis process in two stages: **Intuitive Association** and **Analytic Refinement**. The first stage draws upon clinical experience to make a rough diagnosis. We use dual retrieval of patient cases and disease documents to produce a preliminary list of

diseases. The disease document retriever helps identify diseases that have not been encountered before. Then, the second stage further refines the diagnosis through more detailed problem analysis, enhancing both diagnostic accuracy and interpretability. We construct a multi-disease classifier using our **Disease Knowledge Graph** (DKG) to finalize the differential diagnosis. The detailed process is shown in the left of the Figure 4.3.

Stage 1: Intuitive Association.

The objective of this stage is to make a rough diagnosis. Given the dialogue history U_t , we extract SOAP segments and use them to retrieve relevant patient cases and disease documents. These segments contain information (e.g., patient symptoms and medical history) needed for differential diagnosis. We concatenate the segments $\{S^S\}$, $\{S^O\}$, and $\{S^A\}$ as the query for subsequent dual retrieval. Here, the plan segments are not considered, as their content is not directly relevant to the diagnostic process. We use the aforementioned query to retrieve two types of knowledge: (1) patient cases (in the form of a series of SOAP segments) from the training corpus that exhibit conditions similar to those of the current patient, and (2) disease documents (e.g., clinical findings and diagnostic methods) from a disease corpus, with descriptions that align with the current patient’s conditions. The query and knowledge (patient cases or disease documents) sequences with a “[CLS]” token inserted at the front are separately input to a BERT encoder [25]. We select the last hidden states of “[CLS]” tokens as their representations, $\mathbf{S}^q \in \mathbb{R}^d$, $\mathbf{S}^{\text{Case}} \in \mathbb{R}^d$, and $\mathbf{S}^{\text{Doc}} \in \mathbb{R}^d$ and calculate relevance scores for patient cases and disease documents in relation to the current patient’s condition as follows:

$$s^{\text{Case}} = \langle \mathbf{S}^q, \mathbf{S}^{\text{Case}} \rangle, \quad (4.1)$$

$$s^{\text{Doc}} = \langle \mathbf{S}^q, \mathbf{S}^{\text{Doc}} \rangle, \quad (4.2)$$

where $\langle, \rangle$ represents a similarity function based on the dot product. After retrieving patient cases, each case, which is annotated with one or more diseases, is assigned a relevance score. This score is then applied to the annotated diseases. Since different patient cases can be annotated with the same disease but assigned different scores, each disease can be associated with a set of scores $\{s_{ij}^{\text{Case}}\}_{j=1}^{n_i}$ from various patient cases. We select the maximum of them as the relevance score:

$$s_i^{\text{Case}} = \max(\{s_{ij}^{\text{Case}}\}_{j=1}^{n_i}), \quad (4.3)$$

where n_i denotes the number of cases associated with disease i . After retrieving disease documents, each disease can also be assigned a relevance score s_i^{Doc} , based on the corresponding documents. We normalize both types of scores for all diseases and then sum the two normalized scores to obtain the final disease relevance score:

$$s_i = s_i^{\text{Case}} + s_i^{\text{Doc}}. \quad (4.4)$$

The top- K diseases are selected as a **Preliminary Diagnosis**, denoted as $\{d_i\}_{i=1}^K$, where d_i represents the i -th disease.

We adopt a contrastive loss [17] to optimize the two retrievers, ensuring that the representation of the query and the representations of relevant patient cases or disease documents are brought closer to each other in the embedding space. The loss functions for the patient case and disease document retrievers are defined as:

$$\mathcal{L}_{\text{Case}} = -\log \frac{\exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Case}+} \rangle)}{\exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Case}+} \rangle) + \sum_{\mathbf{S}^{\text{Case}-} \in \mathcal{B}} \exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Case}-} \rangle)}, \quad (4.5)$$

where $\mathbf{S}^{\text{Case}+}$ denotes representations from relevant cases that share at least one diagnosed disease with the current dialogue, and $\mathbf{S}^{\text{Case}-}$ denotes representations from irrelevant cases that are unrelated to the currently discussed diseases. We use irrelevant cases from input batches $\mathcal{B}$ for training. The loss function $\mathcal{L}_{\text{Doc}}$ for the disease document retriever is defined in the same way as:

$$\mathcal{L}_{\text{Doc}} = -\log \frac{\exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Doc}+} \rangle)}{\exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Doc}+} \rangle) + \sum_{\mathbf{S}^{\text{Doc}-} \in \mathcal{B}} \exp(\langle \mathbf{S}^q, \mathbf{S}^{\text{Doc}-} \rangle)}, \quad (4.6)$$

where $\mathbf{S}^{\text{Doc}+}$ denotes representations from documents of ground truth diseases, and $\mathbf{S}^{\text{Doc}-}$ denotes representations from unrelated disease documents.

Stage 2: Analytic Refinement.

This stage further refines the diagnosis via problem-specific analysis, improving diagnostic accuracy and interpretability. We construct a multi-disease classifier trained with multi-task learning to jointly classify affected systems and organs, and to leverage the relationships between systems, organs, and diseases to improve differential diagnosis. The disease knowledge graph is employed to augment the classification.

As depicted in the bottom left of Figure 4.3, we first feed the SOAP Notes from one patient case to a SOAP encoder. Specifically, the SOAP segments $\{S^S\}$, $\{S^O\}$, and $\{S^A\}$ are concatenated after prepending the “[CLS]” token, and fed into a BERT-based SOAP encoder. The last hidden state of the “[CLS]” token is used as the representation of the SOAP Notes, denoted as $\mathbf{S} \in \mathbb{R}^d$. Second, we incorporate entities involved in one-hop relationships with diseases identified in the preliminary diagnosis list and construct the graph G . For all entities in the graph, we use the same encoder used for SOAP Notes to generate their embeddings. Each raw entity embedding $\mathbf{e}^0 \in \mathbb{R}^d$ is calculated by averaging the token embeddings for that entity. We employ Graph Attention Network (GAT) [120] to merge neighboring information as:

$$\alpha_{j\omega}^k = \frac{\exp(\sigma_1(\mathbf{a}^\top [\mathbf{W}^k \mathbf{e}_j^0 \| \mathbf{W}^k \mathbf{e}_\omega^0]))}{\sum_{\mu \in \mathcal{N}_j} \exp(\sigma_2(\mathbf{a}^\top [\mathbf{W}^k \mathbf{e}_j^0 \| \mathbf{W}^k \mathbf{e}_\mu^0]))}, \quad (4.7)$$

$$\mathbf{e}_j = \left[\sigma_2 \left(\sum_{\omega \in \mathcal{N}_j} \alpha_{j\omega}^k \mathbf{W}^k \mathbf{e}_\omega^0 \right) \right]_{k=1}^h, \quad (4.8)$$

where $\mathbf{e}_j \in \mathbb{R}^d$ represents the updated embedding, $\mathbf{a} \in \mathbb{R}^{2d}$ and $\mathbf{W}^k \in \mathbb{R}^{d_h \times d}$ are learnable parameters, σ_1 and σ_2 denote activation function, $\mathcal{N}_j$ is a set of neighboring entities that connect to entity j , h is the number of heads, and d_h is the dimension for each head.

We implement a cross-attention layer to enable interaction between the SOAP representation and the sub-graphs relevant to candidate diseases. The SOAP representation is fused separately with the entity embeddings of each disease-specific sub-graph, resulting in multiple disease-specific representations. These fused representations are then passed through a linear layer followed by a sigmoid layer to produce probabilities for each disease. The cross-attention layer is defined as follows:

$$\mathbf{A}_i = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_i^\top}{\sqrt{d}}\right), \quad (4.9)$$

$$p_i = \text{sigmoid}(\text{MLP}(\mathbf{A}_i\mathbf{V}_i\mathbf{O})), i = 1, \dots, K, \quad (4.10)$$

where $\mathbf{K}_i \in \mathbb{R}^{n_{G_i} \times d}$ and $\mathbf{V}_i \in \mathbb{R}^{n_{G_i} \times d}$ are the linearly projected representations of the concatenated entity embeddings $\{\mathbf{e}_j\}_{j \in G_i}$, $\mathbf{Q} \in \mathbb{R}^d$ is the linearly projected representation of the SOAP representation $\mathbf{S}$, $\mathbf{O} \in \mathbb{R}^{d \times d}$ is an output projection matrix, and $\text{MLP}(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}^n$ is a classification layer that outputs predictions for all diseases in the corpus (Sec. 4.2.3). G_i denotes the sub-graph relevant to the i -th disease in the preliminary diagnosis list, n_{G_i} denotes the number of entities in the sub-graph, and n denotes the number of total diseases. For the i -th disease, we select its corresponding probability from p_i . The selected probabilities for all K diseases are then combined and filtered using an appropriate threshold to obtain the refined differential diagnosis set $\{d'_i\}$.

To further enhance the differential diagnosis, we develop additional system and organ classification modules to augment disease prediction. These two classification modules are described as follows:

$$p^{\text{System}} = \text{sigmoid}(\text{MLP}^{\text{System}}(\mathbf{S})), \quad (4.11)$$

$$p^{\text{Organ}} = \text{sigmoid}(\text{MLP}^{\text{Organ}}(\mathbf{S})). \quad (4.12)$$

A disease is selected only if the corresponding probabilities for the systems and organs associated with that disease are greater than a preset threshold. Then, the final **Differential Diagnoses** are obtained after applying the above filtering.

We optimize the multi-disease classification by minimizing a binary cross-entropy loss. The loss function $\mathcal{L}_d$ is calculated as follows:

$$\mathcal{L}_d = - \sum_i [y_i^d \cdot \log p_i + (1 - y_i^d) \cdot \log(1 - p_i)], \quad (4.13)$$

where y_i^d is the label of i -th disease in the preliminary diagnosis list. Furthermore, we implement system classification and organ classification losses for multi-task learning, denoted as $\mathcal{L}_s$ and $\mathcal{L}_o$. These two losses are also binary cross-entropy losses. The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_d + \lambda_s \mathcal{L}_s + \lambda_o \mathcal{L}_o, \quad (4.14)$$

where λ_s and λ_o are hyperparameters that control the trade-off between the learning of each task.

4.3.2 Diagnosis-guided Response Generation

In this section, we utilize the refined diagnosis to guide medical knowledge retrieval, thereby enhancing response generation with the aid of retrieved disease-related knowledge. Besides, a doctor dialogue act predictor is introduced to guide the flow of the dialogue [139]. The detailed architecture is shown on the right of Figure 4.3.

Dialogue Act Prediction. We develop a multi-act predictor to assist knowledge retrieval and manage the dialogue flow. The input of the predictor is SOAP segments and dialogue history. We add “[CLS]” token at the front of the segment sequence and encode this sequence using a BERT encoder, which is also applied to the dialogue history. The final hidden states of “[CLS]” token in these two encodings are selected as the representation of the segments $\mathbf{H}_t^s \in \mathbb{R}^d$ and dialogue history $\mathbf{H}_t \in \mathbb{R}^d$. We concatenate two representations to predict dialogue acts as follows:

$$p_i^a = \text{sigmoid}(\mathbf{W}^a[\mathbf{H}_t^s; \mathbf{H}_t]), i = 1, \dots, m, \quad (4.15)$$

where $\mathbf{W}^a \in \mathbb{R}^{m \times 2d}$ is a trainable parameter matrix, $[\cdot]$ denotes a concatenation operation, and m is the number of candidate acts. The predicted dialogue acts $\{a_i\}_{i=1}^{n_a}$ are selected through an appropriate threshold (See in Experiments). We also apply a binary cross-entropy loss to optimize the multi-act prediction.

Disease-guided Retrieval. To augment response generation, we retrieve disease-related passages based on the refined differential diagnosis and predicted dialogue acts. These passages are from the medical encyclopedia website, providing external knowledge for response generation. For dialogue acts directly related to a specific aspect of medicine, we choose corresponding passages without requiring retrieval. For example, responses with the act “Inquire about present illness” are closely related to the passage describing clinical manifestations of one disease. For non-medical dialogue acts, we use the current dialogue history to retrieve top relevant passages from a disease corpus through the BM25 algorithm [97].

Response Generation. After retrieving medical knowledge and predicting dialogue acts, we incorporate these two pieces of information to guide response generation. We construct a generation model following the Fusion-in-Decoder (FiD) method [42, 103], which allows the decoder to attend to all encoding representations at the same time when generating a response. The input of the model contains a group of sequences, where the first sequence is the dialogue history sequence U_t concatenated with unique tokens denoting dialogue acts $\{a_i\}$, and other sequences are retrieved knowledge sequences Z_t . Each sequence is concatenated with a special token (“[U]” or “[K]”) at the front to represent the dialogue history or knowledge. Compared with concatenating the dialogue history with knowledge as an input sequence, the above input can leverage longer dialogue history.

We train the generation model by a negative log-likelihood loss. The loss function is

defined as:

$$\mathcal{L}_g = - \sum_{i=1}^N \log p(U_{t,i}^D), \quad (4.16)$$

where $U_{t,i}^D$ is the i -th token’s probability. During training, we use ground truth acts as part of the input and apply ground truth diseases to guide knowledge retrieval. Then in inference, we apply predicted acts and diseases as guidance.

4.4 Experiments

Methods		B-1	B-2	B-4	R-1	R-2	E-P	E-R	E-F1
w/o Pre-training	Seq2Seq	28.55	22.85	15.45	25.61	11.24	16.79	10.44	12.88
	Seq2Seq-Entity	29.13	23.22	15.66	25.79	11.42	23.79	15.89	19.06
	HRED	31.61	25.22	17.05	24.17	9.79	15.56	10.12	12.26
	HRED-Entity	32.84	26.12	17.63	24.26	9.76	21.75	15.33	17.98
	VRBot	29.69	23.90	16.34	24.69	11.23	18.67	9.72	12.78
w/ Pre-trained LM	GPT-2	35.27	28.19	19.16	28.74	13.61	18.29	14.45	16.14
	GPT-2-Entity	34.56	27.56	18.71	28.78	13.62	21.27	17.10	18.96
	BART	34.94	27.99	19.06	29.03	14.40	19.97	14.29	16.66
	BART-Entity	34.14	27.19	18.42	28.52	13.67	23.49	16.90	19.66
	DFMed	42.56	33.34	22.53	29.31	14.21	22.48	22.84	22.66
	IADDx (Ours)	43.17[†]	34.09[†]	23.33[†]	29.60[†]	14.37	21.81	22.90[†]	22.34

Table 4.1: Automatic evaluation results on MedDG. [†] denotes statistically significant differences ($p = 0.05$).

4.4.1 Datasets

We perform our experiments using two medical dialogue datasets: **MedDG** [71] and **KaMed** [58]. MedDG encompasses 17K dialogues, primarily centered on 12 diseases within the gastroenterology department. We partition the dataset into 14862, 1999,

and 999 dialogues for training, validation, and testing. KaMed offers a comprehensive collection of over 63K dialogues spanning nearly 100 hospital departments. We follow [139] to divide datasets into 29159, 1532, and 1539 dialogues for experiments.

4.4.2 Baseline Models

We evaluate IADDx with six baseline models. **Non-Pretrained models:** (1) **Seq2Seq** [110] employs an RNN enhanced by an attention layer. (2) **HRED** [98] leverages a multi-level RNN design. (3) **VRBot** [58] is designed for medical dialogue generation, emphasizing the tracking and predicting of entities. **Pretrained models:** (1) **GPT-2** [91] is a transformer decoder-based language model. (2) **BART** [56] is a transformer-based encoder-decoder model. (3) **DFMed** [139] is a medical dialogue generation model that learns entity and dialogue act flows. For our experiments on the MedDG dataset, we supplement model inputs with predicted entities as described by [71].

4.4.3 Evaluation Metrics

Automatic Evaluation. BLEU [85] and ROUGE [64] scores across varied n -grams (specifically, **B-1/2/4**, **R-1/2**) are utilized as metrics to evaluate the quality of generated responses. In alignment with [71], the precision, recall, and F1 of entities mentioned in the responses are calculated, denoted as **E-P/R/F**. Besides, the diagnosis accuracy is evaluated by a disease F1 score, denoted as **D-F1**.

Human Evaluation. We randomly selected 100 cases and engaged three physicians for manual evaluation. The performance of our IADDx is compared with various baseline models. Drawing upon previous studies [71, 58], we assess the generated responses using three metrics: sentence fluency (FLU), knowledge accuracy (KC),

Methods	B-1	B-2	B-4	R-1	R-2
Seq2Seq	23.52	18.56	12.13	23.56	8.67
HRED	26.75	21.08	13.91	22.93	7.80
VRBot	30.04	23.76	16.36	18.71	7.28
GPT-2	33.76	26.58	17.82	26.80	10.56
BART	33.62	26.43	17.64	27.91	11.43
DFMed	40.20	30.97	20.76	28.28	11.54
IADDx (Ours)	40.98[†]	31.69[†]	21.35[†]	28.31[†]	11.67[†]

Table 4.2: Automatic evaluation results on KaMed. [†] denotes statistically significant differences ($p = 0.05$).

and overall quality (EQ). Each metric is on a 5-point Likert scale (from poorest 1 to excellent 5).

4.4.4 Implementation Details

We apply the MedBERT³, a BERT-base model pre-trained on Chinese medical documents as the backbone for encoders in the differential diagnosis component. The disease corpus used for the intuitive association is extracted from an online medical encyclopedia website named *xiaohe*⁴, which contains medical knowledge certified by specialists. In the intuitive association stage, we selected the top 50 diseases as the preliminary diagnosis results. In the analytic refinement stage, we predict multiple diseases from the list. We set the weights for $\mathcal{L}_d$ and $\mathcal{L}_{expl}$ to 1 and 0.5. The threshold for predicting multiple diseases is set to 0.8 based on the disease F1 scores on the validation dataset. In the response generation component, we retrieve knowledge

³<https://github.com/trueto/medbert>

⁴<https://www.xiaohe.cn/medical>

from the same corpus and select the top 5 passages for generating responses. The dialogue act predictor is also based on the MedBERT. We choose from 10 acts and adopt different thresholds to get the highest act F1 scores for each act. The generator is a pre-trained BART-base model⁵ with a six-layer encoder and a six-layer decoder. We adopt the AdamW optimizer [72] to train the above models and implement all experiments on a single RTX 3090 GPU.

The training details for each model in our framework are described as follows:

For training the patient case retriever in the intuitive association stage, we adopt a learning rate of $1e-5$ and a batch size of 8. We set the gradient accumulation step to 6, employ a 1000-step warm-up, and complete the training in 1 epoch. For training the disease document retriever in the intuitive association stage, we adopt a learning rate of $7e-5$ and a batch size of 8. We set the gradient accumulation step to 6, employ a 1000-step warm-up, and complete the training in 25 epochs.

For training the multi-disease classifier in the analytic refinement stage, we adopt a learning rate of $2e-5$ and a batch size of 12. We set the gradient accumulation step to 2, employ a 1000-step warm-up, and complete the training in 10 epochs.

For training the multi-act predictor in the response generation component, we adopt a learning rate of $1e-5$ and a batch size of 12. We set the gradient accumulation step to 2, employ a 1000-step warm-up, and complete the training in 5 epochs. The threshold for each dialogue act is as follows: Inquire about Present Illness: 0.3; Inquire about History: 0.2; Make a Diagnosis: 0.2; Prescribe Medications: 0.2; Recommend a Laboratory Test: 0.2; Recommend a Treatment Plan: 0.2; Provide Daily Precautions: 0.2; Inform: 0.3; Chitchat: 0.5; Other: 0.05;

For training the response generator, we adopt a learning rate of $4e-5$ and a batch size of 1. We set the gradient accumulation step to 2, employ a 1000-step warm-up, and complete the training in 5 epochs.

⁵<https://huggingface.co/fnlp/bart-base-chinese>

Methods	FLU	KC	EQ
BART	3.77	1.87	3.12
BART-Entity	3.79	2.05	3.41
DFMed	3.91	2.26	3.59
IADDx (Ours)	4.08	2.41	3.83

Table 4.3: Human evaluation results on MedDG.

4.4.5 Evaluation of Dialogue Generation

Automatic Evaluation. The dialogue generation results for the MedDG dataset are presented in Table 4.1, and results for the KaMed dataset can be found in Table 4.2. We observe that our IADDx method outperforms baseline models on most evaluation metrics. In detail, on the MedDG dataset, IADDx surpasses the state-of-the-art method DFMed by 0.61, 0.75, and 0.8 in B-1, B-2, and B-4 metrics, as well as 0.29 and 0.16 in R-1 and R-2 metrics. Additionally, we achieve comparable entity accuracy even without employing entity flow learning like DFMed. The reason is that we use multiple differential diagnoses to retrieve pertinent knowledge and subsequently leverage them to enhance response generation. Similar advantages are evident in the experimental results on the KaMed dataset. IADDx outperforms DFMed by 0.78, 0.72, and 0.59 in B-1, B-2, and B-4 metrics. It indicates that modeling and incorporating differential diagnosis aid in generating coherent, informative, and accurate responses.

Human Evaluation. Table 4.3 shows the results of human evaluation on MedDG. Our method consistently surpasses all baselines on every reported metric. This suggests that by explicitly modeling the differential diagnosis and using it to guide response generation, we can produce more informative and accurate responses. The Fleiss’ kappa [29] score is 0.49, indicating a moderate level of inter-annotator agreement. We

asked three registered physicians to assess the generated responses. The evaluation includes sentence fluency (FLU), knowledge accuracy (KC), and overall quality (EQ). Sentence fluency evaluates whether a sentence includes repetition, truncated phrases, and typos. Knowledge accuracy evaluates the accuracy of entities (e.g., diseases, medicines, tests, symptoms, and departments) contained in the response. Overall quality assesses whether the response is contextually appropriate. For example, it assesses whether the facts described are correct and answer the patient’s questions.

4.4.6 Analysis of Intuitive-then-Analytic Differential Diagnosis

To delve deeper into the efficacy of our approach, we examine multiple variants of our IADDx method as follows: (1) **w/o DDx**, which removes the entire differential diagnosis component and generates responses conditioned solely on the dialogue history and dialogue acts. (2) **w/o Analytic**, which removes the analytic refinement on the preliminary diagnosis and employs the top 5 ranked diseases to guide knowledge retrieval. (3) **w/o DKG**, which removes the diagnosis-oriented graph in the multi-disease classifier and solely adopts the mean representations of SOAP segments and a subsequent linear layer to classify diseases.

Table 4.4 displays the overall ablation results. We see a reduction in performance across all metrics with the ablation variants, emphasizing the critical role of each module in our proposed method. Among these variants, the response quality of *w/o* DDx notably decreases due to the absence of diagnosis-related knowledge. Such knowledge is crucial for providing essential disease information that enhances the informativeness and accuracy of the response. Besides, the differential diagnosis performance of *w/o* Analytic significantly drops compared with the complete DDx. It is because the retrieval-based intuitive association provides a fixed number of potential diseases and cannot ensure that all relevant diseases rank high, inadvertently introducing unrelated diseases. The results of *w/o* DKG demonstrate that the diagnosis-oriented graph

Datasets	Methods	D-F1	B-1	B-4	R-2
MedDG	IADDx	43.50	43.17	23.33	14.37
	w/o DDx	-	42.21	22.08	13.78
	w/o Analytic	37.01	43.02	22.98	14.09
	w/o Graph	42.61	43.13	23.22	14.23
KaMed	IADDx	50.23	40.98	21.35	11.67
	w/o DDx	-	39.03	19.54	10.22
	w/o Analytic	40.26	40.74	21.16	11.44
	w/o Graph	49.52	40.86	21.22	11.52

Table 4.4: Ablation results on two datasets.

plays a role in the differential diagnosis and subsequent diagnosis-guided response generation. Additionally, we examined how the top-K setting affects subsequent response generation. A larger K introduces more noisy disease candidates, though the analytic-refinement step filters out some of this noise. Conversely, a smaller K can miss plausible diseases, causing the response to overlook them. An appropriate K lies between 20 and 30.

4.4.7 Case Study

We present a case study of medical dialogue generation by different models in Table 4.5. We observe that the response generated by our IADDx method offers more accurate diagnoses, along with treatment plans tailored to those diagnoses, compared to responses from the three baseline models. The prescribed medicines are suitable for treating the diagnosed disease GERD. This suggests that IADDx performs an effective differential diagnosis and leverages it to enhance response generation. Besides, although four diseases are included in the differential diagnosis, IADDx can implicitly select the most likely disease to generate a response.

Interpretation of differential diagnosis. Our IADDx offers a comprehensive list of differential diagnoses that can explain the generated response. In this dialogue, IADDx identifies four potential diseases that exhibit similar symptoms. The first and second diseases are both related to GERD and have been utilized to generate the response. Beyond these two, physicians still need to consider the potential of the other diseases and seek to confirm or rule them out in subsequent conversations. IADDx also provides a diagnosis path composed of systems, organs, diseases, and symptoms to interpret the differential diagnosis. We select entities with high attention weights to build the diagnosis path. We observe that the patient’s condition predominantly pertains to the digestive system and the stomach. Additionally, the attention to specific symptom entities lends further support to the potential diseases.

4.5 Chapter Summary

In this chapter, we propose a medical dialogue generation framework with a differential diagnosis, IADDx, which explicitly models the process of differential diagnosis through intuitive association followed by analytic refinement. Moreover, we devise a diagnosis-oriented graph to interpret the differential diagnosis. The diagnosis results are utilized to guide response generation. Experiments on two datasets demonstrate the efficacy of IADDx. Besides, we generate intermediate outcomes and graph-based diagnostic paths to help understand the diagnostic procedure.

Patient: Hello, doctor! Over the past week, I've been experiencing occasional subtle pain in the center of my upper abdomen, and sometimes the pain radiates to my back. Last night, I felt a tightness in my chest and had acid reflux. What could be the issue?

Doctor: Hello, are you currently experiencing acid reflux or heartburn? Burp often?

Patient: I haven't had noticeable acid reflux, but had slight burning sensation. I started burping since yesterday. After taking medicine today, burping became less.

...

Patient: I didn't sleep well last night since I felt discomfort in my upper abdomen, and I woke up several times.

Gold Response: It's likely that you have gastroesophageal reflux disease (GERD). I'd recommend trying Mosapride. For now, avoid greasy, fried, and spicy foods.

BART: I suspect it's GERD.

BART-Entity: Yes, it seems to be caused by an excessive secretion of stomach acid.

DFMed: I believe you might have weak gastrointestinal motility and poor digestive function. I recommend taking Omeprazole, Domperidone, and Mosapride for treatment. Stick to a bland diet and avoid spicy and irritating foods.

Differential Diagnosis: (1) reflux esophagitis, (2) GRED, (3) bile reflux gastritis, (4) duodenogastric reflux

Attended Entities in DKG: [digestive system] → [stomach] → [reflux esophagitis / GRED] → [upper abdominal pain / poststernal burning sensation / acid reflux]

IADDx (Ours): I suspect you have GERD. I recommend taking Rabeprazole, Mosapride, and Magnesium aluminum carbonate for treatment. Over the next few days, be mindful of your diet. Focus on easily digestible foods, and try to avoid acidic, spicy, or cold foods.

Table 4.5: Case study. The responses are generated by different models, where diseases are underlined.

Chapter 5

Aligning Diagnostic Reasoning

Enabling medical dialogue systems to emulate the diagnostic reasoning processes of clinicians has been a long-standing research focus since the early days of the field [38]. A critical issue of the existing systems is that they overlook clinicians’ internal thought processes and overly rely on co-occurrence patterns in the training data, often defaulting to the most common disease. In this chapter, we explore the internal “Clinician Preference” during diagnostic reasoning and align the system’s reasoning process with clinical experts. We propose the Emulation framework, which first applies abductive and deductive reasoning modules to emulate clinicians’ diagnostic reasoning, and then generates thought processes aligned with clinicians’ preferences. Moreover, we construct a new dataset of diagnostic thought processes with the assistance of proprietary large language models (LLMs).

5.1 Introduction

Medical dialogue systems as a fundamental tool in facilitating effective healthcare consultations have garnered sustained attention in recent years [152, 71]. Especially with the rise of large language models (LLMs), these systems have shown promising potential to act as virtual medical assistants that can aid clinicians in accurate and efficient diagnosis [5, 15, 82]. In this research area, how to enable these medical systems to emulate clinicians’ diagnostic reasoning process has been the long-standing research focus ever since its inception [38]. To achieve this, previous research has paid tremendous efforts in constructing high-quality datasets of medical consultation dialogues, based on which they fine-tuned language models, rudimentarily realizing the simulation of clinicians’ diagnostic process that can inquire about symptoms and make a diagnosis [5, 15, 71, 139].

Nonetheless, existing research remains focused on the “output” of the clinician’s reasoning process (e.g., the symptoms and diseases mentioned in clinicians’ utterances), still neglecting the clinician’s internal thought process and decision-making mechanisms. As a result, available systems overly rely on the co-occurrence patterns in the training data, prone to inquire about the most frequent symptoms and diagnose the most common disease. In fact, a real clinician’s thought process goes far beyond this. In clinical medicine, there is a concept known as “Clinician Preference” [104, 38, 80]. It refers to the inclinations that healthcare professionals exhibit during diagnostic reasoning. It has been widely acknowledged that clinician preferences are affected by many factors [104], much more than the symptom and disease frequency. For example, in Figure 5.1, the clinician prefers to discuss examinations in the next response to rule out colitis rather than consider factors that cause irritable bowel syndrome. This is because the location of the discomfort typically indicates a more severe issue in the colon. For the existing systems that are constructed purely through dialogue data fine-tuning, it is difficult to capture such subtle reasoning preferences [136].

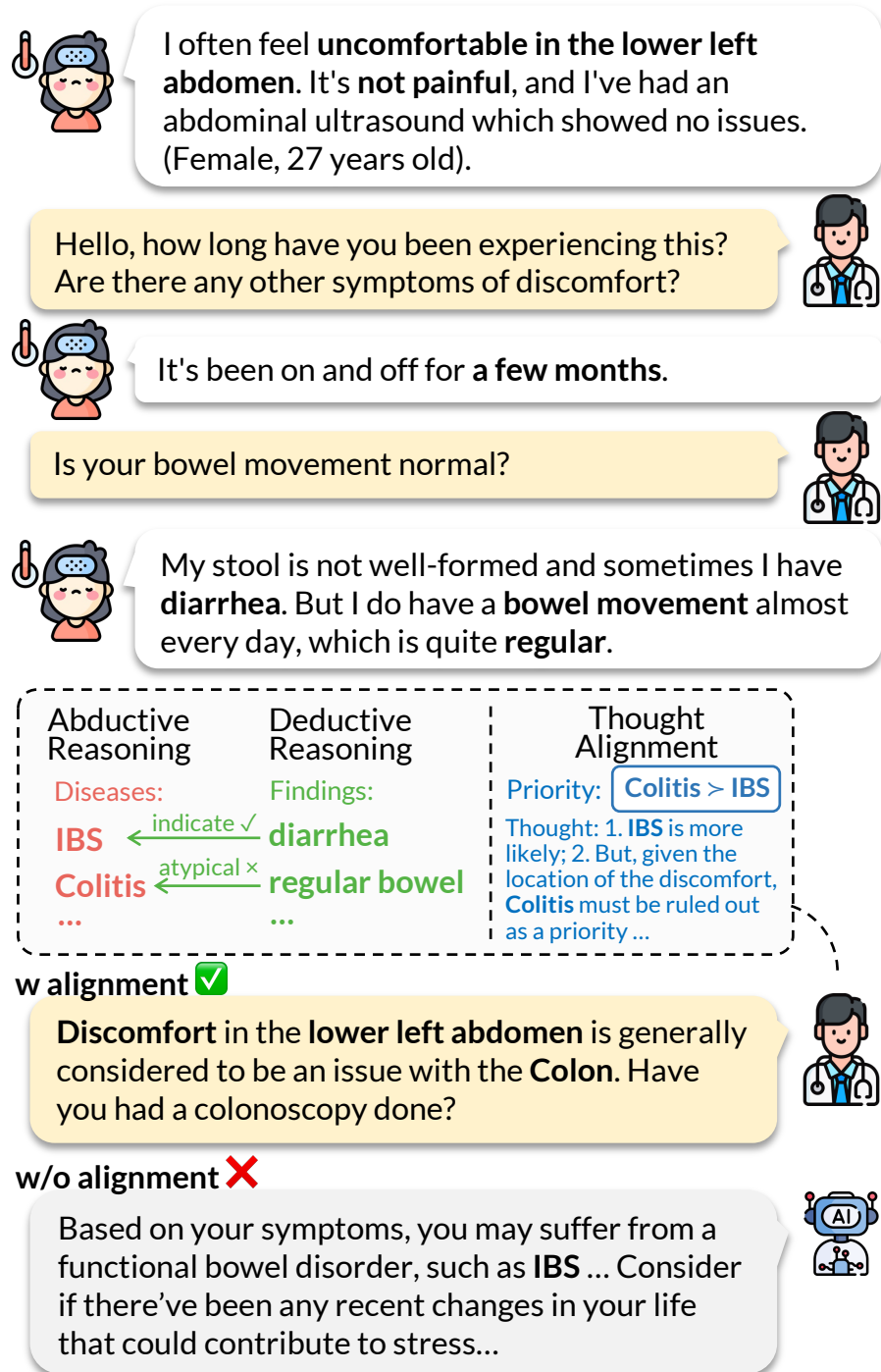


Figure 5.1: An example of a medical dialogue. **IBS** is the abbreviation for irritable bowel syndrome.

In this chapter, we aim to develop a medical dialogue system that can align with the internal diagnostic reasoning process of clinicians. To this end, we must first model a diagnostic analysis process. For the multi-turn medical dialogue, clinicians generally adopt an iterative, abductive, and deductive analysis that discovers an explanation of the patient’s condition and evaluates the effectiveness of the explanation [38]. This analysis provides a robust and comprehensive foundation for accurate diagnoses. Then, we need to align response generation with the clinician preference based on the analysis and dialogue context. The thought process of how clinicians reason and generate responses is a vital resource for learning their preferences. Extracting the thought process and modeling it together with responses can help learn the diagnostic reasoning process.

Based on the above motivation, we propose a novel medical dialogue system framework, EMULATION, which emulates clinicians’ diagnostic reasoning processes to generate an appropriate response that relies on ample diagnostic analysis and aligns with clinician preferences in consultation. First, an abductive reasoning module investigates potential diseases that can explain a patient’s condition. Then, a deductive reasoning module comprehensively analyzes the relation between clinical findings and potential diseases. Finally, the thought alignment module adjusts the potential disease priority that may be discussed next and generates thought processes that align with the clinician preference based on the above analysis. To learn the clinician preference, we build a new diagnostic thought process dataset with the help of an LLM.

Our key contributions are outlined as follows:

- We propose a novel medical dialogue system framework, EMULATION, that emulates clinicians’ diagnostic reasoning processes and aligns with clinician preferences. This is the first work that explores clinician preferences and internal thought processes.
- We build a diagnostic thought process dataset that is employed to align response

generation with clinician preferences.

- Experimental results demonstrate the effectiveness and explainability of EMULATION.

5.2 Preliminary

5.2.1 Problem Formulation

In our work, we conceptualize a medical dialogue as a sequence $U = \{(U_k^P, U_k^D)\}_{k=1}^T$, where each pair (U_k^P, U_k^D) comprises an utterance from a patient followed by an utterance from a doctor. Each doctor’s utterance is annotated with a list of diseases $E_t = \{e_i\}$ that could be relevant to the patient’s condition mentioned in the dialogue. Given a dialogue history $U_t = \{U_1^P, U_1^D, \dots, U_t^P\}$ up to the t -th patient utterance, our system’s objective is to generate a contextually appropriate and medically informed t -th doctor’s utterance U_t^D .

5.2.2 Diagnostic Reasoning

Diagnostic reasoning involves a detailed examination of how doctors think and make decisions in medicine [38]. It serves as a foundational element for advanced cognitive activities, such as formulating diagnostic conclusions and grasping the underlying pathology of diseases. The framework of diagnostic reasoning has been a focal point of medical cognition research [86]. A widely recognized perspective holds that diagnosis represents an iterative, abductive, and deductive process of formulating and evaluating potential explanations for a patient’s abnormal condition [27, 38]. We summarize this perspective into two reasoning processes: **Abductive Reasoning** aims to create a plausible diagnosis to explain observed clinical findings; **Deductive Reasoning** further tests the available diagnoses by determining whether the findings support,

refute, or are unrelated to the diagnoses. In our method, the first process efficiently explores a disease knowledge base to identify several possible diseases that explain the patient’s condition in the current turn of the conversation. The second process comprehensively inspects the relationship between the clinical findings and possible diseases. These two processes are conducted iteratively across the conversation. Besides, the clinician preference for medical decision analysis plays an essential role in the diagnostic reasoning process [104]. This preference highlights behaviors and actions in medical conversations unique to each clinician, which LLMs may lack. It is the nuanced preference difference that distinguishes a superior clinician from other clinicians [150]. We model a **Thought Alignment** process to align response generation with the general clinician preference in consultation.

5.2.3 Disease Annotation

Given GPT-4’s demonstrated effectiveness in several medical licensing examinations [82], we employ it to annotate potential diseases relevant to the patient’s condition automatically. We implement two ways to generate lists of diseases considered **before** and **after** examining the current doctor’s response. For the first list, we construct a prompt with the dialogue history U_t , utilizing the model’s diagnostic skills to identify potential diseases. For the second list, we create a prompt that incorporates both the history U_t and the ground truth response U_t^D , enabling the model to deduce the diseases that doctors might discuss. Then, we link the inferred diseases with those in an external medical knowledge base and obtain two lists of potential diseases, E_t^{pri} and E_t^{post} . We merge two lists as one E_t for each doctor’s response.

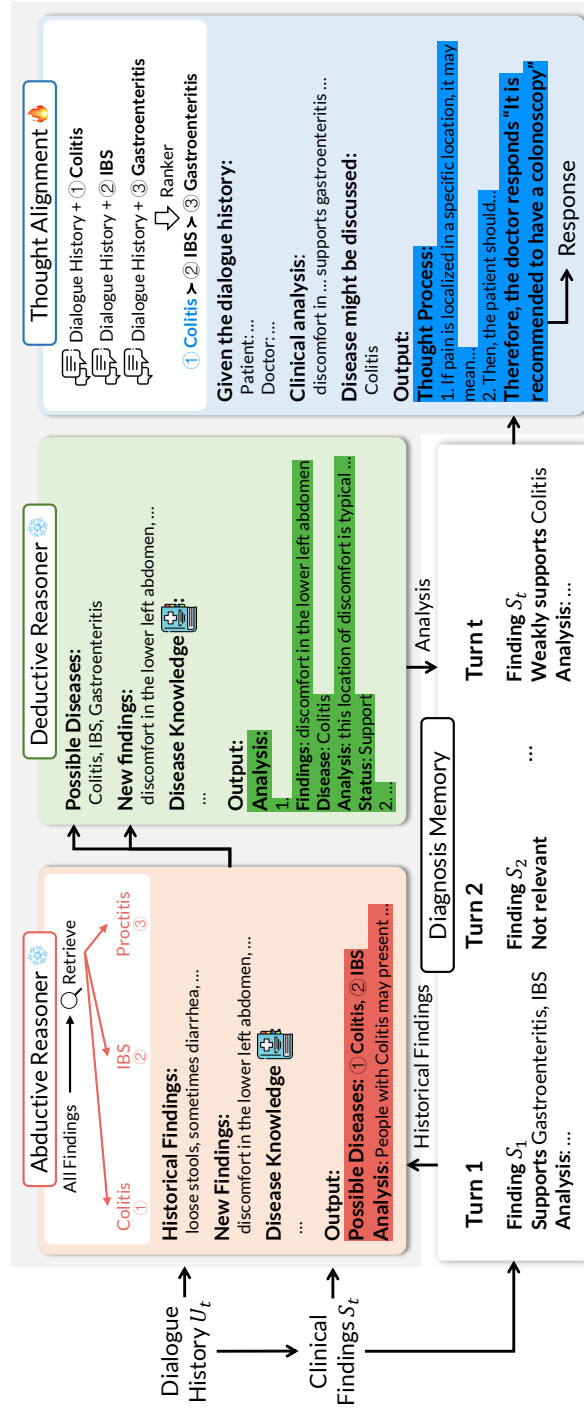


Figure 5.2: Illustration of our Framework EMULATION. The symbols of a snowflake or a flame represent the module mainly operating in a prompt-based way or having undergone fine-tuning, respectively.

5.3 Method

In this section, we introduce a diagnostic reasoning framework, which analyzes patient conditions through abductive reasoning and deductive reasoning [38], and aligns responses with the clinician preferences. As shown in Figure 5.2, our framework includes three modules. The *Abductive Reasoner* (§5.3.1) first generates potential diseases based on the new clinical findings of each turn. Then, the *Deductive Reasoner* (§5.3.2) comprehensively analyzes the relation between potential diseases and new clinical findings. Finally, the *Thought Alignment* (§5.3.3) module adjusts the disease priority that might be discussed in subsequent dialogues and performs thought process modeling to generate the next response.

Given the medical dialogue history U_t at the t -th conversation turn, we first use a pre-trained LLM to summarize new clinical findings from the recent utterances U_{t-1}^D and U_t^P . The clinical findings are presented using the SOAP note [8], a documentation technique utilized by healthcare professionals for recording notes (e.g., chief complaint) in a patient’s medical record. For example, we provide the LLM with a prompt to “*summarize the clinical findings from the recent conversation between doctor and patient, adhering to the structure of the SOAP note*”. Each of the summarized clinical findings is in the phrase format and merged into a set $S_t = \{s_j\}_{j=1}^m$, where m represents the number of clinical findings at the t -th conversation turn.

5.3.1 Abductive Reasoner

The Abductive Reasoner aims to find the most possible diseases that explain the patient’s abnormal condition. Previous studies rely solely on the diagnosis capability of black-box language models or focus on limited diseases in one dataset. We design a two-step pipeline to achieve a comprehensive and explainable abduction. The first step employs a disease retriever to narrow down the scope of the disease. Then, the second

step leverages a pre-trained LLM to generate a potential disease list with detailed explanations based on the clinical findings and external medical knowledge.

Disease Retrieval. We concatenate all findings as the query and retrieve disease documents that are relevant to the findings from an external medical knowledge base. A dense retriever with a BERT encoder is applied to collect relevant documents. The sequences of the query and disease document, each prefixed with a “[CLS]” token, are fed independently into the encoder. The hidden state corresponding to the “[CLS]” token from each sequence is chosen to serve as their respective representations $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{h}_{doc} \in \mathbb{R}^d$. We compute relevance scores through dot product as $\mathbf{h}^\top \cdot \mathbf{h}_{doc}$. The diseases corresponding to the top- K documents ranked by the scores are considered candidate diseases $\hat{E}_t^0 = \{e_i\}_{i=1}^K$. Our approach utilizes Contrastive Learning [31] to train the retrieval model. The loss function applied to the training is specified as follows:

$$\mathcal{L}_{CL} = -\log \frac{\exp(\mathbf{h}^\top \cdot \mathbf{h}_{doc}^+)}{\exp(\mathbf{h}^\top \cdot \mathbf{h}_{doc}^+) + \sum_{\mathcal{C}} \exp(\mathbf{h}^\top \cdot \mathbf{h}_{doc}^-)}, \quad (5.1)$$

where $\mathbf{h}_{doc}^+$ and $\mathbf{h}_{doc}^-$ denotes document representations of the relevant diseases in this turn E_t and the irrelevant ones in the knowledge base $\mathcal{C}$. We adopt in-batch negative sampling for training.

Diagnosis Refinement. After the retrieval, we have assembled a list of potential diseases; however, this approach unavoidably results in the inclusion of diseases irrelevant to the patient’s condition. Furthermore, the disease retrieval, based on vector similarity, can not elucidate the rationale behind the presence of specific diseases within the potential list. In pursuit of an explainable diagnosis, we utilize a pre-trained LLM to identify possible diseases from this list, offering an explanation grounded in clinical findings. The diagnosis knowledge of each disease is provided since it can appropriately reduce the hallucination issue of LLMs. For instance, we prompt the

LLM with “*given the findings and knowledge, select possible diseases that can explain the new clinical findings while satisfying past findings*”. To achieve a stable disease identification, we adopt a majority vote inspired by [125]. Specifically, we randomly divide the list into batches of the same size (since the maximum input tokens of LLMs limit the diseases and knowledge that can be input). Such division is repeated several times to construct different batch groups. We utilize batch groups to generate several refined disease lists and calculate the voting score of each disease as follows:

$$v(e_i) = \sum_{j=1}^B \mathbb{1}(e_i \in \hat{E}_t^j), e_i \in \hat{E}_t^0, \quad (5.2)$$

where B stands for the number of batch groups, and $\hat{E}_t^j$ represents the j -th refined list. The final refined list includes diseases with a voting score exceeding $B/2$, denoted as $\hat{E}'_t = \{e_i\}_{i=1}^{K'}$.

5.3.2 Deductive Reasoner

While the Abductive Reasoner has offered an analysis of the relationship between potential diseases and new clinical findings, this analysis predominantly aims to confirm a disease rather than exclude one or sift through information irrelevant to the current diagnosis. It is crucial for clinicians to thoroughly examine the information presented in the dialogue history. However, previous studies often overlook these non-affirmative analyses in their diagnostic processes, which is not conducive to improving diagnostic accuracy.

We introduce the Deductive Reasoner to evaluate the affirmative and non-affirmative relations between new clinical findings and possible diseases. Given the clinical findings S_t and the refined disease list $\hat{E}'_t$, this reasoner applies a pre-trained LLM with a prompt as “*analyze if the new clinical findings support, oppose, or are irrelevant to the possible diseases*”. We improve the accuracy of analysis by incorporating additional diagnostic knowledge of each disease. The content of specific tags, i.e., findings, disease,

and status, is extracted as diagnosis memory M_t for subsequent response generation.

5.3.3 Thought Alignment

The clinician preference in consultation is critical, as it differentiates an experienced clinician from a novice clinical student. Prior research has always concentrated on diagnosing the most likely disease and discussing extra symptoms or treatment based on that diagnosis, which does not accurately reflect the dynamics of an actual healthcare consultation. To address this, we have developed the Thought Alignment module, which is designed to adapt preferences akin to that of an expert clinician. The module first prioritizes diseases to be discussed in the subsequent dialogue. Then, it models the thought process and response of real clinicians to align with clinician preferences.

Disease Priority Alignment We implement a disease ranker to establish the priority of each disease within the refined disease list $\hat{E}'_t$. The backbone of this ranker is the BERT encoder. For the input, we combine the dialogue history with each potential disease, formatting it as follows: “[CLS] {history} *the next response will discuss: {disease}*”. The relevance score of the dialogue history to the discussed disease is calculated:

$$r(U_t, e_i) = \text{MLP}(\text{repr}(U_t; e_i)), e_i \in \hat{E}'_t, \quad (5.3)$$

where $\text{repr}(\cdot)$ extracts the hidden state of the “[CLS]” token as the representation, and $\text{MLP}(\cdot)$ projects the representation to a scalar. The ranking model is also trained using a contrastive loss function. To better adjust the priority within the refined disease list, the selection of negative samples includes the union of this refined list $\hat{E}'_t$ and the pre-annotated disease list E_t^{pri} . The positive diseases are from E_t^{post} . We compute relevance scores between the dialogue history and all diseases in the refined list and then reorder the list to $\hat{E}''_t$.

Thought Process Alignment Thought processes are a reflection of clinician preferences in consultation. Motivated by studies on distilling multi-step reasoning capabilities [30, 9], we design a thought distillation method to empower a model that can generate thought processes consistent with clinician preferences. Specifically, we first leverage a pre-trained LLM to deduce a plausible thought process for each doctor’s response. The prompt to the LLM is organized as “*complete the thought process based on the dialogue context*”. We augment the prompt with 3 human-annotated thought processes in the Chain-of-Thought (CoT) format. The thought extraction can be defined as:

$$Y_t \sim \text{LLM}(Y_t|U_t, U_t^D, \text{prompt}). \quad (5.4)$$

Then, we gather the thought process for each dialogue turn and develop a thought process alignment model using the autoregressive language modeling approach as our training objective. The loss function can be defined as:

$$\mathcal{L}_G = -\sum_j \log p(y_{jt}|U_t, M_{\leq t}, E_t^{post}, Y_{t,<j}), \quad (5.5)$$

where $M_{\leq t}$ denotes the diagnosis memory until the t -th turn and E_t^{post} denotes the diseases discussed in the next response. During the inference, we utilize the top- K diseases in the list $\hat{E}_t''$ instead of E_t^{post} and their related analyses to generate the thought process. The final response is extracted from “*Therefore, the doctor responds {response}*”.

5.4 Thought Process Corpus

In this section, we present the medical dialogue thought process corpus, MEDIATOR, where each dialogue turn is annotated with a chain-of-thought reasoning path. We utilize the powerful reasoning ability of GPT-4 and its impressive professional skills in the medical domain for automatic annotation. The annotated reasoning path is expected to reflect the thought process of clinicians who analyze the patient’s

Description	MedDG	KaMed
# Thought processes	122K	285K
Avg. # steps	4.17	4.13
Avg. # tokens per step	40.98	41.21
Avg. # total tokens	236.39	239.53

Table 5.1: Statistics of MEDIATOR

condition and determine what to discuss. Two datasets, MedDG [71] and KaMed [58], are annotated following the few-shot prompt in §5.3.3. As shown in Table 5.1, the former is annotated with 122K thought processes and the latter with 285K. Each thought process includes a reasoning path consisting of approximately four steps. An example of the generated thought process is displayed in Table 5.2. The doctor investigates the new clinical findings (i.e., discomfort under certain conditions) and plans to discuss allergy issues after multi-step reasoning.

Human assessments are carried out to evaluate the quality of the generated thought processes. We randomly select 100 samples from each dataset and ask three medical students who have undergone clinical internships to assess them. The evaluation employs three metrics: (1) **Knowledge**: whether the knowledge used in the thought process is accurate; (2) **Consistency**: whether the thought process is consistent with the dialogue history; (3) **Rationality**: whether the thought process starts with premises and uses logical progression to derive responses. Table 5.3 presents the evaluation results, demonstrating that the thought processes in MEDIATOR adequately fulfill the above three criteria.

Patient: What’s going on with my throat feeling itchy and making me want to gag? (Female, 22 years old)

Doctor: How long have you been experiencing this?

Patient: About half an hour. And my temperature is just under 37 degrees.

Doctor: Have you had pharyngitis before?

Patient: No, but I have chronic gastritis.

Doctor: So, this hasn’t happened before?

Patient: Only when I have a cold, I would feel itchy and cough. Also, I tend to feel uncomfortable when the seasons change. Other than that, no.

Thought Process:

1. The patient mentions symptoms of itchy throat and gagging during colds or when the seasons change. This indicates respiratory issues.
2. The doctor considers the possibility of an allergic reaction, as ... Seasonal changes frequently introduce variations in allergens and irritants in the environment, which might explain the patient’s symptoms.
3. The patient’s symptoms are not frequent, but appear to be triggered by specific conditions or events. This further supports the allergies or environmental causes.
4. The patient mentions having chronic gastritis, but the relation to the current symptoms is unclear ... but usually does not cause an itchy throat. So, confirming allergy issues is a priority.

Therefore, the doctor responds, "Have you done any cleaning today? Or been exposed to dust mites?"...

Table 5.2: An example of the generated thought process.

5.5 Experiments

5.5.1 Datasets

Our experiments utilize two medical dialogue datasets, MedDG [71] and KaMed [58]. Dialogues in these datasets exhibit a clear multi-turn context, with each dialogue

Datasets	Knowledge	Consistency	Rationality
MedDG	90%	89%	82%
KaMed	87%	91%	79%

Table 5.3: Human evaluation results of automatically generated thought processes. Values in the table represent average valid percentages.

averaging about 10 turns. The MedDG dataset comprises 17,860 dialogues, focusing on 12 gastroenterology-related diseases. The dataset is divided into 14,862/1,999/999 for training, validation, and testing. The KaMed dataset includes more than 63,000 dialogues, spanning a wide range of diseases across approximately 100 hospital departments. We clean privacy-sensitive content following DFMed [139] and divide the dataset into 29,159/1,532/1,539 for training, validation, and testing.

5.5.2 Baseline methods

We compare our method with two categories of baselines: LLMs equipped with Chinese medical conversation abilities and language models that are fine-tuned on target datasets.

Medical LLMs. (1) **HuatuoGPT-II** [15] is a medical LLM performing state-of-the-art in several Chinese medical tasks. (2) **DISC-MedLLM** [5] is a medical LLM with strong multi-turn consultation capabilities. (3) **GPT-4** [82] is one of the most advanced pre-trained LLMs designed by OpenAI.

Fine-tuned Models. (1) **VRBot** [58] is a medical dialogue generation (MDG) model with entity tracking and predicting. (2) **GPT-2** [91] is a transformer decoder-based language model. (3) **BART** [56] is a transformer-based encoder-decoder model. (4) **Qwen-7B** [4] is a strong base language model focusing on Chinese and English.

Methods	MedDG						KaMed					
	B-1	B-4	R-1	R-2	E-F	B-1	B-4	R-1	R-2	E-F	B-1	E-F
Zero-shot LLMs	HuatuoGPT-II	42.45	24.78	15.85	4.24	9.45	40.89	22.9	18.06	4.65	11.46	11.46
	DISC-MedLLM	40.72	22.6	20.13	6.6	10.15	38.05	20.26	20.48	5.93	13.54	13.54
	GPT-4	42.19	23.32	13.99	3.47	13.15	41.88	23.34	13.94	3.1	13.86	13.86
Fine-tuned Models	VRBot	29.69	16.34	24.69	11.23	12.78	30.04	16.36	18.71	7.28	12.08	12.08
	GPT-2	35.27	19.16	28.74	13.61	16.14	33.76	17.82	26.80	10.56	17.26	17.26
	BART	34.94	19.06	29.03	14.40	16.66	33.62	17.64	27.91	11.43	19.20	19.20
	Qwen-7B	35.11	19.03	30.19	15.01	18.05	34.00	17.66	28.34	12.18	19.88	19.88
	DFMed	42.83	22.90	29.72	14.31	22.92	40.50	20.92	28.33	11.73	22.31	22.31
	EMULATION	42.35	22.76	30.91 [†]	15.17 [†]	24.03 [†]	39.87	19.79	28.54 [†]	12.33 [†]	24.27 [†]	24.27 [†]

Table 5.4: Automatic evaluation results on two datasets. † denotes statistically significant differences ($p < 0.05$).

(5) **DFMed** [139] enhances MDG with entity and dialogue act flow learning.

5.5.3 Implementation Details

All the prompt-based operations in EMULATION are implemented with `gpt-3.5-turbo`, i.e., clinical findings extractions, abductive diagnosis refinement, and deductive diagnosis analysis. We apply the MedBERT¹ pre-trained in the medical domain as the backbone of the disease retriever and the disease alignment model. The corpus of disease documents, utilized for disease retrieval and to enhance both abductive and deductive reasoning, is derived from a specialist-certified online medical knowledge base *xiaohe*². We retrieve the top 50 ($K=50$) relevant diseases for further refinement. The refined disease list is sent to the disease alignment model to obtain the top 5 ($K''=5$) diseases that may be discussed in subsequent responses. Then, we employ the pre-trained language model Qwen-7B-Chat³ to train the thought process alignment model, which has seven billion parameters and proficient Chinese understanding and generation ability. We train the model using the LoRA approach with $r=64$ and $\alpha=16$.

For the medical LLMs, we use the HuatuoGPT2-13B⁴ and DISC-MedLLM⁵ to generate responses in zero-shot way as they have already been fine-tuned on medical dialogue datasets and other large-scale medical knowledge datasets. These two models are all based on language models with 13B parameters. We utilize the default decoding parameters to generate responses. For fine-tuned baseline methods, we employ generation results from the original papers.

For training disease retriever in §5.3.1, we use the MedBERT as the encoder and

¹<https://github.com/trueto/medbert>

²<https://www.xiaohe.cn/medical>

³<https://github.com/QwenLM/Qwen>

⁴<https://github.com/FreedomIntelligence/HuatuoGPT-II>

⁵<https://github.com/FudanDISC/DISC-MedLLM>

Methods	Proactivity	Accuracy	Helpfulness	LQ	Average
DISC-Med	4.52	4.41	4.71	4.98	4.66
DFMed	4.73	4.48	4.79	4.86	4.72
EMULATION	4.69	4.65	4.82	4.99	4.79

Table 5.5: Automatic evaluation results based on GPT-4.

employ contrastive learning. The batch size is 8 with 6 gradient accumulation steps. The learning rate is set at $3e-5$. We train 6 epochs for the MedDG dataset and 10 epochs for the KaMed dataset. We select the checkpoint with the highest disease recall rate in validation datasets. For training the disease priority alignment model, we also use the MedBERT as the encoder and employ contrastive learning. The batch size is 2 with 8 gradient accumulation steps. The learning rate is set at $1e-5$. We train 3 epochs for two datasets and select the checkpoint with the highest IoU in validation datasets.

5.5.4 Automatic Evaluation

We assess the generated responses using three automated metrics: BLEU-1/2/4 (**B-1/4**) [85], evaluating n -gram precision; ROUGE-1/2 (**R-1/2**) [64], assessing n -gram recall; and Entity-F1⁶ (**E-F**) [71], which gauges the accuracy of medical entities, such as diseases and medications.

Table 5.4 displays the response generation results of all baseline methods. Our framework outperforms the baselines in most metrics, especially the medical entity accuracy. It demonstrates the effectiveness of our framework in diagnostic reasoning and response generation. Specifically, EMULATION performs better than available medical LLMs trained on large-scale medical dialogues and knowledge bases in R-1/2 and E-F. It demonstrates that modeling the diagnostic reasoning process can help

⁶<https://github.com/lwgkzl/MedDG>

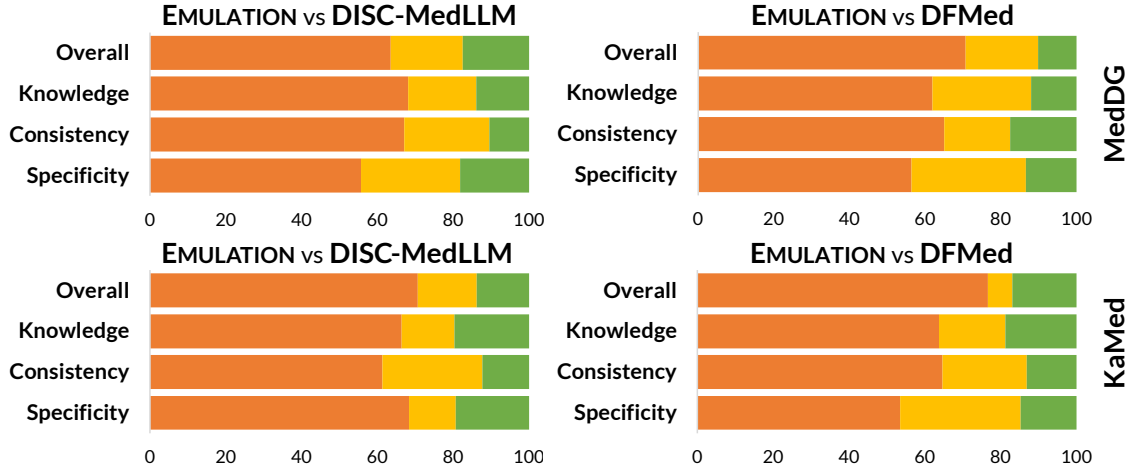


Figure 5.3: Human evaluation results of baseline methods. Red is for **Win**, yellow for **Tie**, and green for **Lose**.

generate more accurate and targeted responses. Besides, our framework achieves better n -gram recall and entity accuracy than the state-of-the-art model DFMed. The B-1/4 scores are a bit lower, which may be because the fine-tuned parameters are too limited to approach the linguistic pattern of the gold utterances. However, benefiting from the construction of diagnostic reasoning processes, EMULATION can generate responses with more consistent content and higher entity accuracy.

We also adopt another multi-dimensional automatic evaluation using GPT-4, which follows the method in DISC-Med [5]. This evaluation focuses on **Proactivity**, **Accuracy**, **Helpfulness**, and **Linguistic Quality (LQ)**. Our EMULATION performs better than other baselines in the above aspects as shown in Table 5.5.

5.5.5 Human Evaluation

We conduct human evaluations on comparison with two baselines: DISC-MedLLM and DFMed. We drew a random sample of 100 cases and had three medical students

Methods	MedDG			KaMed		
	B-4	R-2	E-F	B-4	R-2	E-F
EMULATION	22.76	15.17	24.03	19.79	12.33	24.27
w/o Abd. Reasoning	20.30	13.77	17.47	18.02	11.31	19.75
w/o Ded. Reasoning	22.34	15.07	23.77	19.63	12.05	24.11
w/o Dis. Alignment	21.93	14.69	21.24	18.84	11.87	22.43
w/o Thot. Alignment	22.31	14.95	23.82	19.72	12.12	24.08

Table 5.6: Ablation study on two datasets

evaluate them according to **Knowledge**, **Consistency**, and **Specificity**. Besides, the overall quality is compared. We compare our framework EMULATION with baseline methods from three aspects. (1) Knowledge Accuracy (Knowledge), which assesses whether the response correctly applies the disease knowledge. (2) Consistency, which assesses whether the response is consistent with the dialogue context and doctor’s ground truth response. (3) Specificity, which assesses whether the response provides specific diagnoses, prescriptions, treatment plans, or examination requirements rather than general suggestions. We ask our annotators to compare the responses generated by our method with the responses generated by baseline methods. EMULATION outperforms two baselines in all aspects, as shown in Figure 5.3.

5.5.6 Analysis of Diagnostic Reasoning Process

To further evaluate the effectiveness of our framework, we analyze several variations of our EMULATION as detailed below: (1) **w/o Abd. Reasoning**, which omits abductive reasoning along with any related deductive reasoning, relying solely on dialogue history to form thought processes. (2) **w/o Ded. Reasoning**, which eliminates the process of deductively analyzing clinical findings and potential diseases. (3) **w/o Dis. Alignment**, which forgoes learning the disease priority, opting to select

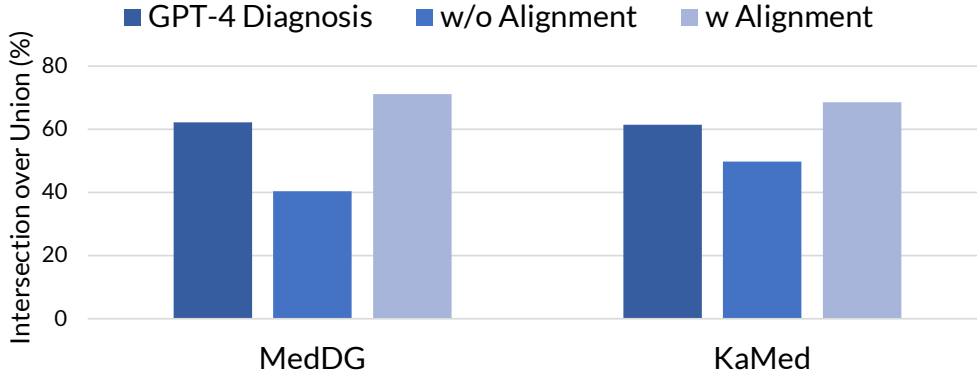


Figure 5.4: The top-5 diagnosis results with or without disease priority alignment.

the most relevant diseases identified through abductive reasoning. (4) **w/o Thot. Alignment**, which directly generates responses based on the results of abductive and deductive reasoning.

Table 5.6 presents the comprehensive results of our ablation study. There is a noticeable decline in effectiveness across various metrics for the ablation models, underscoring the indispensable contribution of each module within our framework. Particularly, the variant *w/o Abd. Reasoning* experiences a significant drop in response quality due to the lack of initial diagnosis and subsequent analysis. This decline is attributed to the typical practice of clinicians to center medical conversations around certain diseases; thus, omitting precise abductive and deductive reasoning processes leads to a lack of focus in the dialogue. Besides, the *w/o Dis. Alignment* variant shows a marked decrease in performance, reinforcing the importance of disease priority alignment with clinician practices for achieving consistent responses. The outcomes from *w/o Ded. Reasoning* and *w/o Thot. Alignment* further affirm that detailed analyses and thought generation enhance the quality of response generation.

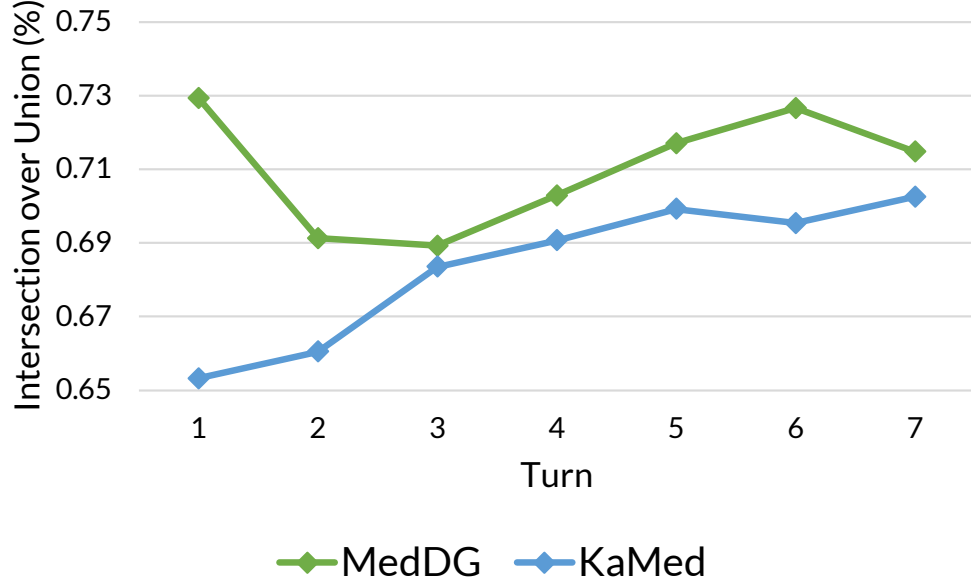


Figure 5.5: The top-5 diagnosis results of EMULATION at different conversation turns.

5.5.7 Analysis of Disease Alignment

We evaluate the diagnosis accuracy to determine the impact of aligning disease priorities. The metric Intersection over Union (IoU) between predicted diseases and plausible diseases E_t^{post} is employed. As shown in Figure 5.4, aligned diagnoses substantially outperform unaligned diagnoses that are derived directly from abductive reasoning. Remarkably, aligned diagnoses are marginally more accurate than the diseases diagnosed by GPT-4, i.e., E_t^{pri} . We also display the diagnosis results in different conversation turns as shown in Figure 5.5. Overall, the diagnosis accuracy continuously improves as the number of conversation turns increases. The diagnostic accuracy on the MedDG dataset was initially high but decreased in the second turn. It is because the patients in this dataset provide relatively sufficient information in the first turn, allowing EMULATION to make a good rough diagnosis. However, with the introduction of additional information, the difficulty of diagnosis increases, and the difference between EMULATION’s diagnosis and that of real doctors becomes greater.

5.6 Case Studies

Figure 5.6 and Figure 5.7 display cases on the MedDG and KaMed datasets, respectively.

5.7 Chapter Summary

This chapter proposes a novel medical dialogue system framework, EMULATION, that emulates clinicians' diagnostic reasoning processes to generate appropriate responses grounded in abductive and deductive diagnostic reasoning analysis and alignment with clinician preferences. Besides, a new diagnostic thought process corpus is presented and utilized to model the clinician preference. Experimental results demonstrate the effectiveness of EMULATION on two datasets. One promising area for future work is applying the EMULATION framework in telemedicine consultations. With the increasing demand for telemedicine, it is essential to explore how this framework can enhance virtual patient-doctor interactions. Future studies could investigate the framework's effectiveness in improving patient satisfaction and overall consultation quality in remote settings.

Given medical dialogue as follows:

Patient: A 28-year-old male previously had bowel movements roughly every 3-4 days with thick stools. However, he had a good appetite before, but in the recent half year, he has been experiencing discomfort in his stomach and intestines, with a decreased appetite, and now has bowel movements every two days, which are thin and soft. A year ago, a gastroscopy revealed gastritis erosion, and tests for *Helicobacter pylori* were negative. (Male, 28 years old).

Doctor: Hello! How long have you been experiencing this situation? Is your diet regular? What kind of food do you like? Do you eat a lot of spicy or cold food?

Patient: This situation has lasted for half a year. I'm busy with work, so I only have an irregular meal at noon every day. Not much spicy or cold food.

Doctor: What other symptoms do you have?

Patient: I often feel discomfort in my stomach and intestines, and the regularity and shape of my bowel movements are completely different, with no other noticeable symptoms. I have previously taken stomach comfort granules and Stomach Ease. Do I need to undergo a colonoscopy or any other examination for my condition? Or what should I pay attention to and treat next?

Clinical findings in the dialogue are:

28 years old, male, thick stools, good appetite, discomfort in stomach and intestines, decreased appetite, thin and soft stools, gastritis erosion, half a year, irregular diet, not much spicy or cold food, different bowel movement regularity and shape

Clinical analysis:

- "Stomach and intestinal discomfort, decreased appetite, change in bowel habits, thin and soft stools." supports "Gastritis, Irritable Bowel Syndrome, Functional Gastrointestinal Disease"
- "Diagnosed with gastritis erosion a year ago." supports "Gastroesophageal Reflux Disease"
- "Helicobacter pylori test results were negative." opposes "Gastritis"
- "Patient has had symptoms for half a year." supports "Gastritis"
- "Patient's diet is irregular, usually only eating lunch due to being busy at work." supports "Irritable Bowel Syndrome"
- "Patient rarely eats spicy or cold food." opposes "Gastroesophageal Reflux Disease"
- "Frequent stomach and intestinal discomfort" supports "Gastritis"
- "Change in bowel movement regularity and shape" supports "Irritable Bowel Syndrome"
- "No other significant symptoms" opposes "Peptic Ulcer"

Possible diseases that might be discussed:

Gastritis, Irritable Bowel Syndrome, Peptic Ulcer, Gastroesophageal Reflux Disease, Functional Gastrointestinal Disease

Thought process:

1. The patient reports frequent stomach and intestinal discomfort and significant changes in bowel movement regularity and shape. These symptoms, along with the patient's history of gastritis erosion, suggest the possibility of chronic gastrointestinal issues.
2. The patient mentioned that they have tried "stomach comfort granules" and "Stomach Ease," which are traditional Chinese medicines commonly used for treating gastrointestinal discomfort. The effectiveness of these treatments could provide clues about the nature of the patient's condition.
3. The patient inquires about the need for a colonoscopy, indicating their concern about the severity of their condition. A colonoscopy could provide valuable information about the patient's gastrointestinal health and help identify any abnormalities. But it is not needed at the moment as there are no symptoms related to colitis.
4. Considering the patient's symptoms, medical history, and response to previous treatments, the doctor contemplates recommending a gastroscopy to further investigate the patient's stomach condition. This could help identify any potential issues causing the patient's symptoms.
5. The doctor may advise the patient to pay attention to their diet, as irregular eating habits could exacerbate gastrointestinal problems. Therefore, the doctor replies, "You could consider a gastroscopy to check. Pay attention to your diet in the meantime." This includes suggesting a gastroscopy to further investigate the patient's stomach condition and advising the patient to pay attention to their diet, as irregular eating habits could exacerbate their symptoms.

Ground Truth Response:

There's no need for a colonoscopy for the time being, but it's crucial to pay attention to your diet regularly.

DFMed:

Your situation is likely due to gastrointestinal dysfunction. It's recommended to have a light diet, keep warm, avoid spicy, stimulating, and greasy foods, and eat more vegetables and fruits. You can take Trimebutine and Bifidobacterium triplex live bacteria tablets together to improve symptoms. If that doesn't work, you may need to go to the hospital for a colonoscopy examination.

DISC-MedLLM:

Your symptoms may be caused by gastritis or gastrointestinal dysfunction. However, persistent symptoms may require further medical evaluation. 1. Diet: Try changing your dietary habits, avoiding overly greasy, spicy, or irritating foods, and increasing the intake of dietary fiber such as whole grains, vegetables, and fruits to help improve digestion. 2. Medication: Continue taking stomach medication as directed by your doctor. If symptoms persist or worsen, the medication treatment plan may need to be adjusted...

Figure 5.6: Case studies on the MedDG dataset.

Given medical dialogue as follows:
Patient: What tests are needed to check for tuberculosis? (Female, 2 years and 9 months old).
Doctor: To check for tuberculosis, you need a chest CT, sputum test, tuberculosis infection T-cell test, and tuberculin skin test, among others.
Patient: How much blood needs to be drawn for the tuberculosis infection T-cell test and tuberculin skin test for children aged 2 to 3? What's the general heart rate range for children of this age?
Doctor: A few milliliters of blood. The likelihood of tuberculosis in toddlers is extremely low.
Patient: Our child was hospitalized for pneumonia, treated for 27 days with no significant improvement, even with meropenem administered.

Clinical findings in the dialogue are:
Female, 2 years and 9 months old, child aged 2 to 3, pneumonia, treated for 27 days with no significant improvement, administered meropenem.

Clinical analysis:
- "The child was hospitalized for pneumonia, treated for 27 days with no significant improvement" supports "Tuberculosis, Pneumonia, Tuberculosis tract infection, Drug-resistant tuberculosis"
- "Meropenem was administered" supports "Tuberculosis, Drug-resistant tuberculosis"

Possible disease might be discussed:
Tuberculosis, Pneumonia, Tuberculosis, Respiratory tract infection, Drug-resistant tuberculosis

Thought process:
1. The patient mentions that their child was hospitalized for pneumonia, treated for 27 days but showed no significant improvement. This raises concerns as it suggests that the current treatment regimen might be ineffective.
2. The patient also mentioned that the child had been treated with meropenem, an antibiotic commonly used for severe bacterial infections. This may indicate that the pneumonia is bacterial, which is common in children.
3. However, despite treatment, the lack of improvement raises concerns about the accuracy of the diagnosis and the effectiveness of the treatment. It's also possible that the bacteria causing the pneumonia have developed resistance to meropenem.
4. Considering these factors, the doctor suggests conducting a tuberculin skin test. This test can help determine whether the child has been infected with the bacteria that cause tuberculosis. While tuberculosis is less common in children, it cannot be ruled out when a child is ill for a long time and unresponsive to treatment.
Therefore, the doctor replies, "A tuberculin skin test can be done." This indicates that, given the lack of improvement with current treatments, a tuberculin skin test is recommended to explore other possible causes for the child's persistent symptoms. This reflects the doctor's thoroughness in considering all possible diagnoses and the importance of re-evaluating the diagnosis when treatment is ineffective.

Ground Truth Response:
I suggest you get tested to rule out tuberculosis.

DFMed:
That means there has been no improvement, further examination is needed.

DISC-MedLLM:
Your child has been hospitalized for pneumonia for 27 days and has been treated with meropenem, but there hasn't been significant improvement. In this situation, we need to reassess your child's condition to determine the next steps in treatment. Meropenem is a potent antibiotic commonly used to treat severe infections, but if there hasn't been noticeable improvement after its use, we need to consider the following. Check for resistance: It's possible that the bacteria causing the infection have developed resistance to meropenem, resulting in poor treatment outcomes. Further bacterial culture and sensitivity testing are needed to identify the most effective antibiotic. Rule out complications: Pneumonia can lead to other complications such as pleural effusion or empyema, which may require specific management.

Figure 5.7: Case studies on the KaMed dataset.

Chapter 6

Enhancing Medical Reasoning with External Knowledge

Large-language models (LLMs) have demonstrated strong performance on diverse medical benchmarks, underscoring their potential to support real-world clinical tasks [11, 121, 138]. Given the breadth and high-stakes nature of medicine, retrieval-augmented generation (RAG) has emerged as a leading approach for closing knowledge gaps and reducing hallucinations by incorporating external medical knowledge [73, 126, 132, 147]. Although RAG performs well on many medical tasks, it still struggles with complex questions that demand intensive reasoning [109]. To address this issue, we construct an explicit thought process that reasons through each medical question and surfaces its implicit knowledge requirements. This thought process is used to retrieve relevant information, which is then incorporated into reasoning to derive the final answer. To optimise both thought and answer generation, we propose a joint learning framework, RAR², that simultaneously enhances **R**easoning-**A**ugmented **R**etrieval and **R**etrieval-**A**ugmented **R**easoning. Extensive experiments on several biomedical question-answering datasets show that RAR² consistently outperforms baselines. Further test-time scaling analysis confirms the robustness and scalability.

6.1 Introduction

The capabilities of Large Language Models (LLMs) in medicine have long attracted substantial research attention [105, 106, 121]. LLMs such as GPT-4o and Baichuan-M1 have demonstrated strong performance across diverse medical benchmarks, highlighting their potential to support real-world clinical tasks [11, 121, 138]. More recently, given the vast scope and high-stakes nature of the medical domain, Retrieval-Augmented Generation (RAG) has emerged as a key approach for mitigating knowledge gaps and hallucinations by leveraging external medical knowledge [73, 126, 132, 147].

While RAG performs well in a range of medical tasks, it still struggles with complex medical questions requiring intensive reasoning [109]. This difficulty arises because the surface-level information in these questions does not reflect their actual knowledge requirements. For example, in Figure 6.1, although the question clearly describes the patient’s symptoms, such as “weak in upper body and arms” and “won’t move legs”, the knowledge necessary to generate the correct answer lies in clinical concepts like “autonomic dysfunction” and “infant botulism”. These concepts must be inferred through detailed analytical reasoning. Existing RAG methods for complex medical reasoning primarily focus on direct query refinement, but fail to construct and optimize a reasoning process that comprehensively uncovers the underlying knowledge requirements [73, 62, 39].

Additionally, few studies have explored enhancing the medical reasoning capabilities of LLMs within the RAG framework [44]. Reasoning with external knowledge is more challenging, as the retrieved content inevitably includes noise information. LLMs should learn to integrate relevant knowledge into the reasoning process while avoiding interference from irrelevant information. Therefore, optimizing retrieval-augmented reasoning is essential for improving both the accuracy and robustness of generation outcomes.

In this work, we construct a thought process to reason through medical questions and

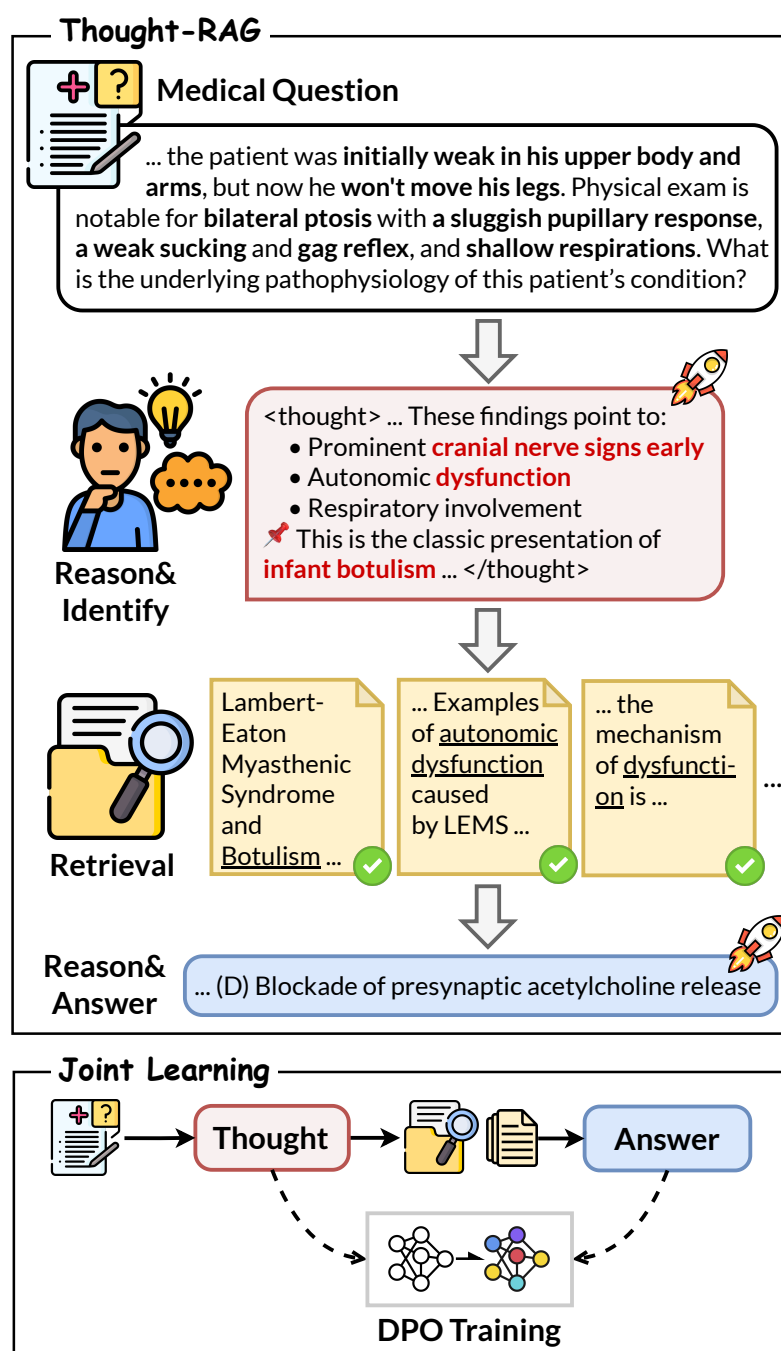


Figure 6.1: Thought-RAG for medical reasoning. Our method jointly optimizes the generation of thought processes and the subsequent retrieval-augmented answer generation.

identify their implicit knowledge requirements. This thought process is directly used to retrieve relevant information, which is subsequently incorporated into reasoning to derive the final answer. To optimize both thought and answer generation, we propose a joint learning framework, **RAR**², that simultaneously improves **Reasoning-Augmented Retrieval** and **Retrieval-Augmented Reasoning**. Specifically, we first construct a training dataset consisting of mixed preference pairs. One type is thought pairs, in which a sampled thought process is annotated based on the outcome of subsequent retrieval-augmented generation. The other type is answer pairs, in which a sampled answer is annotated according to its correctness. We then apply Direct Preference Optimization (DPO) to fine-tune the LLM [92]. These preference pairs enable supervised preference learning, allowing the model to identify relevant knowledge and reason effectively with external information. Extensive experiments across several biomedical question answering datasets demonstrate RAR²'s superiority over existing RAG baselines. Our further test-time scaling analysis validates the scalability of RAR².

In summary, our contributions are as follows:

- We propose a joint learning framework, RAR², that simultaneously improves reasoning-augmented retrieval and retrieval-augmented reasoning. Additionally, we design two test-time scaling strategies to explore the boundaries of our framework.
- We construct a mixed preference dataset to train the LLM to identify implicit knowledge needs and reason effectively with external information.
- Experimental results demonstrate the effectiveness of RAR² in six biomedical question answering datasets. Our framework outperforms baseline under tuning-free and fine-tuned settings.

6.2 Preliminary

6.2.1 Problem Formulation

In this work, we focus on LLM-based medical reasoning for medical question answering. Given a question q and a large medical corpus $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$, a set of top-relevant documents $\mathcal{D}_k$ is retrieved for q , where n is the total number of documents in the corpus and k is the number of retrieved ones. Then, the question and the retrieved documents are used as context for an LLM $\mathcal{M}$ to generate an answer y by step-by-step reasoning. Our goal is to jointly optimize document retrieval and retrieval-augmented medical reasoning.

6.2.2 Medical Corpus

A medical corpus is crucial for retrieval-augmented medical reasoning, as it provides the external knowledge necessary for LLMs to reason effectively about a given question. In this work, we adopt MedCorp, a large-scale medical corpus proposed in [135]. MedCorp integrates four major sources: PubMed¹, StatPearls², medical textbooks [46], and Wikipedia. It comprises 30.4M documents, including clinical guidelines, peer-reviewed research articles, and medical encyclopedic content. To process the documents, we reuse the original chunking strategy provided with the corpus and denote each document chunk as $d_i \in \mathcal{D}$, where $\mathcal{D}$ is the set of all chunks.

6.3 Reasoning Before Retrieval

We begin with a preliminary study on the impact of reasoning prior to retrieval. Rather than fine-tuning a model to generate or refine a query, we instead sample a thought

¹<https://pubmed.ncbi.nlm.nih.gov/>

²<https://www.statpearls.com/>

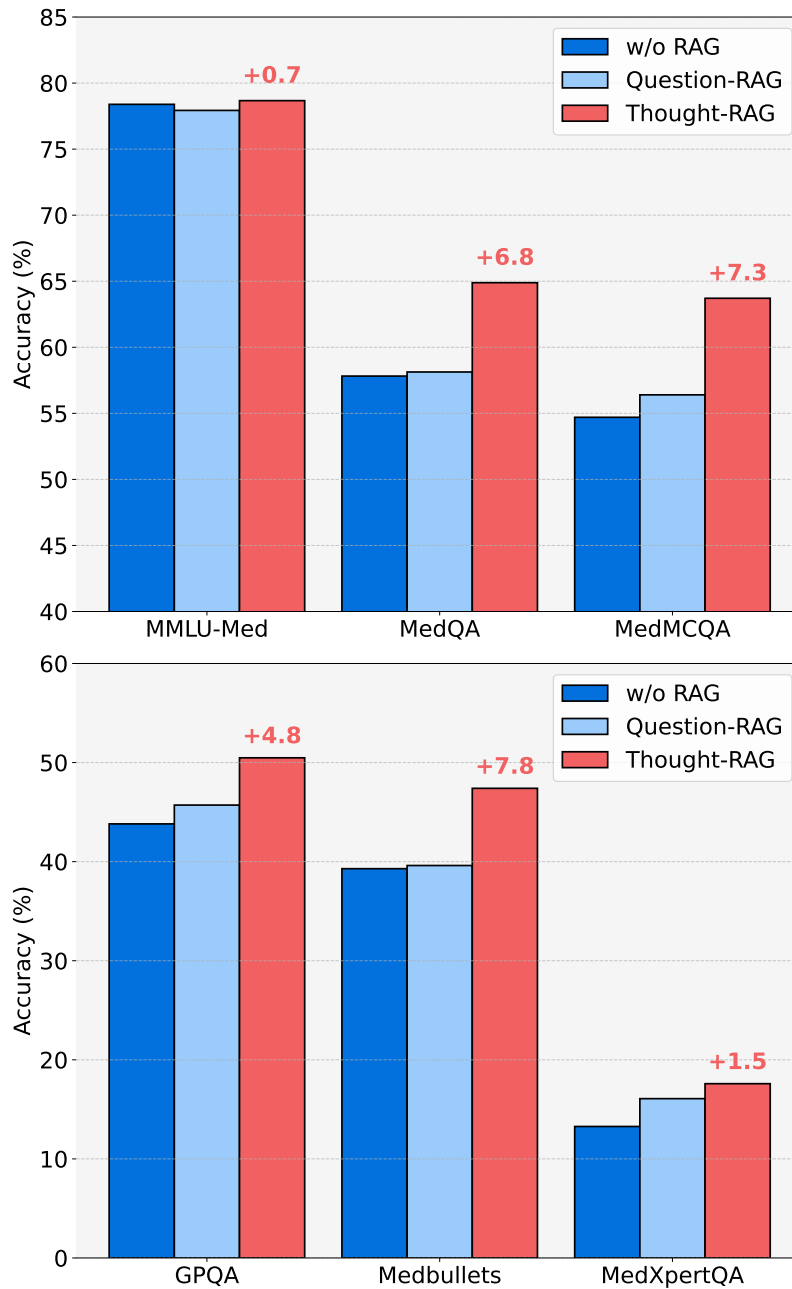


Figure 6.2: The performance comparison between Question-RAG and Thought-RAG on several medical question answering datasets.

process that could guide the model toward a final answer and use it directly as a query to retrieve relevant documents. The thought process is sampled in a zero-shot manner by prompting the LLM with: *“Please reason step by step to identify the relevant*

knowledge that may be involved.” For retrieval, we use the BM25 algorithm [97] to retrieve the top- k documents from the corpus. These documents are then provided as context to the LLM for answer generation. We evaluate this approach—referred to as Thought-RAG—against standard question-based retrieval (Question-RAG) using the Qwen2.5-7B-Instruct model [145] with $k = 32$.

Figure 6.2 shows the performance comparison on several medical question answering datasets. We can observe that Thought-RAG consistently outperforms Question-RAG, with an average improvement of 4.82% across all datasets. Especially in challenging datasets like Medbullets and GPQA, the gains are more significant. Furthermore, for most datasets, the accuracy remains unchanged or even declines using Question-RAG. Similar phenomenon can be found in the MIRAGE benchmark [135]. These observations suggest that reasoning-intensive medical questions often do not explicitly state the specific information needs required for successful retrieval. In contrast, a thought process generated through prior reasoning more accurately captures the underlying knowledge requirements, thereby improving overall RAG performance. Reasoning first and then using the resulting thought process for retrieval is a straightforward yet powerful approach to collect relevant knowledge, but it has received limited attention within the RAG community [107]. Therefore, further optimizing the generation of the thought process holds promise for enhancing retrieval quality and improving the RAG performance.

6.4 Method

In this section, we introduce the RAR² framework, which constructs a mixed preference dataset consisting of thought and answer pairs, and applies DPO to jointly enhance both reasoning-augmented retrieval and retrieval-augmented reasoning for LLMs. As shown in Figure 6.3, we first sample a set of thought processes and answers. The thought processes are generated based on the medical question, while the answers are

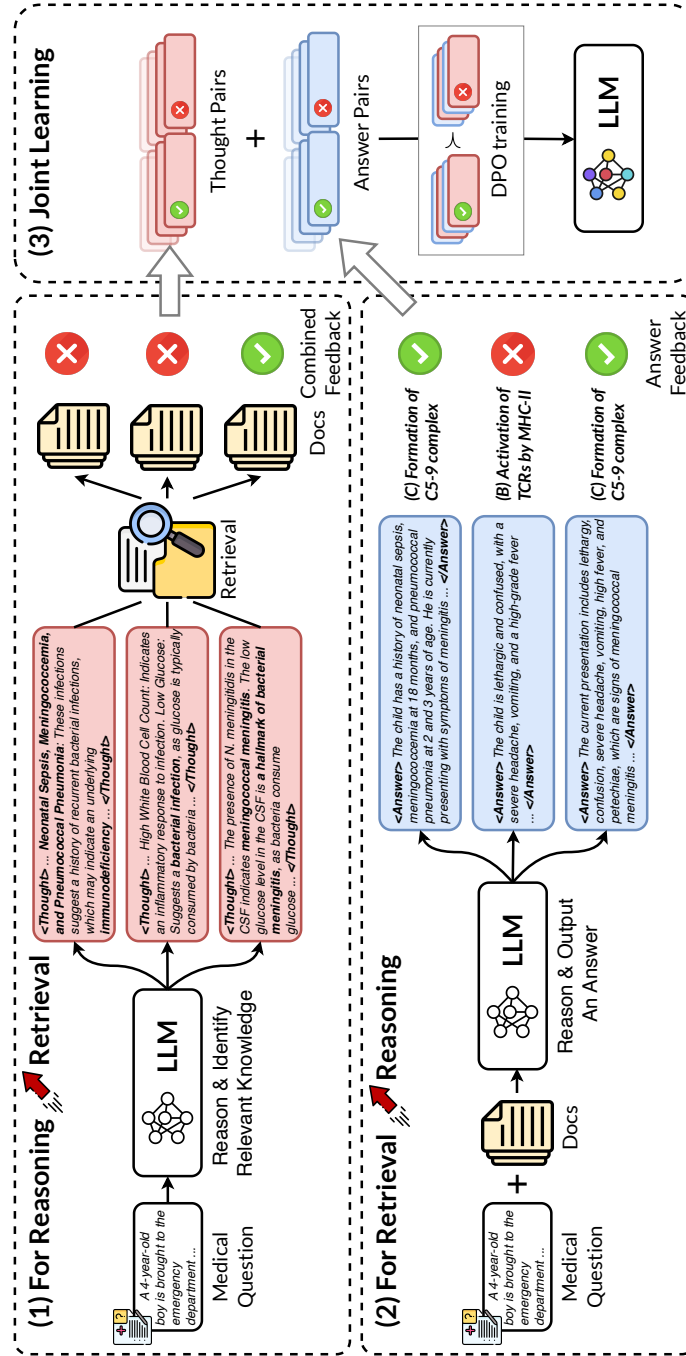


Figure 6.3: Construction of mixed preference pairs. (1) We sample several thought processes and append each to the question to form a query for retrieval. Each thought is then annotated based on whether it leads to a correct answer in subsequent RAG. (2) We sample several answers and annotate each one according to its correctness.

generated using the question along with the corresponding retrieved documents. All samples are annotated based on the correctness of the answer (§6.4.1). We collect samples to form a mixed preference dataset and apply DPO to jointly optimize two types of reasoning processes (§6.4.2). Besides, we design two test-time scaling strategies to examine the boundaries of RAR² (§6.4.3).

6.4.1 Construction of Mixed Preference Pairs

Our goal is to formulate and optimize two types of reasoning processes: one for retrieval and the other for retrieval-augmented generation. Previous work often overlooks the reasoning process required to investigate the underlying knowledge requirements, and few studies focus on jointly optimizing both retrieval and generation. To address this, we design distinct preference pairs tailored to each type of reasoning process.

Thought Pairs. We use an instruction-tuned LLM as the base model to sample thought processes. Given a medical question q , we first prompt the model with “*Please reason step by step to identify the relevant knowledge that may be involved.*” to sample several thought processes:

$$y^t \sim \mathcal{M}(q, \text{Prompt}_t). \quad (6.1)$$

We then annotate them as chosen and rejected based on two criteria. The first criterion appends the prompt “*The answer is:* ” to the end of each thought process and lets the LLM complete the answer. The correctness of the result is denoted as $g^{\text{direct}} \in \{0, 1\}$. The second criterion uses the thought process as a query to retrieve the top- k documents, which are then used to generate an answer. The correctness of this result is denoted as $g^{\text{rag}} \in \{0, 1\}$. To ensure deterministic outputs, we set the temperature to 0 during annotation. The thought samples are labeled as chosen only if both $g^{\text{direct}} = 1$ and $g^{\text{rag}} = 1$. Such annotation ensures that the thought process can lead to a correct answer. We collect one chosen and one rejected thought process to

form a preference pair, denoted as (y^{t+}, y^{t-}) .

Answer Pairs. We use the same base model to sample answer candidates. Given a medical question q and a thought process y^t , we first retrieve top- k documents $\mathcal{D}_k$ from the corpus. Then, we sample a group of answers based on the question and each retrieved document set:

$$y^a \sim \mathcal{M}(q, \mathcal{D}_k, \text{Prompt}_a), \quad (6.2)$$

where Prompt_a is “*Please reason step by step and choose one option from the above*”. The sample is labeled as chosen if its answer is correct. We collect one chosen and one rejected answer to form a preference pair, denoted as (y^{a+}, y^{a-}) .

6.4.2 Joint Learning

After collecting the thought and answer pairs, we construct a mixed preference dataset $\mathcal{Y} = \{(y^{t+}, y^{t-})\} \cup \{(y^{a+}, y^{a-})\}$. We then apply DPO to jointly optimize both types of reasoning processes. Joint training is adopted because the two processes are complementary and can benefit from each other: reasoning to identify relevant knowledge facilitates answer generation, while answer generation, in turn, helps refine the analytical quality of the thought process. Previous studies fail to optimize these reasoning processes or consider their complementary relationship.

For DPO training, we thoroughly shuffle all preference pairs and employ the DPO loss as follows:

$$\mathcal{L} = -\mathbb{E}_{(q, y^+, y^-) \sim \mathcal{Y}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y^+|q)}{\pi_{\mathcal{M}}(y^+|q)} - \beta \log \frac{\pi_\theta(y^-|q)}{\pi_{\mathcal{M}}(y^-|q)} \right) \right], \quad (6.3)$$

where σ is the sigmoid function, π_θ is the policy model, $\pi_{\mathcal{M}}$ is the reference model, and β is the hyperparameter that regulates the extent of deviation from the reference model.

6.4.3 Test-Time Scaling

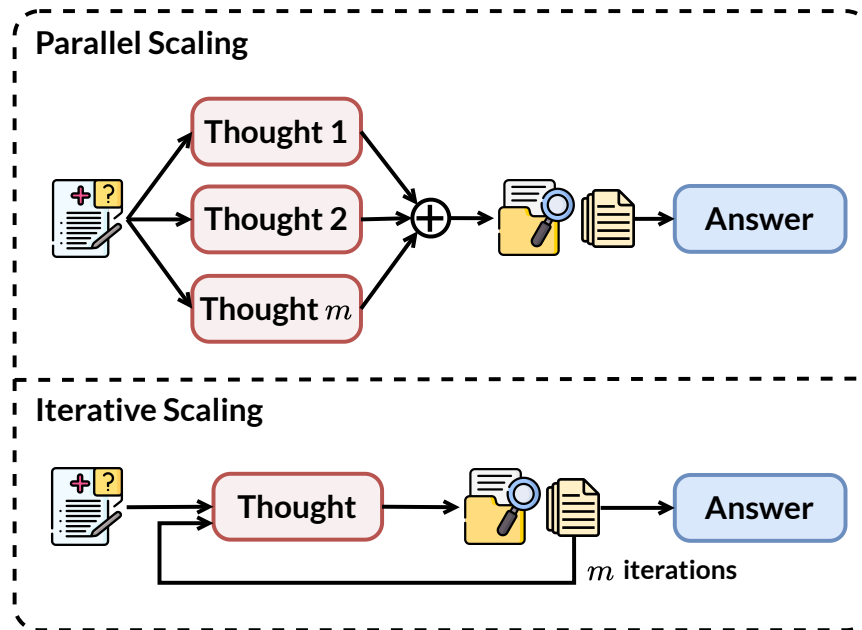


Figure 6.4: Frameworks of test-time scaling. Parallel scaling concatenates multiple thought processes to retrieve documents, while iterative scaling uses the retrieved documents in sequence to enhance thought generation.

To investigate the test-time scaling effect of RAR², we design two scaling strategies, as illustrated in Figure 6.4. Previous studies have paid little attention to the continual improvement of medical reasoning within the RAG framework.

The first strategy (**Paralleling Scaling**) generates m thought processes in parallel and concatenates them to form a single extended thought process. This combined thought is then used to retrieve documents, based on which an answer is generated. The second strategy (**Iterative Scaling**) generates a thought process and retrieves documents, which are then appended to the prompt for the next round of thought generation. This iterative process is repeated m times, after which an answer is generated. We design these two strategies to explore how RAR² scales with the number of generated thought processes. The first strategy merges more useful information to retrieve documents, while the second exploits the retrieved documents to generate more accurate thought

Datasets	#Sample
MedQA	1273
MedMCQA	4183
MMLU-Med	1089
GPQA	105
Medbullets	308
MedXpertQA	2450

Table 6.1: The number of samples for each evaluation dataset

processes.

6.5 Experiments

6.5.1 Evaluation Datasets

We evaluate our method on six biomedical question answering datasets: MedQA, MedMCQA [83], MMLU-Med (MMLU.) [37], Medbullets [10], GPQA [94], and MedXpertQA [162]. Among these, MedQA is the in-domain dataset, as our training data is derived from it. The remaining datasets are considered out-of-domain. The number of answer options across datasets ranges from 4 to 10. Notably, Medbullets, GPQA, and MedXpertQA are more recent and consist of graduate- or expert-level problems. Overall, the diversity of these datasets enables a robust evaluation of the model’s medical reasoning capabilities.

Methods	MedQA [†]	MedMCQA	MMLU	GPQA	Medbullets	MedXpertQA	Avg.
<i>No RAG</i>							
Qwen2.5-7B	57.82	54.70	78.39	43.81	39.29	13.27	47.20
m1-7B	64.34	59.34	78.02	36.19	48.38	16.29	50.43
HuatuoGPT-o1	68.81	64.95	79.87	45.71	50.65	14.98	54.17
<i>Tuning-Free RAG</i>							
MedRAG	54.20	52.35	75.81	42.86	39.29	14.20	46.45
i-MedRAG	62.84	<u>55.18</u>	79.87	<u>51.43</u>	<u>44.81</u>	<u>16.37</u>	51.75
Med-R ²	81.06	49.27	72.39	-	-	-	-
RAR ² (w/o train)	<u>64.89</u>	63.71	<u>78.67</u>	<u>50.48</u>	47.40	17.59	54.27

Table 6.2: Comparisons of RAR² with other medical large language models and tuning-free RAG methods. [†] denotes the in-domain dataset.

6.5.2 Baseline Methods

We compare our framework with six strong baselines in medical reasoning. Qwen2.5-7B-Instruct [145] is used as our base model. Two recent models enhanced for medical reasoning that do not use RAG are included: m1-7B-23K [41] and HuatuoGPT-o1 [12]. For fair comparison, we use HuatuoGPT-o1-7B with the same model size. Additionally, we include three tuning-free RAG methods that have been applied to medical tasks: MedRAG [135], i-MedRAG [135], and Med-R² [73]. Med-R² is developed through a few-shot generation strategy. Finally, we include four RAG methods that involve fine-tuning on specific components: Self-BioRAG [44], SimRAG-8B [142], RAG²-7B [73], and SPO [21]. Baselines with publicly released models are reimplemented for comparison.

Methods	MedQA [†]	MedMCQA	MMLU.
Self-BioRAG	43.60	42.15	53.92
SimRAG-8B	62.92	67.51	75.57
RAG ² -7B	75.64	63.04	78.67
SPO	76.98	71.08	<u>85.49</u>
RAR ² (w/ train)	<u>76.43</u>	<u>65.69</u>	86.32

Table 6.3: Comparisons of RAR² with other RAG methods that involve fine-tuned components. [†] denotes the in-domain dataset.

6.5.3 Implementation Details

We utilize the MedQA training set [45] to implement our method, which comprises 10K medical problems derived from professional medical board exams. Each problem requires selecting the correct answer from four options. Our base model for sampling preference pairs is Qwen2.5-7B-Instruct [145]. We set the number of thought sampling attempts to 15 and the number of answer sampling attempts to 5. The temperature of sampling is set 0.2 and the top- p is set 0.9. For each question, we retrieve the top 32 document chunks using the BM25 algorithm [97] with parameters $k_1 = 1.2$ and $b = 0.75$ as context for sampling solutions. A total of 12K preference pairs are collected after filtering and selection, where 4K are thought pairs and 8K are answer pairs. The model is then trained for 4 epochs with a global batch size of 64 and a learning rate of 1e-6, while the parameter β for the DPO loss is set to 0.2. All experiments are conducted on eight A100 GPUs, and we employ DeepSpeed ZeRO3 to optimize memory usage.

6.5.4 Main Results

We report the main results on six biomedical datasets shown in Table 6.2 and Table 6.3. We calculate the average performance across several datasets as **Avg.** in the final column. We present the results of RAR² under two settings: **w/o train** and **w/ train**. The former uses the base model for inference, while the latter uses the DPO-tuned model.

Table 6.2 presents the results in comparison with medically enhanced LLMs and tuning-free RAG methods. Our framework significantly improves the medical reasoning capabilities of the LLM over its backbone model, Qwen2.5-7B. It outperforms the backbone on all six datasets, achieving an average accuracy gain of 7.07%. Additionally, RAR² performs competitively with state-of-the-art medical language models: it surpasses m1-7B and achieves comparable average accuracy to HuatuoGPT-o1. Notably, on challenging MedXpertQA, RAR² achieves accuracy gains of 1.3% and 2.61% over m1-7B and HuatuoGPT-o1, respectively. Compared to other tuning-free RAG methods, RAR² also achieves higher average accuracy. Specifically, it outperforms i-MedRAG, which uses iterative query refinement, and Med-R², which employs a more complex retrieval pipeline.

Table 6.3 presents the results in comparison with other RAG methods that are applied to medical tasks and involve fine-tuned components. Our RAR² outperforms most baseline methods. Notably, on MMLU-Med, it achieves the highest accuracy among all models. SPO performs well on MedQA and MedMCQA, which may be attributed to its use of larger and unopen-source medical knowledge sources and additional test sample selection. Overall, the results on both in-domain and out-of-domain datasets demonstrate that our framework can help general LLMs consistently improve their medical reasoning abilities for both retrieval and answer generation.

Methods	MedQA	MedMCQA	MMLU.	GPQA	Medbullets	MedXpertQA	Avg.
RAR ²	76.43	65.69	86.32	56.19	57.14	20.98	60.46
- w/o training	64.89	63.71	78.67	50.48	47.40	17.59	54.27
- w/o answer	74.39	63.88	83.75	53.33	53.57	18.90	57.97
- w/o thought	74.86	64.79	84.85	54.29	55.19	19.51	58.92

Table 6.4: Ablation study on several medical question answering datasets.

6.5.5 Ablation Studies

We demonstrate the effectiveness of RAR² under different training settings, as detailed below: (1) **w/o training**, which adopts the thought-based RAG framework without any fine-tuning; (2) **w/o answer**, which removes answer pairs and optimizes only the generation of thought processes; and (3) **w/o thought**, which removes thought pairs and optimizes only the generation of answers. The results are shown in Table 6.4.

As shown in the table, RAR² achieves the best performance across all three selected datasets, demonstrating the effectiveness of our proposed method. Compared to the setting without training, both *w/o answer* and *w/o thought* yield significant performance improvements, indicating that optimizing the generation of either thought processes or answers is crucial for RAR²’s effectiveness. More importantly, optimizing both types of preference pairs yields complementary gains, enhancing the performance of both reasoning-augmented retrieval (i.e., thought generation) and retrieval-augmented reasoning (i.e., answer generation).

We also investigate the impact of the number of retrieved documents on the performance of RAR². As shown in Figure 6.5, RAR² achieves its best overall performance when retrieving 32 documents, although the optimal number varies slightly across different datasets.

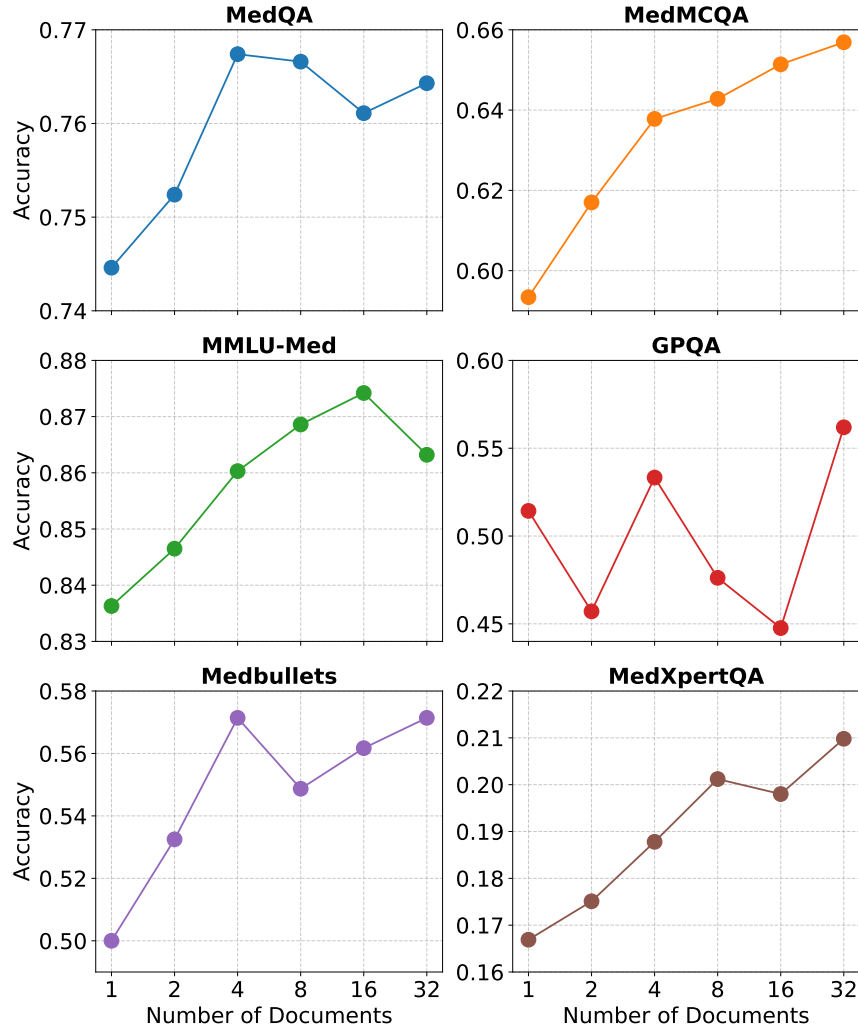


Figure 6.5: The performance comparison with a different number of retrieved documents.

6.5.6 Impact of Test-Time Scaling

We investigate the impact of test-time scaling on RAR^2 , as shown in Figure 6.6 and Figure 6.7. The total number of generated thought processes is scaled from 1 to 8. For parallel scaling, the sampling temperature is set to 1.0, and the top- p value is also set to 1.0. For iterative scaling, the number of thoughts depends on the number of iterations.

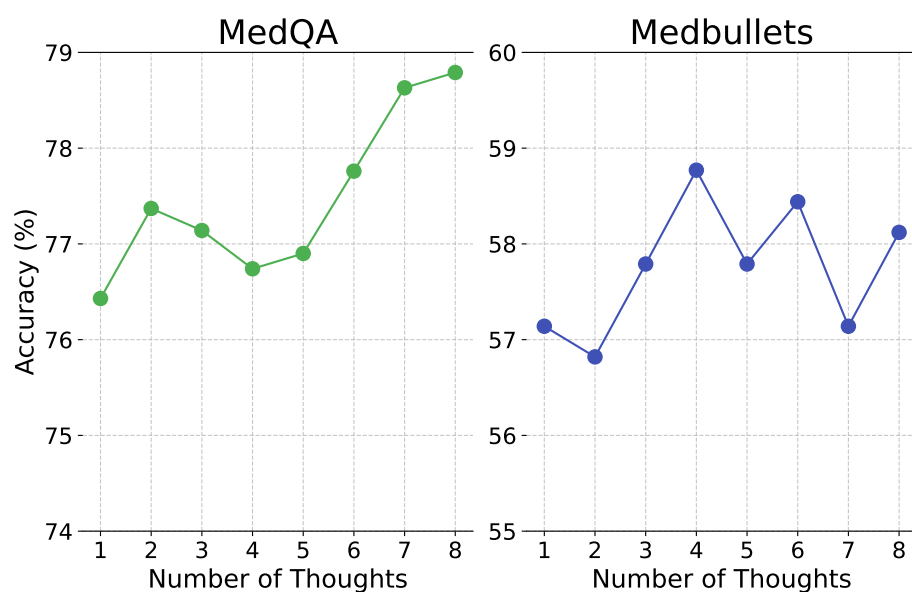


Figure 6.6: Accuracies with the Parallel Scaling strategy on the MedQA and Medbullets datasets.

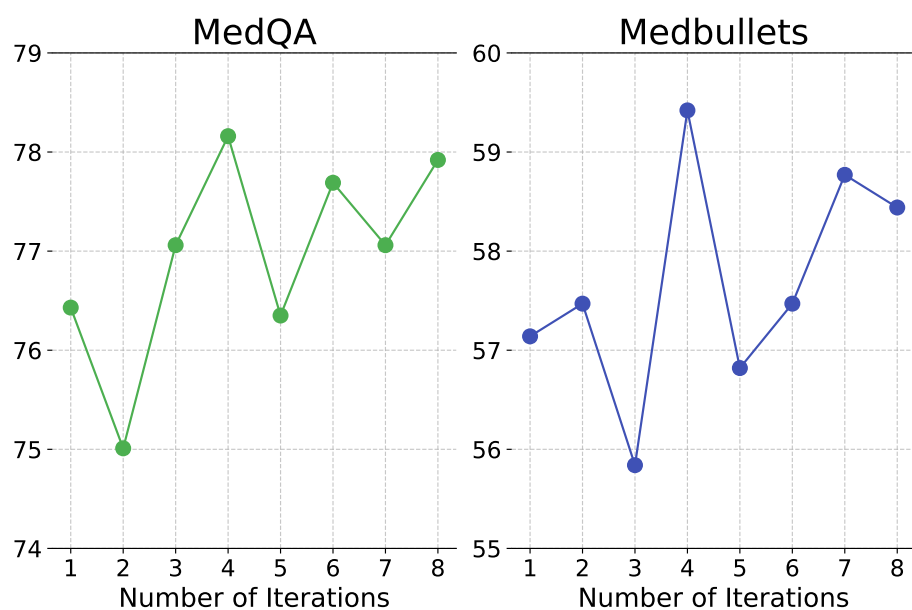


Figure 6.7: Accuracies with the Iterative Scaling strategy on the MedQA and Medbullets datasets.

Figure 6.6 and 6.7 shows the results with increasing iterations on the MedQA and Medbullets datasets, where one is the in-domain dataset and the other is a challenging out-of-domain dataset. For Parallel Scaling, accuracy shows an upward trend as the number of thought processes increases. The improvement is particularly stable on MedQA. A similar effect is observed in Iterative Scaling, where accuracy generally increases with the number of iterations, achieving a maximum improvement of approximately 2% on both datasets.

6.6 Chapter Summary

In this chapter, we propose RAR², a novel joint learning framework designed to enhance LLM reasoning capabilities within the RAG framework, specifically targeting complex medical questions. Unlike existing RAG approaches, which rely primarily on direct query refinement, RAR² explicitly constructs and optimizes a thought-based reasoning process to uncover implicit knowledge requirements. We introduce a mixed preference dataset comprising thought pairs and answer pairs and leverage DPO for joint training. Extensive experiments on multiple biomedical question answering datasets demonstrate that RAR² outperforms state-of-the-art RAG baselines, both with and without fine-tuning. Moreover, our analysis of test-time scaling strategies validates the scalability and robustness of RAR², highlighting its potential to significantly improve medical reasoning and decision-making in clinical scenarios.

Chapter 7

Conclusions and Future Work

This thesis advances the development of clinical process-aware medical dialogue systems, aiming to align with clinical practice with respect to accuracy, reliability, and safety. We articulate and address three primary research questions: (1) How to model the critical information flow (e.g., medical entities and dialogue acts) in medical dialogue and incorporate key entities and acts into response generation; (2) How to develop a comprehensive differential diagnosis (DDx) and leverage it to improve the reliability and interpretability of system responses; (3) How to emulate clinicians' internal reasoning processes and improve a system's reasoning capabilities with external medical knowledge. By examining dialogue structure, differential diagnosis, and internal reasoning, we tackle core challenges in enabling medical dialogue systems to emulate essential aspects of the clinical process.

Taken together, our studies validate the proposed methods and highlight their promise for building clinical process-aware medical dialogue systems. Aligning system behavior with key stages of the clinical process yields gains in accuracy, reliability, and safety, thereby increasing their viability as medical assistants or virtual doctors. Designing systems to emulate clinical processes not only satisfies real-world operational needs but also establishes a reference point for the development of clinician-like dialogue

systems.

7.1 Summary of Contributions

The following sections summarize the main contributions of this thesis.

7.1.1 Dual Flow Enhanced Medical Dialogue Generation

- We propose a novel MDS framework, DFMed, which models the transitions of medical entities and dialogue acts via step-by-step interweaving.
- We summarize the dialogue acts in the medical consultation scenario grounded on the medical documentation standards, SOAP notes, including make a diagnosis, prescribe medications, etc.
- Experimental results show the superiority of DFMed over the previous frameworks and demonstrate the effectiveness of introducing the medical entity and dialogue act flows.

7.1.2 Differential Diagnosis Guided Medical Dialogue System

- We propose a medical dialogue generation framework, IADDx, which explicitly models a differential diagnosis with intuitive-then-analytic processes and incorporates diagnosis to guide response generation.
- We build a multi-disease classifier with multi-task learning of affected systems and organs. A disease knowledge graph composed of systems, organs, diseases, and symptoms is adopted to interpret the diagnostic process.
- We conduct extensive experiments on two medical dialogue datasets, demonstrating the effectiveness and interpretability of IADDx.

7.1.3 Diagnostic Reasoning Aligned Medical Dialogue System

- We propose a novel medical dialogue system framework, EMULATION, that emulates clinicians’ diagnostic reasoning processes and aligns with clinician preferences. This is the first work that explores clinician preferences and internal thought processes.
- We build a diagnostic thought process dataset that is employed to align response generation with clinician preferences.
- We propose a joint learning framework, RAR², that simultaneously improves reasoning-augmented retrieval and retrieval-augmented reasoning. Additionally, we design two test-time scaling strategies to explore the boundaries of our framework.
- We construct a mixed preference dataset to train the LLM to identify implicit knowledge needs and reason effectively with external information.
- Experimental results demonstrate the effectiveness of EMULATION and RAR² in several medical dialogue and medical question answering datasets. Our framework outperforms baseline in conducting diagnostic reasoning with or without external medical knowledge.

7.2 Safety Considerations

Automated medical reasoning systems present several distinct safety hazards that must be named and managed explicitly. Clinical errors can manifest as incorrect or delayed diagnoses, unsafe dosing (e.g., unit-conversion mistakes or dose extrapolation outside labeled ranges), and guideline violations in complex, co-morbid cases—each with direct potential for patient harm. Overconfidence and automation bias pose a second hazard: models may produce fluent, confident outputs that are wrong,

and clinicians may overweight such outputs, especially when confidence is poorly calibrated or uncertainty is hidden in the interface. Third, data issues can compromise validity: label noise from heuristic coding or inconsistent chart abstraction; shortcut learning where proxies (e.g., device brand, room location) stand in for true physiology; spurious correlations tied to workflow artifacts; and domain/temporal shift across sites, scanners/devices, populations, or time, all of which degrade performance when deployment conditions diverge from training. Fourth, equity risks arise when error rates or calibration vary across age, sex, race/ethnicity, language proficiency, socioeconomic status, or rare-disease cohorts, potentially entrenching disparities despite strong average metrics. Fifth, security and privacy threats—including data leakage through logs, membership/model inversion attacks that reconstruct sensitive attributes, and exposure via third-party integrations—can erode trust and regulatory compliance. Finally, for generative components, hallucination and fabrication may yield unsupported clinical claims, invented citations, or synthesized patient details that sound plausible but lack evidentiary grounding.

Systems in this thesis are designed strictly as clinical decision support for licensed healthcare professionals; it augments—not replaces—clinical judgment, and it must not be used for autonomous diagnosis, prescribing, or patient communication. Its intended use is limited to providing evidence-linked differentials, risk estimates, and checklist-style safety reminders for patients who match the training and validation populations and data modalities described in this thesis. The interface requires clinician review and acknowledgement before any recommendation can affect care, and it presents confidence, rationale, and source provenance to support human verification. To prevent misuse, the system enforces hard boundaries: requests that are out of scope—such as patient cohorts not represented in training (e.g., pediatrics if trained on adults), care settings or devices outside validation, language/format mismatches, missing critical variables, suspected data corruption, or tasks explicitly excluded (e.g., emergent triage, dose calculation beyond labeled ranges)—trigger an abstention with

a clear explanation and route to escalation (e.g., specialist review, pharmacist check, or alternative workflow). Soft boundaries apply when uncertainty or distribution-shift detectors indicate elevated risk; in such cases the system de-emphasizes outputs, highlights caveats, and requires stronger human confirmation. All blocks and escalations are logged for audit, and boundary criteria are reviewed periodically as datasets, indications, and regulatory guidance evolve.

7.3 Future Work

While this thesis establishes key groundwork for clinical process-aware medical dialogue systems, important directions remain open for further refinement and expansion:

- **Multimodal and Contextual Understanding:** One important frontier is extending medical dialogue systems beyond text into multimodal domains. Clinical interactions are inherently multimodal: a doctor draws not only on spoken language but also on lab results, medical images, vital signs, and the patient’s electronic health records (EHRs) to make decisions. Currently, most MDSs handle only textual input (e.g., typed symptom descriptions or chat transcripts). This limitation can lead to suboptimal or incorrect assessments; for example, a chatbot managing a dermatology query may misdiagnose a rash if it cannot access an image of the lesion. Future dialogue agents should be able to process and reason over clinical images (e.g., X-rays, CT/MRI scans, photographs) and other data such as lab values and medication histories, in addition to natural language. Incorporating these sources could substantially improve diagnostic accuracy and scope. Recent surveys highlight multimodal medical dialogue as an emerging challenge, noting that missing visual information can impede correct diagnosis. Accordingly, research is needed on architectures that fuse textual context with visual and structured data, for example by

combining language models with computer vision models and modules that handle structured EHR data.

- **Personalization and Adaptive Interaction:** Every patient is unique, and every clinician has a distinct communication style, so a one-size-fits-all dialogue system is unlikely to serve all users optimally. Adaptive personalization is an emerging direction for tailoring medical dialogues to individual needs. Current healthcare conversational agents implement simple personalization features (e.g., using a patient’s name or remembering past interactions), but these are often rudimentary and not grounded in robust frameworks. A review study shows that many healthcare chatbots included personalized content such as tailored feedback or reminders, which improved user satisfaction and engagement, yet these features were added ad hoc without theoretical guidance and with limited evaluation of their impact [50]. Going forward, personalization should be treated as a first-class design goal. This entails developing methods for user modeling in medical dialogues (for example, estimating a patient’s health literacy, emotional state, or linguistic preferences) and dynamically adjusting responses accordingly. A patient-facing MDS could simplify its language and increase empathy for an anxious, low-literacy user, whereas the same system might use more technical language and faster pacing for a health-savvy user. Similarly, a clinician-facing assistant might adapt to a physician’s specialty or workflow, providing more detailed explanations for junior clinicians and concise summaries for experts, and formatting outputs to individual preferences (e.g., SOAP notes).
- **Integration into Human–AI Collaboration:** Human–AI collaboration paradigms will be critical to study in this setting. Rather than treating the MDS as an autonomous decision-maker, future research is likely to frame it as a cooperative tool that works in tandem with healthcare professionals. This entails defining how control and responsibility are shared between human and system. For example, an AI might draft a clinical plan that the physician reviews and

edits, analogous to a junior trainee writing a note that a senior doctor finalizes. Determining the optimal division of labor (what the AI should handle versus what the human should handle) remains an open question that requires input from clinicians. Early real-world deployments underscore the importance of user-centered design: if an AI tool is not intuitive or adds steps to a clinician’s workflow, it will likely be ignored or rejected. Therefore, future MDS development should involve iterative co-design with clinicians, incorporating feedback to ensure the system genuinely streamlines tasks. For patient-facing systems, workflow integration means aligning the chatbot with the care continuum, such as linking a triage bot to appointment systems so it can schedule follow-up visits, or enabling a support bot to share conversation summaries with the patient’s doctor to keep the care team informed. In sum, the future of MDSs is not standalone apps in isolation but embedded, interoperable agents that operate within healthcare environments and alongside humans.

- **Improving Transparency, Interpretability, and Explainability:** As generative models play a larger role in healthcare, their operation must become more transparent and interpretable. Clinicians and patients must be able to understand and trust the reasoning behind AI-provided answers. A near-term priority is to build systems that explain their recommendations in human-understandable ways, encompassing both post hoc explainability (providing reasons or evidence for a given response) and intrinsic interpretability (designing models that reason in structured, inspectable ways). Regulations and best practices increasingly hold that clinicians should have a sufficient grasp of the medically relevant factors underlying a model’s output before using AI advice. Accordingly, future dialogue models will likely include features that elucidate their decision processes. Technically, strategies such as chain-of-thought prompting are an active area of research. Simply asking a model to explain itself can produce a rationale, but naive explanations may be fabricated or biased. Additional work is needed

on interpretability tools, such as methods to visualize which parts of the input influenced the answer or to trace how patient data are used within a model’s internal layers. Explaining large neural models remains challenging given their complexity and stochastic behavior, but progress here is essential. Ultimately, explainability is a requirement for responsible deployment: transparent and interpretable dialogue systems will enable clinicians to validate AI recommendations, foster patient trust, and meet regulatory expectations for accountability in AI-driven healthcare.

References

- [1] Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. Medhalu: Hallucinations in responses to healthcare queries by large language models. *CoRR*, abs/2409.19492, 2024.
- [2] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626, 2025.
- [3] Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023.
- [5] Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie

- Peng, Xuanjing Huang, and Zhongyu Wei. Disc-medllm: Bridging general large language models and real-world medical consultation. *CoRR*, abs/2308.14346, 2023.
- [6] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju S. Patel, Oluseyi Fayanju, Shreya J. Shah, Ethan Goh, Dong-han Yao, Brian Soetikno, Eduardo Pontes Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle A. Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, François Grolleau, Kameron C. Black, Bethel Mieso, Aydin Zahedivash, Wen-wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan H. Chen, Emily Alsentzer, Keith E. Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. Medhelm: Holistic evaluation of large language models for medical tasks. *CoRR*, abs/2505.23802, 2025.
- [7] Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021–e2526021, 2025.
- [8] Susan Cameron and Imani Turtle-Song. Learning to write case notes using the soap format. *Journal of Counseling & Development*, 80(3):286–292, 2002.
- [9] Hyunjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin

- Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5606–5632. Association for Computational Linguistics, 2023.
- [10] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *CoRR*, abs/2402.18060, 2024.
- [11] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *CoRR*, abs/2402.18060, 2024.
- [12] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuoogpt-o1, towards medical complex reasoning with llms. *CoRR*, abs/2412.18925, 2024.
- [13] Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 4432–4440. AAAI Press, 2022.
- [14] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuoogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *CoRR*, abs/2406.19280, 2024.

-
- [15] Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. Huatuogpt-ii, one-stage training for medical adaption of llms. *CoRR*, abs/2311.09774, 2023.
- [16] Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280, 2025.
- [17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.
- [18] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 12 2022. btac817.
- [19] Xi Chen, Huahui Yi, Mingke You, Weizhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digit. Medicine*, 8(1), 2025.
- [20] Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 5134–5143. Computer Vision Foundation / IEEE, 2025.

- [21] Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications. *CoRR*, abs/2501.02460, 2025.
- [22] Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. Med42-v2: A suite of clinical llms. *CoRR*, abs/2408.06142, 2024.
- [23] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.
- [24] Pat Croskerry. A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028, 2009.
- [25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [26] Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. PlugMed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5050–5066, Singapore, December 2023. Association for Computational Linguistics.
- [27] Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

-
- [28] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [29] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [30] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR, 2023.
- [31] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [32] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. Medalpaca - an open-source collection of medical conversational AI models and training data. *CoRR*, abs/2304.08247, 2023.
- [33] Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. BP4ER: bootstrap prompting for explicit reasoning in medical dialogue generation. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and

- Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2480–2492. ELRA and ICCL, 2024.
- [34] Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. Dialmed: A dataset for dialogue-based medication recommendation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 721–733. International Committee on Computational Linguistics, 2022.
- [35] Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. DialMed: A dataset for dialogue-based medication recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 721–733, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [36] Stefan Hegselmann, Georg von Armin, Tillmann Rheude, Noel Kronenberg, David A. Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful EHR encoders. *CoRR*, abs/2502.17403, 2025.
- [37] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [38] Keith J Holyoak and Robert G Morrison. *The Cambridge handbook of thinking and reasoning*. Cambridge University Press, 2005.

-
- [39] Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1321–1335. Association for Computational Linguistics, 2024.
- [40] Zefa Hu, Haozhi Zhao, Yuanyuan Zhao, Shuang Xu, and Bo Xu. T-agent: A term-aware agent for medical dialogue generation. In *International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, June 30 - July 5, 2024*, pages 1–8. IEEE, 2024.
- [41] Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models, 2025.
- [42] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics, 2021.
- [43] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5):105:1–105:36, 2022.
- [44] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024.
- [45] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter

- Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081, 2020.
- [46] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081, 2020.
- [47] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics, 2019.
- [48] Hao-Cheng Kao, Kai-Fu Tang, and Edward Y. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2305–2313. AAAI Press, 2018.
- [49] Dennis Kasper, Anthony Fauci, Stephen Hauser, Dan Longo, J Jameson, and Joseph Loscalzo. *Harrison’s principles of internal medicine, 19e*. McGraw-hill New York, NY, USA:, 2015.
- [50] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11):e15360, 2019.

-
- [51] Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online, August 2021. Association for Computational Linguistics.
- [52] Suzanne Kurtz, Juliet Draper, and Jonathan Silverman. *Teaching and learning communication skills in medicine*. CRC press, 2017.
- [53] Wuyang Lan, Wenzheng Wang, Changwei Ji, Guoxing Yang, Yongbo Zhang, Xiaohong Liu, Song Wu, and Guangyu Wang. Clinicalgpt-r1: Pushing reasoning capability of generalist disease diagnosis with large language model. *CoRR*, abs/2504.09421, 2025.
- [54] Robert S Ledley and Lee B Lusted. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130(3366):9–21, 1959.
- [55] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020.
- [56] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

- [57] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [58] Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. Semi-supervised variational reasoning for medical dialogue generation. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 544–554. ACM, 2021.
- [59] Rumeng Li, Xun Wang, and Hong Yu. Llamacare: An instruction fine-tuned large language model for clinical NLP. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10632–10641. ELRA and ICCL, 2024.
- [60] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.
- [61] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang.

- Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.
- [62] Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. Rgar: Recurrence generation-augmented retrieval for factual-aware medical question answering, 2025.
- [63] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3):100943, 2024.
- [64] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [65] Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13362–13370. AAAI Press, 2021.
- [66] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *CoRR*, abs/2502.09838, 2025.
- [67] Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*

- Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [68] Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Min. Anal.*, 7(4):1116–1128, 2024.
- [69] Philip R. Liu, Sparsh Bansal, Jimmy Dinh, Aditya Pawar, Ramani Satishkumar, Shail Desai, Neeraj Gupta, Xin Wang, and Shu Hu. Medchat: A multi-agent framework for multimodal diagnosis with large language models. *CoRR*, abs/2506.07400, 2025.
- [70] Wenge Liu, Jianheng Tang, Xiaodan Liang, and Qingling Cai. Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. *Neurocomputing*, 442:260–268, 2021.
- [71] Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *ArXiv*, abs/2010.07497, 2020.
- [72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [73] Keer Lu, Zheng Liang, Da Pan, Shusen Zhang, Xin Wu, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. Med-r²: Crafting trustworthy LLM physicians through retrieval and reasoning of evidence-based medicine. *CoRR*, abs/2501.11885, 2025.

-
- [74] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6), 2022.
- [75] Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. *CoRR*, abs/2505.20099, 2025.
- [76] Douglas W Maynard and John Heritage. Conversation analysis, doctor–patient interaction and medical communication. *Medical education*, 39(4):428–435, 2005.
- [77] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, 2025.
- [78] Mark B Mengel, Warren Holleman, and Scott A Fields. *Fundamentals of clinical practice*. Springer Science & Business Media, 2002.
- [79] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Machine Learning for Health, ML4H@NeurIPS 2023, 10 December 2023, New Orleans, Louisiana, USA*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 2023.
- [80] Elizabeth Murray, Lance Pollack, Martha White, and Bernard Lo. Clinical decision-making: physicians’ preferences and experiences. *BMC Family Practice*, 8(1):1–10, 2007.

- [81] Usman Naseem, Surendrabikram Thapa, Qi Zhang, Liang Hu, and Mehwish Nasim. MDKG: graph-based medical knowledge-guided dialogue generation. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2164–2168. ACM, 2023.
- [82] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- [83] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H. Chen, Tom J. Pollard, Joyce C. Ho, and Tristan Naumann, editors, *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 2022.
- [84] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning, CoNLL 2023, Singapore, December 6-7, 2023*, pages 314–334. Association for Computational Linguistics, 2023.
- [85] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [86] Vimla L Patel and Marco F Ramoni. Cognitive models of directional inference in expert medical reasoning. 1997.

-
- [87] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *CoRR*, abs/2311.18681, 2023.
- [88] Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree, 2025.
- [89] Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5925–5941. Association for Computational Linguistics, 2023.
- [90] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *CoRR*, abs/2402.13963, 2024.
- [91] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [92] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [93] Justin T Reese, Leonardo Chimirri, Yasemin Bridges, Daniel Danis, J Harry

- Caufield, Kyran Wissink, Julie A McMurry, Adam SL Graefe, Elena Casiraghi, Giorgio Valentini, et al. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. *medRxiv*, 2024.
- [94] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023.
- [95] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.
- [96] Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149, 2025.
- [97] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.
- [98] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press, 2016.
- [99] Tianqi Shang, Weiqing He, Charles Zheng, Lingyao Li, Li Shen, and Bingxin

- Zhao. Dynamicare: A dynamic multi-agent framework for interactive and open-ended medical decision-making. *CoRR*, abs/2507.02616, 2025.
- [100] Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2840–2861. Association for Computational Linguistics, 2024.
- [101] Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. MidMed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [102] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics, 2021.
- [103] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics, 2021.

- [104] Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. *Skills for communicating with patients*. crc press, 2016.
- [105] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [106] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617, 2023.
- [107] Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. *CoRR*, abs/2411.00300, 2024.
- [108] Scott D.C. Stern, Adam S. Cifu, and Diane Altkorn. *Symptom to Diagnosis: An Evidence-Based Guide*, 4e. McGraw-Hill Education, 2020.
- [109] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Serkan Ö. Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *CoRR*, abs/2407.12883, 2024.

-
- [110] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.
- [111] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 599–621. Association for Computational Linguistics, 2024.
- [112] Arsène Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaétan Marceau-Caron, and Joumana Ghosn. Towards trustworthy automatic diagnosis systems by emulating doctors’ reasoning with deep reinforcement learning. In *NeurIPS*, 2022.
- [113] Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Disentangling reasoning and knowledge in medical large language models. *CoRR*, abs/2505.11462, 2025.
- [114] Abhisek Tiwari, Shreyangshu Bera, Preeti Verma, Jaithra Varma Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. Seeing is believing! towards knowledge-infused multi-modal medical dialogue generation. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

- and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 14513–14523. ELRA and ICCL, 2024.
- [115] Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*, abs/2305.12031, 2023.
- [116] Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. Conversational recommendation via hierarchical information modeling. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2201–2205. ACM, 2022.
- [117] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, Philip Andrew Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Sementurs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *CoRR*, abs/2307.14334, 2023.
- [118] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.
- [119] Deeksha Varshney, Niranshu Behera, Prajeet Katari, and Asif Ekbal. Medprom: Bridging dialogue gaps in healthcare with knowledge-enhanced generative models. *ACM Transactions on Computing for Healthcare*, 2025.

-
- [120] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [121] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, Da Pan, Fei Kou, Fei Li, Fuzhong Chen, Guosheng Dong, Han Liu, Hongda Zhang, Jin He, Jinjie Yang, Kangxi Wu, Kegeng Wu, Lei Su, Linlin Niu, Linzhuang Sun, Mang Wang, Pengcheng Fan, Qianli Shen, Rihui Xin, Shunya Dang, Songchi Zhou, Weipeng Chen, Wenjing Luo, Xin Chen, Xin Men, Xionghai Lin, Xuezhen Dong, Yan Zhang, Yifei Duan, Yuyan Zhou, Zhi Ma, and Zhiying Wu. Baichuan-m1: Pushing the medical capability of large language models, 2025.
- [122] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *CoRR*, abs/2306.09968, 2023.
- [123] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *CoRR*, abs/2304.06975, 2023.
- [124] Wenxuan Wang, Zizhan Ma, Meidan Ding, Shiyi Zheng, Shengyuan Liu, Jie Liu, Jiaming Ji, Wenting Chen, Xiang Li, Linlin Shen, and Yixuan Yuan. Medical reasoning in the era of llms: A systematic review of enhancement techniques and applications, 2025.
- [125] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International*

- Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [126] Yubo Wang, Xueguang Ma, and Wenhua Chen. Augmenting black-box llms with medical textbooks for biomedical question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1754–1770. Association for Computational Linguistics, 2024.
- [127] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [128] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*, 31(9):1833–1843, 2024.
- [129] Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards evaluating and building versatile large language models for medicine. *npj Digit. Medicine*, 8(1), 2025.
- [130] Jiageng Wu, Xian Wu, and Jie Yang. Guiding clinical reasoning with large language models via knowledge seeds. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 7491–7499. ijcai.org, 2024.
- [131] Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *CoRR*, abs/2403.06611, 2024.

-
- [132] Junde Wu, Jiayuan Zhu, and Yunli Qi. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. *CoRR*, abs/2408.04187, 2024.
- [133] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1062–1069. AAAI Press, 2020.
- [134] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Kuttichi Keloth, Xingyu Zhou, Huan He, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Me llama: Foundation large language models for medical applications. *CoRR*, abs/2402.12749, 2024.
- [135] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6233–6251. Association for Computational Linguistics, 2024.
- [136] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *CoRR*, abs/2304.01097, 2023.
- [137] Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. Re-thinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [138] Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6796–6814. Association for Computational Linguistics, 2024.
- [139] Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. Medical dialogue generation via dual flow modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6771–6784, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [140] Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. Medical dialogue generation via intuitive-then-analytical differential diagnosis. *CoRR*, abs/2401.06541, 2024.
- [141] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7346–7353. AAAI Press, 2019.
- [142] Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang, and Qi He. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. *CoRR*, abs/2410.17952, 2024.

-
- [143] Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D. Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Carl Yang, Yang Xie, and Wenqi Shi. Medagentgym: Training LLM agents for code-based medical reasoning at scale. *CoRR*, abs/2506.04405, 2025.
- [144] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. Remedi: Resources for multi-domain, multi-service, medical dialogues. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3013–3024. ACM, 2022.
- [145] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.
- [146] Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. Kg-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2024, Bangkok, Thailand, August 16, 2024*, pages 155–166. Association for Computational Linguistics, 2024.

- [147] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2, 2025.
- [148] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19368–19376. AAAI Press, 2024.
- [149] Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. Finemedlm-o1: Enhancing the medical reasoning ability of LLM from supervised fine-tuning to test-time training. *CoRR*, abs/2501.09213, 2025.
- [150] Kun-Hsing Yu, Elizabeth Healey, Tze-Yun Leong, Isaac S Kohane, and Arjun K Manrai. Medical artificial intelligence and human values. *New England Journal of Medicine*, 390(20):1895–1904, 2024.
- [151] Hongyi Yuan and Sheng Yu. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *Artificial Intelligence in Medicine*, 148:102748, 2024.
- [152] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods*

- in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, November 2020. Association for Computational Linguistics.
- [153] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. Huatuoogpt, towards taming language model to be a doctor. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10859–10885. Association for Computational Linguistics, 2023.
- [154] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.
- [155] Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Llm eval-med: A real-world clinical benchmark for medical llms with physician validation. *CoRR*, abs/2506.04078, 2025.
- [156] Jianqiao Zhao, Yanyang Li, Wanyu Du, Yangfeng Ji, Dong Yu, Michael R. Lyu, and Liwei Wang. Floweval: A consensus-based dialogue evaluation framework using segment act flows. *CoRR*, abs/2202.06633, 2022.
- [157] Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. Medical dialogue response generation with pivotal information recalling. In Aidong Zhang and Huzefa Rangwala, editors, *KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4763–4771. ACM, 2022.

- [158] Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language model for cardiac disease diagnosis, 2025.
- [159] Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinform.*, 38(16):3995–4001, 2022.
- [160] Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 886–896, Online, August 2021. Association for Computational Linguistics.
- [161] Shuang Zhou, Wenya Xie, Jiayi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. Automating expert-level medical reasoning evaluation of large language models. *CoRR*, abs/2507.07988, 2025.
- [162] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *CoRR*, abs/2501.18362, 2025.