通过
知识优化和动态提示调整提高医疗对话生成效果

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

医学对话系统（MDS）已成为实现与患者进行多轮、上下文感知对话的重要在线平台。然而，现有的 MDS 往往难以（1）识别相关的医学知识和（2）生成个性化且医学上准确的回应。为了解决这些挑战，我们提出了 MedRef，这是一种新颖的 MDS，结合了知识精炼和动态提示调整。首先，我们采用知识精炼机制来过滤掉不相关的医学数据，以改善对响应中关键信息的预测。此外，我们设计了一种综合的提示结构，结合了历史细节和明显细节。为了实现对各种患者状况的实时适应性，我们实施了两个关键模块：三元组过滤器和示例选择器，提供适当的知识和系统提示中的示例。在 MedDG 和 KaMed 基准上的广泛实验表明，在生成质量和医学实体准确性方面，MedRef 在表现上优于最先进的基线，强调了其在现实医疗应用中的有效性和可靠性。

1 介绍

医疗对话系统（MDS）已经成为一个重要的研究焦点，旨在通过与患者进行多轮和情境感知的对话来支持医疗专业人员（Shi et al., 2024）。与一般对话系统不同，MDS 必须使用医学领域的知识来理解和回应（Wei et al., 2018; Xu et al., 2019; Xia et al., 2020），为初步评估和护理提供有价值的支持，特别是在资源有限的环境中（Graham et al., 2014）。

尽管 MDS 有很大的前景，但在提供准确和符合情境的回复方面仍存在若干挑战。一个关键挑战是在多轮对话中有效地追踪病人不断变化的健康状态。如图 1 所示，医生在连续的对话轮次中逐步深化对病人病情的理解。类似地，MDS 必须在对话推进过程中保持连贯性。一个常见的方法是从医学知识图谱（MedKG）(Li et al., 2021; Zhao et al., 2022) 中检索相关医学实体（症状、诊断、治疗）*e.g.*,。然而，这种检索增强生成（RAG）方法常常引入不相关的知识，从而降低回复质量。

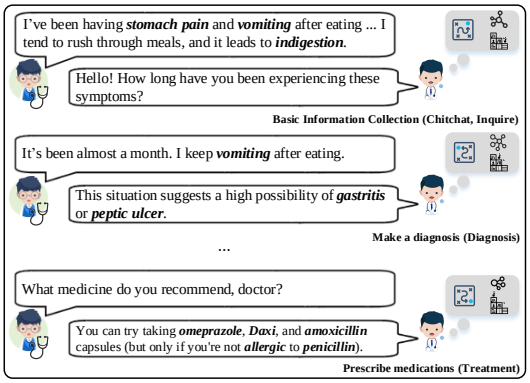

Figure 1: 医学对话生成的一个例子。

同时，大型语言模型（LLM）极大地提高了 MDS 的流畅性，但仍然对提示的结构和内容很敏感。有效的 MDS 提示必须（1）引导模型的注意力集中在关键的医学实体和对话行为上，并且（2）包含相关的对话演示以提供指导。至关重要的是，这些提示应该动态调整以反映实时的患者信息，而这在现有的 MDS 研究中是尚未深入探讨的。

为了解决这些挑战，我们的目标是：（1）优化检索到的知识以指导更准确的响应，（2）动态调整系统提示以符合特定的患者状况。因此，我们提出了 MedRef，一种具有知识优化和动态提示调整的新型 MDS。首先，我们通过纳入上下文医学实体明确表示患者的状况。受 (Xu et al., 2023) 的启发，我们采用了一个实体-动作联合预测模块来获取期望的实体和动作。为减少检索实体的噪声，我们引入了一种知识优化机制，以实现更准确的实体预测和基于知识的响应生成。在此基础上，我们为每一轮对话构建了一个全面的提示结构。此系统提示主要包括以下关键组成部分：（1）任务指令：指导系统响应生成过程的高层次指令。（2）历史细节：对话上下文和识别出的医学实体的总结。（3）明显细节：预测的实体和行为，以及提供响应生成医疗依据的相关知识三元组。（4）相关示例：用于响应格式的示例对话。为了提高响应能力，我们整合了一种实时更新提示内容的动

* Equal contribution.
† Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

态提示调整策略。具体来说，我们利用三元组筛选器和示例选择器仅保留最相关的知识和示例。这使我们的系统能够在整个对话过程中生成准确、上下文相关和针对患者的响应。

我们在两个广泛使用的基准上进行了大量实验：MedDG (Liu et al., 2020) 和 KaMed (Li et al., 2021)。实验结果表明，我们的 MedRef 相较于最新的基准在生成质量和医学实体准确性方面具有优势。消融实验进一步验证了我们框架中各个模块的有效性。

总而言之，我们的贡献可以总结如下：

我们提出了 MedRef，这是一种新颖的医疗对话系统，可以同时解决知识冗余和提示适应，以生成更准确和上下文感知的响应。

- 我们引入了一种知识精炼机制，以过滤掉检索知识中的不相关信息，从而增强医疗实体预测和响应基础。
- 我们制定了一种动态提示调整策略，可以根据患者的状态实时调整提示组件，以提高个性化和连贯性。

2 相关工作

2.1 医疗对话系统

医学对话系统（MDS）通常被认为是一种任务导向型对话系统，旨在协助诊断和治疗 (Valizadeh and Parde, 2022; Varshney et al., 2022; Sun et al., 2022, 2024)。然而，由于隐私和伦理问题，在该领域收集大规模医学数据集的进展常常受到限制。为了解决这一问题，Zeng et al. (2020) 发布了 MedDialog，一个大规模的中英文医学对话数据集，具有较多的对话会话和相对较短的对话轮次。Liu et al. (2020) 引入了 MedDG，其在每个话语中加入了医学实体注释，便于更精细的分析。早期关于 MDS 的研究依赖于基于模板的方法来执行各种任务，如信息提取 (Peng et al., 2024; Zhang et al., 2020)、关系预测 (Du et al., 2019; Lin et al., 2019; Xia et al., 2021) 和槽填充 (Shi et al., 2020)。最近，回应生成成为关注的焦点，利用序列到序列模型 (Bahdanau et al., 2014; Vaswani et al., 2017; See et al., 2017) 和诸如 BioBERT (Lee et al., 2020)、MedBERT (Rasmy et al., 2021)、GPT-2 (Radford et al., 2019) 和 DialoGPT (Zhang et al., 2019) 等预训练模型。MDS 需要整合医学知识以提供准确的回答。在此基础上，VRBot (Li et al., 2021) 通过拟定病人的状态和医生的动作生成回应。MedPIR (Zhao et al., 2022) 通过回忆重要信息作为前缀来生成回应。DFMed (Xu et al., 2023) 使用增强的双流框架来顺序建模医学实体和对话行为。

2.2 Knowledge-Grounded Dialogue Generation

知识驱动对话（KGC）的目标是生成基于从知识图检索到的背景知识的回应 (Speer et al., 2017; Ghazvininejad et al., 2018; Li et al., 2020; Chen et al., 2020)。背景知识通常从结构化和非结构化的来源中获取。KGC 中使用的非结构化知识主要是文档或段落 (Dinan et al., 2018; Zhang et al., 2018; Kim et al., 2020; Zhao et al., 2020)。另一方面，结构化 KGC 依赖于知识三元组或图来预测关键实体 (Liu et al., 2018; Tuan et al., 2019; Xu et al., 2020)。鉴于医学对话对领域特定知识的依赖，KGC 方法已广泛应用于使用医学知识图（MedKG）来支持信息丰富的回应 (Li et al., 2021; Zhao et al., 2022)。

然而，现有的方法通常会从 MedKG 中检索到不相关的信息，无法与患者的具体情况对齐。因此，我们提出了一种知识精炼机制，以改进实体预测和响应生成。

3 方法

3.1 问题表述

假设一个医疗对话会话 $c = \{u_1, r_1, u_2, r_2, \ldots, u_T, r_T\}$ 总共包含 T 回合的发言，其中 u_t 和 r_t 分别表示在第 t 回合中患者的发言和医生的回应。每回合的对话上下文记为 t，它决定了当前医生回应 r_t 的生成。每个发言引入多个医疗实体，并且每个医生的回应进一步用对话行为标注。历史医疗实体 $\bar{x}_t$ 和对话行为 $\bar{a}_t$ 在 $\bar{c}_t$ 中引导回应 r_t 的生成。此外，常用一个医疗知识图谱 G 来检索相关知识以协助生成回应。因此，MDS 的目标是在每回合 t，基于对话上下文 $\bar{c}_t$，历史实体 $\bar{x}_t$，历史行为 $\bar{a}_t$ 和来自 G 的相关知识来生成医生回应 r_t。

为了有效跟踪患者的健康状况并生成适当的响应，必须在 MDS 中编码对话历史的关键组成部分。在上下文 $\bar{c}_t$ 中，每个患者的发言表示为 $u_i = (u_{i,1}, u_{i,2}, \cdots, u_{i,|U_i|})$，包含 $|U_i|$ 个标记，每个医生的发言表示为 $(r_{j,1}, r_{j,2}, \cdots, r_{j,|R_j|})$，包含 $|R_j|$ 个标记。为了捕捉它们的语义内容，我们首先应用嵌入层 f_{emb}，分别为患者和医生的发言生成标记级别的嵌入 e_{u_i} 和 e_{r_j}。鉴于任务的医疗性质，我们采用专门用于医疗领域的预训练模型 MedBERT 作为我们的编码器骨干。嵌入的发言由这个编码器 f_{enc} 处理，以合并顺序对话信息，最终输出 $e_{\bar{c}_t}$ 作为后续模块的上下文表示。编

https://github.com/trueto/medbert

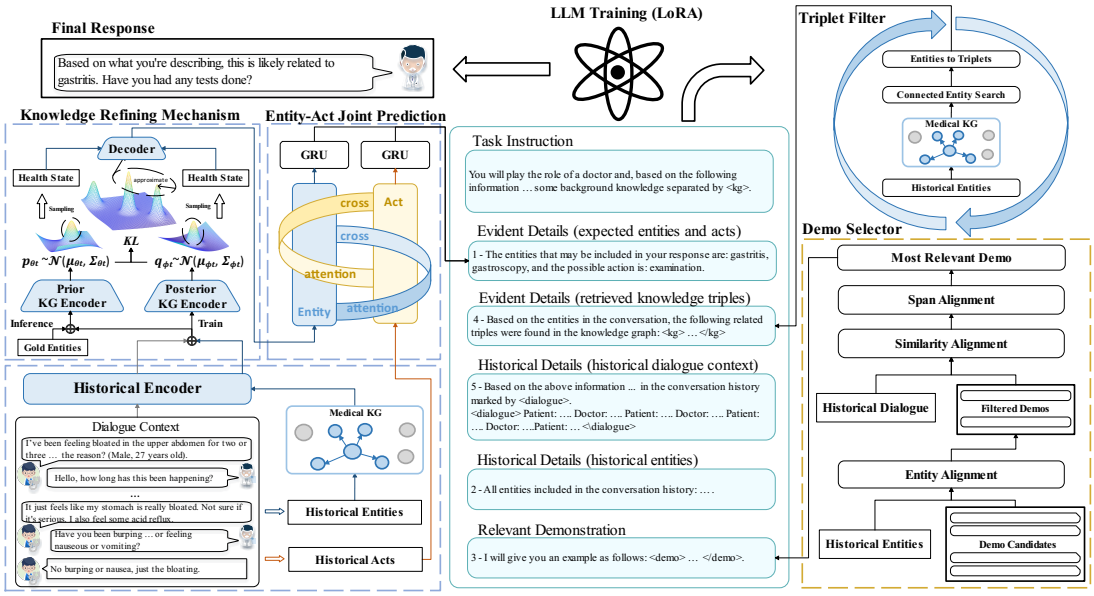

Figure 2: 我们 MedRef 的系统概览，涉及编码对话历史、完善检索的知识以及联合预测实体和行为。三元组过滤器和示例选择器被用来增强最终响应生成的提示。

码过程可以形式化为：

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

然后我们从医学知识图谱 G 中检索相关实体以指导生成准确的响应。具体来说，我们构建一个完全包含 m 历史实体 $\overline{x}_t$ 及其单跳邻居的子图 $G^0_{\overline{x}_t} = \{G^0_{\overline{x}_{t_1}}, \ldots, G^0_{\overline{x}_{t_m}}\}$。然后，我们通过图注意网络（GAT）(Velickovic et al., 2018)，f_{gat} 使用 f_{enc} 和结构信息来编码这些实体。这产生了子图表示：

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

此外，对话行为抓住了每个回复的交流意图（*e.g.*, 症状询问、疾病诊断和治疗建议）。历史对话行为被编码为行为级别的表示 $e_{\bar{a}_t}$。这些丰富的表示共同提供了准确生成回复所需的上下文信息。

3.2 知识精炼机制

由于确定性检索，检索到的实体可能噪声大或过于宽泛。为了解决这个问题，我们使用了一种知识精炼机制，该机制通过建模潜在变量 z_t 来过滤无关的知识。我们首先根据对话上下文 $\bar{c}_t$ 和检索到的实体 $G^0_{\overline{x}_t}$ 估计先验分布 $p_\theta(z_t|\bar{c}_t, G^0_{\overline{x}_t})$。为了引导先验保留有用的知识，我们通过结合来自目标响应 r_t 的真实实体 x_t 定义了后验分布 $q_\phi(z_t|\bar{c}_t, G^0_{\overline{x}_t}, x_t)$。先验和后验都被建模为高斯分布，并通过独立的编码器进行参数化：

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

，其中 μ_θ 、Σ_θ 、μ_ϕ 和 Σ_ϕ 是从不同的知识编码器网络中计算出来的。一旦采样了潜在因子 z_t ，它会通过知识解码器 f_{dec} ，然后其输出与原始实体嵌入 $e^{G_0}_{\overline{x}_t}$ 结合，生成精炼的表示：

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

。这个精炼后的嵌入 $e^{G}_{\overline{x}_t}$ ，具有更少的噪声和更高的相关性，用于更好地预测响应中的期望实体。

基于提炼的知识，我们可以重构响应中的实体。为了捕捉医疗实体（症状、疾病和治疗）与对话行为（症状询问、疾病诊断和治疗建议）之间的高度对应关系，我们利用一个联合预测模块来获取目标响应中的预期实体和行为。我们首先使用交叉注意力模块 f_{ca} 来建模上下文、提炼实体和历史行为之间的交互，然后通过 GRU f_{gru} 来获取新的表示：接着，我们通过线性变换层结合 sigmoid $\sigma(\cdot)$ 作为激活函数，计算出第 t 轮中的实体和行为的预测概率。其中 $W_x \in \mathbb{R}^{|X| \times d}$ 和 $b_x \in \mathbb{R}^{|X|}$ ；$W_a \in \mathbb{R}^{|A| \times d}$ 和 $b_a \in \mathbb{R}^{|A|}$ 。$|X|$ 和 $|A|$ 是候选实体和行为的数量，而 d 是隐藏大小。

3.3 动态提示调整

3.3.1 提示设计

为了更好地激励大语言模型生成准确且针对特定患者的回复，我们设计了一种全面的提示结

构。如图 2 所示，系统提示 $\mathcal{P} = [\mathcal{I}; \mathcal{H}; \mathcal{K}; \mathcal{E}]$ 包含以下关键组件：

任务说明 $\mathcal{I}$ 概述了响应病人的任务，并解释了其余提示的结构。历史细节 $\mathcal{H}$ 总结了对话历史中的关键要素，包括对话上下文 $\bar{c}_t$ ，以及依次列出的历史实体 $\bar{x}_t$ 和行为 $\bar{a}_t$ 。明显细节 $\mathcal{K}$ 提供生成响应的医学知识，包括预测的实体和行为，以及来自 MedKG 的相关知识三元组。相关演示 $\mathcal{E}$ 提供了一个上下文示例来指导响应格式。

为了实现对不同患者状况的实时适应，我们通过引入三重滤波器和演示选择器模块来整合动态提示调整策略，以优化提示中配备的知识和演示。

3.3.2 三重态滤波器

为了从检索到的实体中获得可靠的知识三元组，我们设计了一个迭代过滤过程。

首先，检索到的单跳子图 $G^0_{\bar{x}_t}$ 被转换为一组三元组 $Tri^0_{\bar{x}_t}$ 。接下来，我们计算这些三元组中每个实体的频率，并按降序对其进行排序。基于这些频率，我们通过设置一个阈值 τ 来动态调整保留的三元组。在且仅在三元组的头实体和尾实体的频率都不小于 τ 时，这些三元组才会被保留。

$$Tri^{\tau}_{\bar{x}_t} = \{(e_{head}, r, e_{tail}) | \min(\#e_{head}, \#e_{tail}) \geq \tau\} \quad (5)$$

最初，τ 被设置为 1，并在每次迭代中递增，逐渐减少保留的三元组数量。一旦 $Tri^{\tau}_{\bar{x}_t}$ 中三元组的数量不超过预定义的最大值 M ，过程就终止。当前的 $Tri^{\tau}_{\bar{x}_t}$ 随后作为提示中的最终显著细节的一部分被使用。

3.3.3 演示选择器

为了选择与系统提示最相关的示例，我们引入了一个多步骤的对齐过程。

我们首先根据第一次病人发言中的实体注释将所有训练对话组织成子集。具体来说，我们构建了多实体子集 $S_E = \{S_{E_1}, \ldots, S_{E_K}\}$ ，其中每个子集 S_{E_k} 包含第一次发言中包含相同 n 实体 $E_k = \{x_1, \ldots, x_n\}$ 的对话案例。同时，我们创建了单实体子集 S_e ，其中每个子集 $S_{e'}$ 包含那些提到共同实体 e' 的对话案例的第一次发言。

给定一个当前对话上下文 $\bar{c}_t$ ，我们需要检查其第一句 u_1 是否完全匹配 S_E 中的任何实体集。如果匹配，我们检索对应的子集作为候选示例集 S_{demo} 。否则，我们回退到单实体子集，并从 S_e 中选择至少与 u_1 共享一个实体的所有会话。

为了优化示例选择，我们计算当前首句 u_1 与 S_{demo} 中的语义相似度。我们分别对每个候选句进行编码，然后应用余弦相似度来识别最接近的对话 c_{full} 作为示例参考。

为了提高上下文相关性和减少提示长度，我们使用大小为 ξ 的滑动窗口从 c_{full} 中提取一个重点片段。设 c_{full} 的总话语序列为 $\{u_1, r_1, u_2, r_2, \ldots, u_T, r_T\}$ ，并将起始索引表示为 $i_s = 2t - 1$ ，对应当前对话轮次 t 。最终的示例 $\mathcal{E}$ 在三种情况下被截取：（1）如果 $i_s \leq \xi$ ，我们从 c_{full} 中选择前 2ξ 个话语；（2）如果 $\xi < i_s < T - \xi$ ，我们选择从索引 $i_s - \xi$ 到 $i_s + \xi$ 的话语；（3）如果 $T - \xi \leq i_s$ ，我们选择最后的 2ξ 个话语。

为了优化 MedRef 的不同模块，我们设计了一个两阶段的训练目标。首先，我们预训练实体行为联合预测模块，以为后续的响应生成做好准备。对于预测医学实体，我们计算预测 $\hat{x}_t$ 与真实实体标签 x_t 之间的二元交叉熵（BCE）损失 $\mathcal{L}_x$ 。类似地，基于交叉熵损失 $\mathcal{L}_a$ 训练对话行为预测。这些损失函数可以表述为：

$$\begin{aligned} \mathcal{L}_x &= -\sum_{t=1}^{T}\sum_{i=1}^{|X|}[x_{t_i}\log(\hat{x}_{t_i}) + (1 - x_{t_i})\log(1 - \hat{x}_{t_i})], \\ \mathcal{L}_a &= -\sum_{t=1}^{T}\sum_{j=1}^{|A|}[a_{t_j}\log(\hat{a}_{t_j}) + (1 - a_{t_j})\log(1 - \hat{a}_{t_j})], \end{aligned} \quad (6)$$

为了确保知识提炼的一致性，我们最小化先验 p_θ 和后验 q_ϕ 之间的 Kullback-Leibler (KL) 散度：

$$\mathcal{L}_{kl} = \sum_{t=1}^{T} D_{KL}(q_\phi(z_t|\mu_\phi, \Sigma_{\phi_t}) || p_\theta(z_t|\mu_\theta, \Sigma_{\theta_t})). \quad (7)$$

我们为每个损失分配权重 λ_x 、λ_a 和 λ_{kl} ，因此该阶段的总体损失函数是一个加权组合：

$$\mathcal{L} = \lambda_x\mathcal{L}_x + \lambda_a\mathcal{L}_a + \lambda_{kl}\mathcal{L}_{kl}. \quad (8)$$

接下来，在固定预测模块的情况下，我们对负责响应生成的医疗大语言模型进行微调。通过最大化系统响应的对数似然性，基于语言模型的损失由以下公式给出：

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T}\log\sum_{k} p_{gen}(r_{t_k}|r_{t_{<k}}, \mathcal{P}). \quad (9)$$

4 实验设置

4.1 数据集

我们在两个广泛使用的基准 MedDG 和 Kamed 上进行了实验评估。MedDG 包含超过 17,000 条医学对话，标注有 160 个医学实体，分为五类：疾病、症状、药物、检查和属性。官方分

Table 1: MedDG 和 KaMed 数据集的比较结果。“B” =BLEU，“R” =ROUGE，“E-F1” =entity-F1。粗体/下划线数字表示相对于第二好的显著提升 (p -值 <0.01)。

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

为 14,862（训练）、1,999（验证）和 999（测试）个会话。Kamed 包括超过 63,000 个跨越 100 多个科室的对话。根据 DFMed (Xu et al., 2023)，我们移除了隐私敏感数据，得到 29,159（训练）、1,532（验证）和 1,539（测试）个会话。对话行为被标记为 7 种类型：闲聊、告知、询问、提供日常预防建议、陈述所需医学测试、做出诊断和开具药物。

我们将 MedRef 与以下三种基线进行比较：（1）基于 DL 的方法：Seq2Seq (Sutskever et al., 2014)，带注意力的 RNN；VRBOT (Li et al., 2021)，患者状态和医生行为跟踪模型。(2)基于 PLM 的方法：GPT-2 (Radford et al., 2019)，BART (Lewis, 2019)，通用生成模型；DFMed (Xu et al., 2023)，利用交织实体和动作的双流模型。(3) 基于 LLM 的方法：Chatglm3-6B (Du et al., 2022)，在医学对话上微调的通用 LLM；中经 (Yang et al., 2024)，中医对话模型；华佗 GPT-II (Chen et al., 2023)（Baichuan-7B），DISC-MedLLM (Bao et al., 2023)（Baichuan-13B），专业化的医学 LLM；GPT-4o (Hurst et al., 2024)，先进的闭源 LLM。

4.2 评估指标

为了评估模型生成响应的质量，我们使用 BLEU (Papineni et al., 2002) 和 ROUGE (Lin, 2004) 来评估词汇相似性，并使用实体-F1 分数来衡量实体级别的准确性。

我们关注三个关键的人类评估指标：流畅性（FLU）评估谈话的自然和流畅程度；知识准确性（KC）着重于医学术语的正确性；整体质量（OQ）则考虑整体回复的有效性。

我们使用 ChatGLM3-6B 作为我们的响应生成器的主干，其通过 LoRA (rank=8, 矩阵 =32, dropout=0.1) 进行微调，优化器为 AdamW (学习率 =5e-5)。MedBERT 被用于实体和行为预测（学习率 =3e-5，批量大小 =8）。我们从 CMeKG 中最多检索 25 个三元组。滑动窗口大小为 2。损失权重设置为相应的值。

5 实验结果

5.1 总体表现

如表 1 所示，MedRef 在多个指标上始终优于所有基线，证明其在生成高质量、医学依据明确的回答方面的有效性。相比于 GPT-4o，MedRef 实现了 +1.32 % BLEU-1、+16.08 % ROUGE-1 和 +11.05 % Entity-F1，展示了更好的词汇对齐、流畅性和医学准确性。这个优势源于任务特定的微调，而 GPT-4o 的闭源性质限制了其对医学对话细微差别的适应性。MedRef 倾向于生成流畅的表达，与人类撰写的回答很好地契合，提升了其 ROUGE 和实体-F1 分数，反映了内容的丰富性和相关性。

然而，MedRef 在 KaMed 的 BLEU 分数上略逊于 HuatuoGPT-II 和 GPT-4o。这种差异可能源于数据集的复杂性和这些模型的响应风格偏差。首先，KaMed 涵盖了更广泛的临床场景，涉及超过 100 个科室，这增加了所需医学知识的复杂性，并使得学习高覆盖率的表示更加具有挑战性。此外，HuatuoGPT-II 和 GPT-4o 往往会生成冗长的问答式回复。虽然这种冗长能够增加与参考的词汇级重叠（从而提高 BLEU 分数），但它往往引入无关或冗余内容，导致实体 F1 分数显著降低。其次，HuatuoGPT-II 和 GPT-4o 倾向于采用问答式的方法来处理患者咨询，通常生成非常长的冗余和无意义的文本回复。这种响应趋势不足以使 BLEU 指标略有提高，但显著降低了实体 F1 分数。

为了研究所提出系统中每个模块的贡献，我

Table 2: MedRef 在 MedDG 和 KaMed 数据集上的消融结果。

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

们进行了一项全面的消融研究，其中包括以下用于比较的变体：(1) w/o KRM 删除了知识精炼机制。(2) w/o Demo 删除了由示例选择器匹配的示例 $\mathcal{E}$ 。(3) w/o Kg 删除了从 MedKG 检索到的知识三元组。(4) E-A & Cxt 仅保留预测的实体和动作以及对话上下文，不提供示例或外部知识，并且不使用 KRM。(5) Cxt 仅使用对话上下文，没有任何附加指导或知识。

表 2 中的消融结果显示，所有模型变体都表现出明显的性能下降，强调了每个组件的重要性。特别是，没有 KRM 的模型在所有评估指标上都有显著下降，突显出其在过滤冗余知识和提高实体预测准确性上的双重作用。此外，其他模型变体相对于完整模型的性能下降说明了提示完整性的重要性，并且还表明，选择到我们提示中的检索知识和示例比以前更相关。

5.2 Analysis of Triplet Filter and Demo Selector

为了进一步验证我们的三元组过滤器和演示选择器模块的有效性，我们引入了两个新的模型变体：（1）弱 Kg：此变体不是完全从提示中移除知识三元组（没有 Kg），而是绕过过滤规则，直接从知识图中检索与最近一次话语中实体连接的三元组，随机从一跳连接中选择 M 个三元组。（2）弱 Demo：在这个变体中，演示示例是随机选择的，没有任何对齐过程来确保相关性。

图 3 中的结果显示，两个变体在关键指标上都表现出显著的性能下降。值得注意的是，我们观察到，仅仅增加知识三元组的数量而不应用三元组过滤器，会损害模型的性能。这表明，知识的不加区分使用可能引入噪音，使模型不堪重负，并降低其生成准确响应的能力。同样，随机选择的示例也导致生成质量的下降，这突显了示例选择器对齐过程的重要性。这些发现证实了三元组过滤器和示例选择器对于提高生成医学对话的准确性和相关性都是必不可少的。

5.3 案例研究

图 4 展示了来自 MedDG 数据集的一个运行示例，展示了多轮对话。

在第一回合中，MedRef 展示了其专注于关键实体“痔疮”和“疼痛”的能力，生成的回复与真实情况非常吻合。相比之下，DFMed 和基线模型都未能完全捕捉到这些实体，导致回复不完整。这凸显了 MedRef 在实体预测能力上的优越性及生成更全面询问以更好地解决患者关注问题的能力。

在第二轮交流中，DFMed 正确预测了实体“痔疮”，但忽略了患者早先的陈述“我从未得过痔疮”，导致了一个矛盾的回应。相比之下，MedRef 保持了一致性与患者当前的健康信息一致，从而导致了更准确且在情境上更合适的诊断。

在第五轮中，MedRef 依然能够提供相关且信息丰富的回应。这主要归功于其对检索知识的有效利用以及从整体上下文中推断信息的能力。此示例进一步展示了 MedRef 的稳健性，展示了其处理显式实体提示缺失的情况的能力，但仍能提供有意义且准确的对话。

总的来说，这些案例强调了 MedRef 的优势，不仅在于预测相关的医疗实体，还在于在整个对话中保持上下文的一致性，从而导致更可靠和以患者为中心的互动。这说明了 MedRef 如何超越现有的基准模型，这些模型通常在保持上下文一致性和全面解决患者问题上存在困难。

除了自动评估外，我们还与一个专门团队进行了人工评估实验。志愿者都是拥有丰富医学对话标注经验的医学博士和硕士生，他们在过去几年中致力于相关项目，并能确保正确判断的可靠性。评估人员的任务是对回答进行评分，并根据上述三个指标（流畅性 FLU、知识性 KC、整体质量 OQ）使用 1（差）到 5（优

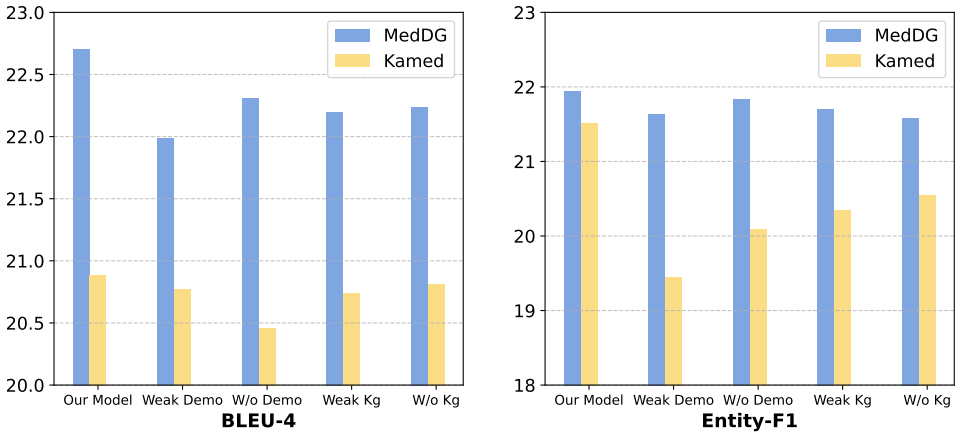

Figure 3: 三重态滤波器和演示选择器的比较结果。

	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: 一个运行案例将 MedRef 与基线进行比较，突出了 MedRef 可以预测更准确的医学实体并生成更相关的响应。

秀）的分级进行评估。

如表 3 所示，MedRef 在所有三个指标上始终优于其他基线模型。值得注意的是，MedRef 的得分与真实响应最接近，这表明其与专家预期的高度一致性。这强化了 MedRef 的专业化设计理念，特别是实体感知机制的整合和动态提示调整，导致更可靠和上下文相关的响应。

从此次评估中获得的一个关键见解是，我们框架的提示设计和动态调整显著提升了大型语言模型（LLMs）的生成质量。结果表明，仅仅使用通用提示微调 LLMs 不足以处理 MDS 的复杂性。相反，MedRef 采用了量身定制的提示策略和知识精炼，使其能够生成不仅更流畅且具备更高医学准确性的回应。这些发现突出了我们系统的优势，表明实体预测、知识精炼和上下文感知提示的结合能够比简单微调策略生成更高质量的医学对话。

在本文中，我们提出了 **Med**ical dialogue system with knowledge **Ref**ining and dynamic prompt adjustment (MedRef)。我们引入了一种变分知识细化机制，以便更准确地预测医学实

Table 3: 用于人工评估的比较结果。每个指标的范围为 1 到 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	3.55	3.68	3.79

体和驱动知识的响应。我们还开发了一种动态提示调整方法，该方法能够根据患者不断变化的病情实时调整系统提示，以确保生成更加个性化和语境相关的多轮医学对话。在两个基准上的大量实验验证了 MedRef 可以在文本生成和基于医学实体的指标方面实现最佳性能。这些发现强调了 MedRef 在提高医学对话系统的质量和可靠性方面的潜力，为促进在医疗环境中更具语境感知和医学准确的互动铺平道路。

6

局限性 虽然我们的模型在医学对话生成中达

到了一流的性能，但仍有两个关键限制为未来的改进提供了机会：(1) 与文本医学知识不同，跨模态知识数据尚未被充分探索以增强对病人病情的捕捉。(2) 当前医学对话系统的情感支持能力仍然是被动的而不是主动的。在保持医学准确性的同时，需要适当的安慰策略。

7

伦理考虑 医疗对话系统的开发和部署优先考虑用户安全、隐私和在医疗保健中负责任地使用人工智能。用于训练的所有数据均已匿名化。该系统被明确旨在作为辅助工具，而不是专业医疗建议的替代品，应与合格的医疗保健提供者的咨询一同使用。

8

致谢

本项工作得到了北京优秀青年科学家计划NO. BJJWZYJH012019100020098，以及智能社会治理平台，“双一流”建设创新交叉平台的重大创新 & 规划，来自中国人民大学的中央高校基本科研业务费和中国人民大学公共计算云资金，中国人民大学世界一流大学（学科）建设基金的支持。

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*.

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.

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.