Enhancing Medical Dialogue Generation through Knowledge Refinement and Dynamic Prompt Adjustment

Hongda Sun[1*] Jiaren Peng[2*] Wenzhong Yang[3] Liang He[4] Bo Du[5] Rui Yan[167†]
[1]Gaoling School of Artificial Intelligence, Renmin University of China [2]Sichuan University
[3]School of Computer Science and Technology, Xinjiang University [4]Tsinghua University
[5]School of Computer Science, Wuhan University [6]School of Artificial Intelligence, Wuhan University
[7]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
{sunhongda98, ruiyan}@ruc.edu.cn, jiarenpeng666@gmail.com
yangwenzhong@xju.edu.cn, heliang@tsinghua.edu.cn, dubo@whu.edu.cn

Abstract

Medical dialogue systems (MDS) have emerged as crucial online platforms for enabling multi-turn, context-aware conversations with patients. However, existing MDS often struggle to (1) identify relevant medical knowledge and (2) generate personalized, medically accurate responses. To address these challenges, we propose MedRef, a novel MDS that incorporates knowledge refining and dynamic prompt adjustment. First, we employ a knowledge refining mechanism to filter out irrelevant medical data, improving predictions of critical medical entities in responses. Additionally, we design a comprehensive prompt structure that incorporates historical details and evident details. To enable real-time adaptability to diverse patient conditions, we implement two key modules, Triplet Filter and Demo Selector, providing appropriate knowledge and demonstrations equipped in the system prompt. Extensive experiments on MedDG and KaMed benchmarks show that MedRef outperforms state-of-the-art baselines in both generation quality and medical entity accuracy, underscoring its effectiveness and reliability for real-world healthcare applications.

1 Introduction

Medical dialogue systems (MDS) have emerged as a pivotal research spotlight, aiming to support healthcare professionals through multi-turn and context-aware conversations with patients (Shi et al., 2024). Unlike general dialogue systems, MDS must understand and respond using medical domain knowledge (Wei et al., 2018; Xu et al., 2019; Xia et al., 2020), offering valuable support for preliminary assessments and nursing care, particularly in resource-constrained environments (Graham et al., 2014).

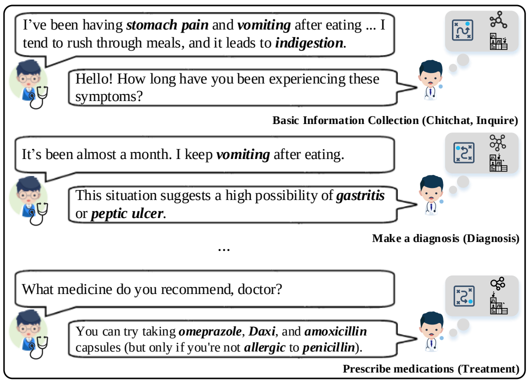

I've been having ***stomach pain*** and ***vomiting*** after eating ... I tend to rush through meals, and it leads to ***indigestion***.
Hello! How long have you been experiencing these symptoms?
Basic Information Collection (Chitchat, Inquire)
It's been almost a month. I keep ***vomiting*** after eating.
This situation suggests a high possibility of ***gastritis*** or ***peptic ulcer***.
Make a diagnosis (Diagnosis)
...
What medicine do you recommend, doctor?
You can try taking ***omeprazole***, ***Daxi***, and ***amoxicillin*** capsules (but only if you're not ***allergic*** to ***penicillin***).
Prescribe medications (Treatment)

Figure 1: An example of medical dialogue generation.

Despite the promise of MDS, several challenges remain in delivering accurate and contextually appropriate responses. One key challenge is effectively tracking a patient's evolving health state throughout multi-turn interactions. As displayed in Figure 1, doctors gradually refine their understanding of a patient's condition over successive turns. Similarly, MDS must maintain coherence as the conversation progresses. A common approach involves retrieving relevant medical entities (*e.g.*, symptoms, diagnoses, treatments) from a medical knowledge graph (MedKG) (Li et al., 2021; Zhao et al., 2022). However, such retrieval-augmented generation (RAG) methods often introduce irrelevant knowledge, which degrades response quality.

Meanwhile, large language models (LLMs) have greatly improved MDS fluency, but remain sensitive to the prompt structure and content. Effective prompts for MDS must (1) direct the model's attention to critical medical entities and dialogue acts, and (2) include relevant conversation demonstrations for guidance. Crucially, these prompts should dynamically adapt to reflect real-time patient information, which is underexplored by existing MDS.

To address these challenges, we aim to (1) refine the retrieved knowledge for more accurate response guidance and (2) dynamically adjust system

[*] Equal contribution.

[†] Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

Findings of the Association for Computational Linguistics: ACL 2025, pages 25715–25726
July 27 - August 1, 2025

prompts to align with specific patient conditions. Therefore, we propose MedRef, a novel MDS with knowledge refining and dynamic prompt adjustment. First, we explicitly represent the patient's condition by incorporating contextual medical entities. Inspired by (Xu et al., 2023), we adopt an entity-action joint prediction module to obtain the expected entities and acts. To mitigate noise from retrieved entities, we introduce a knowledge refining mechanism to enable more accurate entity prediction and knowledge-driven response generation. Building upon this, we construct a comprehensive prompt structure tailored to each dialogue turn. This system prompt mainly includes the following key components: **(1) Task instruction:** A high-level directive guiding the system's response generation process. **(2) Historical details:** A summary of the dialogue context and identified medical entities. **(3) Evident details:** Predicted entities and acts, and relevant knowledge triplets to provide medical evidence for response generation. **(4) Relevant demonstration:** An example conversation for response formatting. To enhance responsiveness, we integrate a dynamic prompt adjustment strategy that updates prompt contents in real time. Specifically, we leverage the Triplet Filter and Demo Selector to retain only the most relevant knowledge and demonstrations. This enables our system to generate accurate, contextually grounded, and patient-specific responses throughout the dialogue.

We conduct extensive experiments on two widely used benchmarks: MedDG (Liu et al., 2020) and KaMed (Li et al., 2021). Experimental results demonstrate the superiority of our MedRef compared with state-of-the-art baselines in both generation quality and medical entity accuracy. Ablation studies further validate the effectiveness of each module in our framework.

To sum up, our contributions can be summarized as follows:

- We propose MedRef, a novel medical dialogue system that jointly addresses knowledge redundancy and prompt adaptation for more accurate and context-aware response generation.
- We introduce a knowledge refining mechanism to filter out irrelevant information in retrieved knowledge, enhancing medical entity prediction and response grounding.
- We develop a dynamic prompt adjustment strategy that adapts prompt components in real time to the patient's condition for improved personalization and coherence.

2 Related Work

2.1 Medical Dialogue System

Medical dialogue systems (MDS) are typically treated as a type of task-oriented dialogue system designed to assist in diagnosis and treatment (Valizadeh and Parde, 2022; Varshney et al., 2022; Sun et al., 2022, 2024). However, progress in this area is often limited in collecting large-scale medical datasets due to privacy and ethical concerns. To address this, Zeng et al. (2020) released MedDialog, a large-scale Chinese-English medical dialogue dataset, which features a larger number of conversation sessions with relatively short turns. Liu et al. (2020) introduced MedDG with medical entity annotations in each utterance, facilitating more fine-grained analysis. Early studies on MDS rely on template-based methods for various tasks like information extraction (Peng et al., 2024; Zhang et al., 2020), relation prediction (Du et al., 2019; Lin et al., 2019; Xia et al., 2021), and slot filling (Shi et al., 2020). More recently, response generation has gained focus, leveraging sequence-to-sequence models (Bahdanau et al., 2014; Vaswani et al., 2017; See et al., 2017) and pre-trained models like BioBERT (Lee et al., 2020), MedBERT (Rasmy et al., 2021), GPT-2 (Radford et al., 2019), and DialoGPT (Zhang et al., 2019). MDS require integration of medical knowledge for accurate responses. Building upon this, VRBot (Li et al., 2021) formulates patient states and physician actions for response generation. MedPIR (Zhao et al., 2022) recalls pivotal information as a prefix to generate responses. DFMed (Xu et al., 2023) uses a dual flow enhanced framework to sequentially model the medical entities and dialogue acts.

2.2 Knowledge-Grounded Dialogue Generation

Knowledge-grounded conversations (KGC) aim to generate responses based on background knowledge retrieved from knowledge graphs (Speer et al., 2017; Ghazvininejad et al., 2018; Li et al., 2020; Chen et al., 2020). The background knowledge is generally retrieved from structured and unstructured sources. The unstructured knowledge used in KGC is mainly documents or paragraphs (Dinan et al., 2018; Zhang et al., 2018; Kim et al., 2020; Zhao et al., 2020). Structured KGC, on the other hand, relies on knowledge triplets or graphs to predict key entities (Liu et al., 2018; Tuan et al., 2019; Xu et al., 2020). Given the dependence of medical dialogues on domain-specific knowledge, KGC methods have been widely applied using medical knowledge graphs (MedKG) to support informed responses (Li et al., 2021; Zhao et al., 2022).

However, existing approaches often retrieve irrelevant information from MedKG, misaligning with a patient's specific condition. Therefore, we propose a knowledge refining mechanism for improved entity prediction and response generation.

3 Method

3.1 Problem Formulation

Suppose a medical conversation session $c = \{u_1, r_1, u_2, r_2, \ldots, u_T, r_T\}$ lasts for a total of T turns of utterances, where u_t and r_t represent the patient's utterance and the doctor's response at the t-th turn. The dialogue context at each turn t is denoted as $\overline{c}_t = \{u_1, r_1, \ldots, u_{t-1}, r_{t-1}, u_t\}$, which conditions the generation of the current doctor response r_t. Each utterance introduces multiple medical entities, and each doctor's response is further annotated with dialogue acts. The historical medical entities $\overline{x}_t$ and dialogue acts $\overline{a}_t$ within $\overline{c}_t$ guide the generation of the response r_t. Moreover, a medical knowledge graph G is commonly used to retrieve relevant knowledge to aid in response generation. Therefore, the objective of MDS is to generate the doctor response r_t at each turn t, conditioned on the dialogue context $\overline{c}_t$, historical entities $\overline{x}_t$, historical acts $\overline{a}_t$ and relevant knowledge from G.

3.2 Input Representation

To effectively track the patient's health condition and generate appropriate responses, it is essential to encode the key components of the dialogue history in the MDS. In the context $\overline{c}_t$, each patient utterance is denoted as $u_i = (u_{i,1}, u_{i,2}, \cdots, u_{i,|U_i|})$ with $|U_i|$ tokens, and each doctor utterance as $(r_{j,1}, r_{j,2}, \cdots, r_{j,|R_j|})$ with $|R_j|$ tokens. To capture their semantic content, we first apply an embedding layer f_{emb}, yielding token-level embeddings e_{u_i} and e_{r_j} for patient and doctor utterances, respectively. Given the medical nature of the task, we adopt MedBERT , a pre-trained model specialized in medical domains, as our encoder backbone. The embedded utterances are processed by this encoder f_{enc} to incorporate sequential dialogue information, and the final output $e_{\overline{c}_t}$ serves as the contextual representation for subsequent modules. The encoding process can be formalized:

$$\begin{aligned} e_{u_i} &= f_{emb}(u_{i,1}, u_{i,2}, \cdots, u_{i,|U_i|}), \\ e_{r_j} &= f_{emb}(r_{j,1}, r_{j,2}, \cdots, r_{j,|R_j|}), \\ e_{\overline{c}_t} &= f_{enc}(e_{u_1}, e_{r_1}, \cdots, e_{u_t}). \end{aligned} \tag{1}$$

We then retrieve related entities from a medical knowledge graph G to guide accurate response generation. Specifically, we construct a subgraph $G^0_{\overline{x}_t} = \{G^0_{\overline{x}_{t_1}}, \ldots, G^0_{\overline{x}_{t_m}}\}$ that totally contains m historical entities $\overline{x}_t$ and their one-hop neighbors. Then we encode these entities using f_{enc} and structural information via a graph attention network (GAT) (Velickovic et al., 2018), f_{gat}. This yields the subgraph representation:

$$e^{G_0}_{\overline{x}_t} = f_{gat}(f_{enc}(G^0_{\overline{x}_{t_1}}, \ldots, G^0_{\overline{x}_{t_m}})). \tag{2}$$

Moreover, the dialogue acts capture the communicative intents of each response (*e.g.,* symptom inquiry, disease diagnosis, and treatment suggestion). The historical dialogue acts are encoded as act-level representations $e_{\overline{a}_t}$. These enriched representations collectively provide the contextual information for accurate response generation.

3.3 Knowledge Refining Mechanism

Retrieved entities can be noisy or overly broad due to deterministic retrieval. To address this, we use a knowledge refinement mechanism that models a latent variable z_t to filter irrelevant knowledge. We first estimate the prior distribution $p_\theta(z_t|\overline{c}_t, G^0_{\overline{x}_t})$ based on the dialogue context $\overline{c}_t$ and retrieved entities $G^0_{\overline{x}_t}$. To guide the prior toward retaining useful knowledge, we define a posterior distribution $q_\phi(z_t|\overline{c}_t, G^0_{\overline{x}_t}, x_t)$ by incorporating the ground-truth entities x_t from the target response r_t. Both prior and posterior are modeled as Gaussian distributions and parameterized via separate encoders:

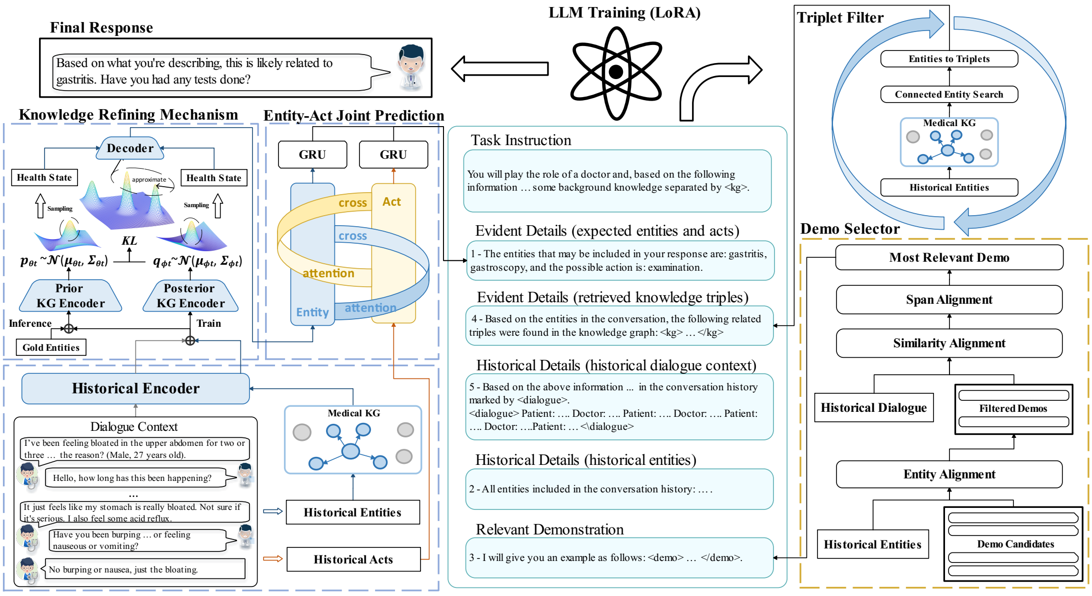
$$\begin{aligned} p_\theta(z_t|\overline{c}_t, G^0_{\overline{x}_t}) &= \mathcal{N}(\mu_\theta(e_{\overline{c}_t}, e^{G_0}_{\overline{x}_t}), \Sigma_\theta(\overline{c}_t, e^{G_0}_{\overline{x}_t})), \\ q_\phi(z_t|\overline{c}_t, G^0_{\overline{x}_t}, x_t) &= \mathcal{N}(\mu_\phi(\overline{c}_t, e^{G_0}_{\overline{x}_t}, x_t), \Sigma_\phi(\overline{c}_t, e^{G_0}_{\overline{x}_t}, x_t)), \end{aligned} \tag{3}$$

where μ_θ, Σ_θ, μ_ϕ, and Σ_ϕ are computed from separate knowledge encoder networks. Once the latent factor z_t is sampled, it is passed through the knowledge decoder f_{dec}, and its output is combined with the original entity embedding $e^{G_0}_{\overline{x}_t}$ to produce the refined representation:

$$e^{G}_{\overline{x}_t} = f_{dec}(z_t) + e^{G_0}_{\overline{x}_t} \tag{4}$$

This refined embedding $e^{G}_{\overline{x}_t}$, with reduced noise and improved relevance, is used to better predict the expected entities in the response.

https://github.com/trueto/medbert



Based on what you're describing, this is likely related to gastritis. Have you had any tests done?

I've been feeling bloated in the upper abdomen for two or three … the reason? (Male, 27 years old).

Hello, how long has this been happening?

…

It just feels like my stomach is really bloated. Not sure if it's serious. I also feel some acid reflux.

Have you been burping … or feeling nauseous or vomiting?

No burping or nausea, just the bloating.

You will play the role of a doctor and, based on the following information … some background knowledge separated by <kg>.

1 - The entities that may be included in your response are: gastritis, gastroscopy, and the possible action is: examination.

4 - Based on the entities in the conversation, the following related triples were found in the knowledge graph: <kg> … </kg>

5 - Based on the above information ... in the conversation history marked by <dialogue>.
<dialogue> Patient: …. Doctor: …. Patient: …. Doctor: …. Patient: …. Doctor: ….Patient: … <\dialogue>

2 - All entities included in the conversation history: ….

3 - I will give you an example as follows: <demo> … </demo>.

Figure 2: System overview of our MedRef, involving encoding dialogue history, refining retrieved knowledge, and jointly predicting entities and acts. The triplet filter and demo selector are used to enhance the prompt for final response generation.

3.4 Entity-Act Joint Prediction

Based on the refined knowledge, we can reconstruct the entities in the response. To capture the high correspondence between medical entities (symptoms, diseases, and treatments) and dialogue acts (symptom inquiry, disease diagnosis, and treatment suggestions), we leverage a joint prediction module to obtain the expected entities and acts in the target response. We first model interactions between context, refined entities, and historical acts using a cross-attention module f_{ca}, followed by a GRU f_{gru} to obtain new representations:

$$\begin{aligned} e_t^{CG} &= f_{ca}(e_{\overline{c}_t}, e^G_{\overline{x}_t}), \\ e_t^{CGA} &= f_{ca}(e_t^{CG}, e_{\overline{a}_t}), \\ \widetilde{e}^G_{\overline{x}_t} &= f_{gru}(e^G_{\overline{x}_t} \oplus e_t^{CG} \oplus e_t^{CGA}). \end{aligned} \tag{5}$$

$$\begin{aligned} e_t^{CA} &= f_{ca}(e_{\overline{c}_t}, e_{\overline{a}_t}), \\ e_t^{CAG} &= f_{ca}(e_t^{CA}, e^G_{\overline{x}_t}), \\ \widetilde{e}_{\overline{a}_t} &= f_{gru}(e_{\overline{a}_t} \oplus e_t^{CA} \oplus e_t^{CAG}). \end{aligned} \tag{6}$$

We then compute prediction probabilities for entities and acts in the t-th turn via linear transformation layers along with sigmoid $\sigma(\cdot)$ as activation functions.

$$\begin{aligned} \widehat{x}_t &= \sigma(W_x \widetilde{e}^G_{\overline{x}_t} + b_x), \\ \widehat{a}_t &= \sigma(W_a \widetilde{e}_{\overline{a}_t} + b_a), \end{aligned} \tag{7}$$

where $W_x \in \mathbb{R}^{|X| \times d}$ and $b_x \in \mathbb{R}^{|X|}$; $W_a \in \mathbb{R}^{|A| \times d}$ and $b_a \in \mathbb{R}^{|A|}$. $|X|$ and $|A|$ are the numbers of candidate entities and acts, and d is the hidden size.

3.5 Dynamic Prompt Adjustment

3.5.1 Prompt Design

To better motivate LLMs to generate accurate and patient-specific responses, we design a comprehensive prompt structure. As shown in Figure 2, the system prompt $\mathcal{P} = [\mathcal{I}; \mathcal{H}; \mathcal{K}; \mathcal{E}]$ contains the following key components:

Task instruction $\mathcal{I}$ outlines the task that responds to the patient and explains the structure of the remaining prompts. **Historical details** $\mathcal{H}$ summarize key elements in the dialogue history, including the dialogue context $\overline{c}_t$, and sequentially listed historical entities $\overline{x}_t$ and acts $\overline{a}_t$. **Evident details** $\mathcal{K}$ provide medical knowledge for generating responses, containing predicted entities and acts, and relevant knowledge triplets from MedKG. **Relevant demonstration** $\mathcal{E}$ provides an in-context example to guide response formatting.

To enable real-time adaptation for varying patient conditions, we integrate a dynamic prompt adjustment strategy by introducing the Triplet Filter and Demo Selector modules to refine the equipped knowledge and demonstrations in the prompt.

3.5.2 Triplet Filter

To obtain reliable knowledge triplets from retrieved entities, we design an iterative filtering process.

First, the retrieved one-hop subgraph $G^0_{\overline{x}_t}$ is transformed into a set of triplets $Tri^0_{\overline{x}_t}$. Next, we compute the frequency of each entity in these triplets and sort them in descending order. Based on these frequencies, we dynamically adjust the triplets retained by setting a threshold τ. Those triplets can be kept if and only if their head and tail entities both have frequencies not less than τ.

$$Tri^{\tau}_{\overline{x}_t} = \{(e_{head}, r, e_{tail}) | \min(\#e_{head}, \#e_{tail}) \geq \tau\} \quad (8)$$

Initially, τ is set to 1 and is incremented in each iteration, gradually reducing the number of retained triplets. The process terminates once the number of triplets in $Tri^{\tau}_{\overline{x}_t}$ does not exceed a predefined maximum M. The current $Tri^{\tau}_{\overline{x}_t}$ is then used as part of the final evident details in the prompt.

3.5.3 Demo Selector

To select the most relevant demonstration for the system prompt, we introduce a multi-step alignment process.

Entity alignment. We begin by organizing all training conversations into subsets based on entity annotations in the first patient utterance. Specifically, we construct multi-entity subsets $S_E = \{S_{E_1}, \ldots, S_{E_K}\}$, where each subset S_{E_k} contains conversation cases whose first utterance includes the same n entities $E_k = \{x_1, \ldots, x_n\}$. In parallel, we create single-entity subsets S_e, where each subset $S_{e'}$ contains cases with first utterances that mention the shared entity e'.

Given a current dialogue context $\overline{c}_t$, we need to check whether its first utterance u_1 exactly matches any entity set in S_E. If so, we retrieve the corresponding subset as the candidate demo set S_{demo}. Otherwise, we fall back to the single-entity subsets and select all sessions from S_e that share at least one entity with u_1.

Similarity alignment. To refine the demo selection, we compute the semantic similarity between the current first utterance u_1 and those in S_{demo}. We encode each candidate separately and then apply cosine similarity to identify the closest conversation c_{full} as the demonstration reference.

Span alignment. To improve contextual relevance and reduce prompt length, we extract a focused span from c_{full} using a sliding window of size ξ. Let the total utterance sequence of c_{full} be $\{u_1, r_1, u_2, r_2, \ldots, u_T, r_T\}$, and denote the start index as $i_s = 2t - 1$, corresponding to the current dialogue turn t. The final demonstration $\mathcal{E}$ is intercepted in three cases: (1) If $i_s \leq \xi$, we select the first 2ξ utterances from c_{full}; (2) If $\xi < i_s < T - \xi$, we select the utterances from index $i_s - \xi$ to $i_s + \xi$; (3) If $T - \xi \leq i_s$, we select the last 2ξ utterances.

3.6 Model Optimization

To optimize different modules of MedRef, we design a two-stage training objective. We first pretrain the entity-act joint prediction module in preparation for subsequent response generation. For predicting medical entities, we compute the binary cross-entropy (BCE) loss $\mathcal{L}_x$ between predictions $\widehat{x}_t$ and ground-truth entity labels x_t. Similarly, dialogue act prediction is trained based on the cross-entropy loss $\mathcal{L}_a$. These loss functions can be formulated as:

$$\begin{aligned}
\mathcal{L}_x &= -\sum_{t=1}^{T}\sum_{i=1}^{|X|}[x_{t_i}\log(\widehat{x}_{t_i}) + (1 - x_{t_i})\log(1 - \widehat{x}_{t_i})], \\
\mathcal{L}_a &= -\sum_{t=1}^{T}\sum_{j=1}^{|A|}[a_{t_j}\log(\widehat{a}_{t_j}) + (1 - a_{t_j})\log(1 - \widehat{a}_{t_j})],
\end{aligned} \quad (9)$$

To ensure consistency in knowledge refining, we minimize the Kullback-Leibler (KL) divergence between the prior p_θ and posterior q_ϕ:

$$\mathcal{L}_{kl} = \sum_{t=1}^{T} D_{KL}(q_\phi(z_t|\mu_\phi, \Sigma_{\phi_t})||p_\theta(z_t|\mu_\theta, \Sigma_{\theta_t})). \quad (10)$$

We assign the weights λ_x, λ_a, and λ_{kl} to each loss, and the overall loss function for this stage is a weighted combination:

$$\mathcal{L} = \lambda_x\mathcal{L}_x + \lambda_a\mathcal{L}_a + \lambda_{kl}\mathcal{L}_{kl}. \quad (11)$$

Next, with the prediction module fixed, we finetune the medical LLM responsible for response generation. By maximizing the log-likelihood of the system responses, the language model based loss is given by:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T}\log\sum_{k} p_{gen}(r_{t_k}|r_{t_{<k}}, \mathcal{P}). \quad (12)$$

4 Experimental Setup

4.1 Datasets

We conduct experimental evaluations on two widely used benchmarks, MedDG and Kamed. MedDG contains over 17,000 medical dialogues annotated with 160 medical entities across 5 categories: diseases, symptoms, medications, examinations, and attributes. It is officially split into 14,862 (train), 1,999 (validation), and 999 (test) sessions.

Table 1: Comparison results on MedDG and KaMed datasets. "B"=BLEU, "R"=ROUGE, "E-F1"=entity-F1. Bold/underline numbers denote significant improvements (p-value<0.01) over the second-best.

Category	Method	*MedDG*						*KaMed*					
		B-1	**B-2**	**B-4**	**E-F1**	**R-1**	**R-2**	**B-1**	**B-2**	**B-4**	**E-F1**	**R-1**	**R-2**
DL-based	Seq2Seq	28.55	22.85	15.45	12.88	25.61	11.24	23.52	18.56	12.13	-	23.56	8.67
	VRBot	29.69	23.90	16.34	12.78	24.69	11.23	30.04	23.76	16.36	12.08	18.71	7.28
PLM-based	GPT-2	35.27	28.19	19.16	16.14	28.74	13.61	33.76	26.58	17.82	17.26	26.80	10.56
	BART	34.94	27.99	19.06	16.66	29.03	14.40	33.62	26.43	17.64	19.20	27.91	11.43
	DFMed	41.74	32.93	22.48	21.54	28.90	13.71	39.59	30.53	20.30	21.33	27.67	11.21
LLM-based	DISC-MedLLM	40.72	-	22.60	10.15	20.13	6.6	38.05	-	20.26	13.54	20.48	5.93
	GPT-4o	42.19	-	**23.32**	13.15	13.99	3.47	**41.88**	-	23.34	13.86	13.94	3.1
	HuatuoGPT-II	39.03	32.56	23.02	8.67	10.94	1.76	40.35	**32.93**	**23.92**	12.00	13.84	2.74
	Zhongjing	26.65	21.75	15.02	6.43	13.14	2.82	27.48	22.35	15.52	6.44	13.70	3.05
	Chatglm3-6B	33.16	26.51	17.97	17.43	29.27	13.69	32.03	25.20	16.68	20.56	28.02	12.12
	MedRef	**43.51**	**33.82**	23.04	**22.70**	**30.07**	**14.52**	40.47	31.62	21.28	**21.96**	**28.14**	**12.42**

Kamed includes over 63,000 dialogues spanning 100+ departments. Following DFMed (Xu et al., 2023), we remove privacy-sensitive data, resulting in 29,159 (train), 1,532 (validation), and 1,539 (test) sessions. Dialogue acts are labeled into 7 types: *Chitchat, Inform, Inquire, Provide Daily Precaution, State a Required Medical Test, Make a Diagnosis, and Prescribe Medications.*

4.2 Baselines

We compare MedRef against the following three types of baselines: (1) **DL-based methods**: Seq2Seq (Sutskever et al., 2014), RNN with attention; VRBOT (Li et al., 2021), patient state and physician action tracking model. (2) **PLM-based methods**: GPT-2 (Radford et al., 2019), BART (Lewis, 2019), general-purpose generative models; DFMed (Xu et al., 2023), dual flow model leveraging interwoven entities and acts. (3) **LLM-based methods**: Chatglm3-6B (Du et al., 2022), general LLM fine-tuned on medical dialogues; Zhongjing (Yang et al., 2024), traditional Chinese medicine dialogue model; HuatuoGPT-II (Chen et al., 2023) (Baichuan-7B), DISC-MedLLM (Bao et al., 2023) (Baichuan-13B), specialized medical LLM; GPT-4o (Hurst et al., 2024), advanced closed-source LLM.

4.3 Evaluation Metrics

Automatic evaluation. To evaluate the quality of the model's generated responses, we utilize **BLEU** (Papineni et al., 2002) and **ROUGE** (Lin, 2004) for assessing lexical similarity, and **entity-F1** score to measure entity-level accuracy.

Human evaluation. We focus on three key human evaluation metrics: **fluency (FLU)** measures how naturally and smoothly the conversation flows; **knowledge accuracy (KC)** focuses on the correctness of the medical terms; and **overall quality (OQ)** considers the holistic response effectiveness.

4.4 Implementation Details

We use ChatGLM3-6B as the backbone of our response generator, which is fine-tuned with LoRA (Hu et al., 2021) (rank=8, α=32, dropout=0.1) using AdamW (lr=5e-5). MedBERT (Rasmy et al., 2021) is used for entity and act prediction (lr=3e-5, batch size=8). We retrieve up to M=25 triplets from CMeKG (Byambasuren et al., 2019). The sliding window size is ξ=2. The loss weights are set to $\lambda_x = 1$, $\lambda_a = 0.05$, $\lambda_{kl} = 0.05$.[1]

5 Experimental Results

5.1 Overall Performance

As shown in Table 1, MedRef consistently outperforms all baselines across multiple metrics, demonstrating its effectiveness in generating high-quality, medically grounded responses. Compared to GPT-4o, MedRef achieves +1.32% BLEU-1, +16.08% ROUGE-1, and +11.05% Entity-F1, demonstrating superior lexical alignment, fluency, and medical accuracy. This advantage stems from task-specific fine-tuning, whereas GPT-4o's closed-source nature limits its adaptability to medical dialogue nuances. MedRef tends to generate fluent utterances

[1] Our code is available at https://github.com/simon-p-j-r/MedReF.

Table 2: Ablation results of MedRef on MedDG and KaMed datasets.

Method	*MedDG*						*KaMed*					
	B-1	**B-2**	**B-4**	**E-F1**	**R-1**	**R-2**	**B-1**	**B-2**	**B-4**	**E-F1**	**R-1**	**R-2**
MedRef	**43.51**	**33.82**	**23.04**	**22.70**	**30.07**	**14.52**	**40.47**	**31.62**	**21.28**	**21.96**	**28.14**	**12.42**
w/o KRM	42.58	33.45	22.70	21.94	29.88	14.23	40.29	31.10	20.88	21.51	27.95	11.92
w/o Demo	41.80	32.87	22.31	21.84	29.69	13.93	39.07	30.34	20.46	20.09	27.35	11.90
w/o Kg	41.76	32.83	22.24	21.58	29.86	13.93	39.82	30.96	20.81	20.55	28.09	11.87
E-A&Cxt only	41.63	32.75	22.30	21.30	28.68	13.27	39.30	30.38	20.42	20.81	26.72	11.22
Cxt only	33.16	26.51	17.97	17.43	29.27	13.69	32.03	25.20	16.68	20.56	28.02	12.12

that align well with human-authored responses, contributing to its superior ROUGE and entity-F1 scores, reflecting content richness and relevance.

However, MedRef slightly underperforms HuatuoGPT-II and GPT-4o on BLEU scores on KaMed. This discrepancy may arise from the dataset complexity and the response style bias of these models. First, KaMed spans a broader range of clinical scenarios, encompassing over 100 departments, which increases the complexity of the required medical knowledge and makes learning high-coverage representations more challenging. Besides, HuatuoGPT-II and GPT-4o often generate verbose, QA-style replies. While this verbosity can increase token-level overlap with references (thereby inflating BLEU scores), it tends to introduce irrelevant or redundant content, leading to much lower entity-F1 scores. Second, HuatuoGPT-II and GPT-4o tend to adopt a QA-style approach to addressing patient inquiries, often generating very long text responses with redundancy and nonsense. This response trend is not enough to the point that it helps to slightly improve the BLEU indicator, but significantly reduces the entity F1 score.

5.2 Ablation Study

To investigate the contribution of each module in the proposed system, we conduct a comprehensive ablation study that includes the following variants for comparison: (1) **w/o KRM** removes knowledge refinement mechanism. (2) **w/o Demo** removes the demonstration $\mathcal{E}$ matched by the demo selector. (3) **w/o Kg** removes the knowledge triplets retrieved from MedKG. (4) **E-A&Cxt only** retains only the predicted entities and actions along with the dialogue context; no demonstrations or external knowledge are provided, and the KRM is not used. (5) **Cxt only** uses only the dialogue context, without any additional guidance or knowledge.

The ablation results in Table 2 show that all variant models exhibit noticeable performance declines, underscoring the importance of each component. In particular, w/o KRM suffers the most significant drop across all evaluation metrics, highlighting its dual role in filtering out redundant knowledge and improving entity prediction accuracy. Moreover, the performance degradation of other model variants relative to the full model illustrates the importance of prompt integrity and also shows that the retrieval knowledge and demonstrations selected into our prompt are more relevant than before.

5.3 Analysis of Triplet Filter and Demo Selector

To further verify the effectiveness of our triplet filter and demo selector modules, we introduce two new model variants: (1) **Weak Kg**: Instead of entirely removing the knowledge triplets from the prompt (w/o Kg), this variant bypasses the filtering rule and directly retrieves the triplets connected to entities in the most recent utterance from the knowledge graph, randomly selecting M triplets from the one-hop connections. (2) **Weak Demo**: In this variant, demonstration examples are selected randomly, without any alignment process to ensure relevance.

The results in Figure 3 show that both variants exhibit significant performance drops across key metrics. Notably, we observe that merely increasing the quantity of knowledge triplets, without applying the triplet filter, harms the model's performance. This suggests that indiscriminate use of knowledge can introduce noise, overwhelming the model and reducing its ability to generate accurate responses. Similarly, the random selection of demonstrations also leads to a decline in generation quality, highlighting the importance of the demo

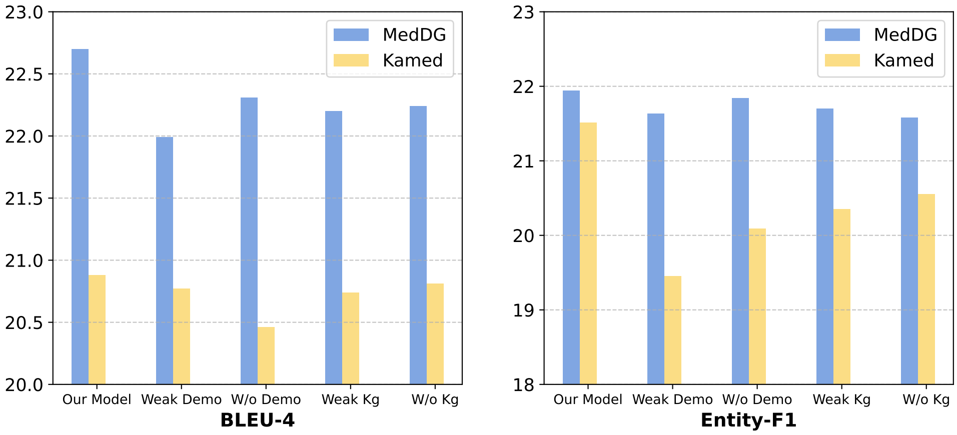


Figure 3: Comparison results of triplet filter and demo selector.

selector's alignment process. These findings confirm that both the triplet filter and demo selector are essential for improving the accuracy and relevance of the generated medical dialogues.

5.4 Case Study

Figure 4 illustrates a running example from the MedDG dataset, showcasing the dialogue across multiple turns.

In Turn-1, MedRef demonstrates its ability to focus on the key entities “hemorrhoids” and “pain”, producing a response that closely matches the Ground Truth. In comparison, both DFMed and the Baseline models fail to fully capture these entities, leading to incomplete responses. This highlights MedRef's superior entity prediction capabilities and its ability to generate more comprehensive inquiries that better address patient concerns.

In Turn-2, DFMed correctly predicts the entity “hemorrhoids” but ignores the patient's earlier statement “I've never had hemorrhoids”, leading to a contradictory response. In contrast, MedRef remains consistent with the patient's current health information, thus leading to a more accurate and contextually appropriate diagnosis.

In Turn-5, MedRef still manages to provide a relevant and informative response. This is largely due to its effective use of the retrieved knowledge and its ability to infer information from the overall context. This example further demonstrates MedRef's robustness, showcasing its ability to handle situations where explicit entity cues are absent, yet still deliver meaningful and accurate dialogue.

Overall, these cases emphasize the advantages of MedRef in not only predicting relevant medical entities but also in maintaining contextual coherence throughout the conversation, leading to more reliable and patient-centered interactions. This illustrates how MedRef surpasses existing baselines, which often struggle with maintaining context consistency and addressing patient concerns comprehensively.

5.5 Human Evaluation

In addition to automatic evaluation, we conduct human evaluation experiments with a dedicated team. The volunteers are all medical doctoral and master's students with extensive experience in annotating medical conversations, who have been working on related projects for the past few years, and can ensure the reliability of the correct judgments. The evaluators are tasked with scoring the responses and rating three aforementioned metrics (FLU, KC, OQ) using a scale from 1 (poor) to 5 (excellent).

As shown in Table 3, MedRef consistently outperforms other baseline models across all three metrics. Notably, the scores for MedRef are the closest to the ground-truth responses, suggesting its higher level of alignment with expert expectations. This reinforces the idea that MedRef's specialized design, particularly the integration of entity-aware mechanisms and dynamic prompt adjustment, leads to more reliable and contextually relevant responses.

A key insight from this evaluation is that our framework's prompt design and dynamic adjustments significantly enhance the generation quality of large language models (LLMs). The results indicate that simply fine-tuning LLMs with generic prompts is insufficient for the complex nature of MDS. In contrast, MedRef leverages tailored prompt strategies and knowledge refinement,

	Chinese	English
Turn-1	患者：今天排便忽然发现***便血***了，鲜血，不痛，粪便也不是很硬，也不***拉稀***，排便也顺畅，以前从来没有出现过这种情况，可能是什么原因导致这种症状的？（女，22岁）。	Patient: "Today, I suddenly noticed ***blood*** in my stool—bright red, without any pain. The stool wasn’t particularly hard, and I don’t have ***diarrhea***. My bowel movements have been smooth ... What could be causing this symptom?" (Female, 22 years old).
	Ground Truth：你好，这种情况多长时间了？肛周有没有***疼痛***。有没有***痔疮***。 DFMed：你好，这种情况多长时间了？以前有***痔疮***吗？ Baseline：你好，这种情况有多长时间了？ Our Model：你好，这种情况多长时间了，有没有***痔疮***，大便干结，肛门***疼痛***等症状。	Ground Truth: "Hello ... Do you have any ***pain*** around the anus? Any history of ***hemorrhoids***?" DFMed: "Hello, how long has this been happening? Have you had ***hemorrhoids*** before?" Baseline: "Hello, how long has this been going on?" Our Model: "Hello ... Any symptoms like ***hemorrhoids***, hard stools, or anal ***pain***?"
Turn-2	患者：没有***疼痛***，没有得过***痔疮***。今天第一次发现这种症状。 ✖	Patient: "No ***pain***, and I've never had ***hemorrhoids***. This is the first time I've noticed this symptom today." ✖
	Ground Truth：还在出血吗。可以看下肛门口。考虑***痔疮***。 DFMed：有没有***痔疮***。大便时肛门疼痛吗。 Baseline：有没有吃辛辣刺激性食物。 Our Model：大便干燥吗？便血考虑***痔疮***或者肛裂。	Ground Truth: "Is there still bleeding? Can you check around the anus? ***Hemorrhoids*** are a possibility." DFMed: "Do you have ***hemorrhoids***? Does your anus hurt during bowel movements?" Baseline: "Have you been eating spicy or irritating foods?" Our Model: "Is your stool dry? The bleeding could be due to ***hemorrhoids*** or an anal fissure."
...	...	...
Turn-5	患者：还行，江苏苏州，医院可以去的。我还想问一下，我去挂号是挂肛肠科还是消化科呢。	Patient: "It’s okay. I’m in Suzhou, Jiangsu, and I can go to the hospital. I’d like to ask, should I register for the proctology department or the gastroenterology department?"
	Ground Truth：直接去急诊外科或者肛肠科。 DFMed：消化科。✖ Baseline：肛肠科。 Our Model：肛肠科。	Ground Truth: "Go directly to the emergency surgery department or the proctology department." DFMed: "Gastroenterology department." ✖ Baseline: "Proctology department." Our Model: "Proctology department."

Figure 4: A running case comparing MedRef with baselines, highlighting that MedRef predicts more accurate medical entities and generates more relevant responses.

allowing it to generate responses that are not only more fluent but also exhibit higher medical accuracy. These findings highlight the advantage of our system, demonstrating that the combination of entity prediction, knowledge refining, and context-aware prompts enables the generation of higher-quality medical dialogues compared to simple fine-tuning strategies.

Table 3: Comparison results for human evaluation. Each metric ranges from 1 to 5.

Method	**FLU**	**KC**	**OQ**
Ground-truth	**3.70**	**3.75**	**3.95**
DFMed	3.42	3.57	3.65
E-A&Cxt only	2.91	3.05	3.14
MedRef	<u>3.55</u>	<u>3.68</u>	<u>3.79</u>

6 Conclusion

In this paper, we propose **Med**ical dialogue system with knowledge **Ref**ining and dynamic prompt adjustment (MedRef). We introduce a variational knowledge refining mechanism for more accurate medical entity predictions and knowledge-driven responses. We also develop a dynamic prompt adjustment method that adapts system prompts in real-time to the patient’s evolving condition, ensuring more personalized and contextually relevant multi-turn medical dialogue generation. Extensive experiments on two benchmarks verify that MedRef can achieve the best performance in terms of both text generation and medical entity-based metrics. These findings underscore MedRef’s potential to improve the quality and reliability of MDS, paving the way for more context-aware and medically sound interactions in healthcare settings.

Limitations

While our model achieves state-of-the-art performance in medical dialogue generation, two key limitations present opportunities for future improvement: (1) Unlike textual medical knowledge, cross-modal knowledge data has not been fully explored to enhance the capture of patient conditions. (2) The emotional support capabilities of current MDS are still passive rather than active. Appropriate comforting strategies are needed while maintaining medical accuracy.

Ethical Considerations

The development and deployment of the medical dialogue system prioritize user safety, privacy, and the responsible use of AI in healthcare. All data used for training is anonymized. The proposed system is clarified to be intended as an assistive tool, not a replacement for professional medical advice, and should be used in conjunction with consultation from qualified healthcare providers.

Acknowledgments

This work was supported by the Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and Public Computing Cloud, Renmin University of China, the fund for building world-class universities (disciplines) of Renmin University of China.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. 2019. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*, 33(10):1–9.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, and 1 others. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3426–3437.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2019. Learning to infer entities, properties and their relations from clinical conversations. *arXiv preprint arXiv:1908.11536*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lisa Graham, Mohammad Moshirpour, Michael Smith, and Behrouz H Far. 2014. Designing interactive health care systems: Bridging the gap between patients and health care professionals. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 235–239. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. 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.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *arXiv preprint arXiv:2010.07497*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Jiaren Peng, Hongda Sun, Wenzhong Yang, Fuyuan Wei, Liang He, and Liejun Wang. 2024. One small and one large for document-level event argument extraction. *arXiv preprint arXiv:2411.05895*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8838–8845.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Hongda Sun, Hongzhan Lin, and Rui Yan. 2024. Collaborative synthesis of patient records through multi-visit health state inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19044–19052.

Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2022. Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. *Advances in Neural Information Processing Systems*, 35:27837–27849.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *arXiv preprint arXiv:1910.00610*.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behra, and Asif Ekbal. 2022. Cdialog: A multi-turn covid-19 conversation dataset for entity-aware dialog generation. *arXiv preprint arXiv:2212.06049*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. 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.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. 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.

Yuan Xia, Chunyu Wang, Zhenhui Shi, Jingbo Zhou, Chao Lu, Haifeng Huang, and Hui Xiong. 2021. Medical entity relation verification with large-scale machine reading comprehension. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3765–3774.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1062–1069.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6771–6784.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2002.10348*.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022. Medical dialogue response generation with pivotal information recalling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4763–4771.