

# Faithful Medical Dialogue Generation using Homo-Heterogeneous Exemplar-based In-Context Knowledge Grounding

Priyanshu Priya, Hardik Goyal, Asif Ekbal  
Department of Computer Science and Engineering  
Indian Institute of Technology Patna, India  
{priyanshu\_2021cs26,hardik\_2001cs27,asif}@iitp.ac.in

## Abstract

The growing reliance on tele-healthcare has heightened the demand for accessible and professional health support. Artificial Intelligence (AI)-assisted medical dialogue systems have emerged as key solutions, with Large Language Models (LLMs) advancing the generation tasks. However, their susceptibility to hallucination leads to inaccurate and unreliable information, posing major challenges. To address this, we propose a novel approach to mitigate hallucinations in LLMs by integrating external knowledge and in-context learning mechanisms for faithful medical dialogue generation (MDG). In particular, we devise an In-context Medical Knowledge-grounded Dialogue Generator (IMKDG), a novel plug-and-play retrieval-based framework that leverages external medical knowledge, in-context learning (ICL), and retrieval methods to enable LLMs to generate faithful responses, thereby enhancing their performance on the MDG task. We utilize large-scale medical knowledge based on the Unified Medical Language System (UMLS) to retrieve knowledge pertinent to the dialogue context. Further, to enhance the LLMs' ICL capability for the MDG task, we propose the Homo-Heterogeneous Exemplar Selection (H2ES) method, a novel in-context exemplar retrieval method based on both dialogue context and medical knowledge. Automatic and human evaluations on the MedDialog-EN and CDialog datasets across various LLMs demonstrate the efficacy of the proposed framework in mitigating hallucinations.

**Keywords:** Faithful Medical Dialogue Generation, Hallucination, Knowledge Grounding, In-context Learning

## 1. Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency in various medical assessments and multiple-choice question-answering (QA) tasks (Singhal et al., 2022; Gao et al., 2023; Wu et al., 2024; Thirunavukarasu et al., 2023). The application of LLMs for e-medical consultation has garnered considerable attention in recent years (Estimates, 2020). However, the exploration of LLMs in medical dialogue generation (MDG), which closely simulates real-world medical interactions, remains relatively limited. Despite the impressive capabilities of LLMs, they often produce erroneous or fabricated information, a phenomenon referred to as hallucination (Rawte et al., 2023). This issue compromises their reliability and poses a significant challenge for professionals utilizing them in applications, such as medical dialogue generation systems. Figure 1 depicts a case of hallucination during the MDG task.

Hallucinations may seem harmless in casual conversation, but in medical settings, where precision is paramount, they pose significant risks. Misinformation can have dire consequences, affecting patient care and potentially endangering lives. Therefore, the precision and faithfulness of information from LLMs are critical as they affect healthcare decisions, diagnoses, and treatment strategies. Consequently, developing tech-

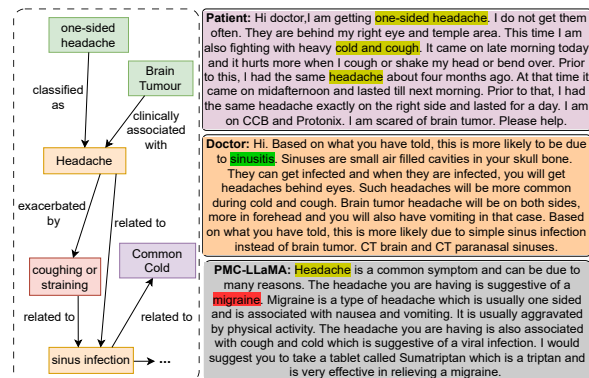


Figure 1: An instance of dialogue (right) where the medical agent acquires accurate knowledge (left) yet generates the hallucinated response. One-sided headed/headache with heavy cough and cold is related to sinus infection rather than a migraine, which contradicts the provided knowledge.

niques to assess and address hallucinations in LLMs is imperative for an enriched and faithful MDG task. While recent efforts have been made to mitigate hallucinations in LLMs (Shuster et al., 2021; Li et al., 2023; Liu et al., 2022a), there remains a scarcity of comprehensive studies in this direction for the faithful MDG task. Incorporating external and authoritative knowledge enhances accuracy and faithfulness in LLM outputs (Kang et al.,

2023; Ji et al., 2023; Priya et al., 2025; Hossain et al., 2025). However, merely providing the LLM with pertinent knowledge is insufficient. As depicted in Figure 1, the LLM may still generate inaccurate responses despite having pertinent knowledge. This underscores the need to effectively utilize knowledge to prevent hallucinations in LLMs. With the remarkable few-shot learning capabilities of LLMs, recently, the research focus has shifted towards in-context learning (ICL) (Brown et al., 2020; Agrawal et al., 2022). ICL enables the generation of desired responses by integrating task definitions, in-context exemplars, and target context. The effectiveness of ICL hinges on a sophisticated exemplar-retrieval approach informed by both dialogue context and domain knowledge.

Driven by these considerations, we introduce a novel plug-and-play retrieval-based framework, named In-context Medical Knowledge-grounded Dialogue Generator (IMKDG). IMKDG aims to retrieve pertinent medical knowledge and high-quality in-context exemplars for a particular dialogue context to guide LLMs toward faithful dialogue generation. IMKDG embodies 3 key modules: (i) a *Medical Knowledge Enrichment* (MKE) module that identifies medical entities in context and retrieves relevant knowledge from the medical knowledge graph based on Unified Medical Language System (UMLS) (Bodenreider, 2004) to enrich dialogue with accurate medical insights, (ii) a *Homo-Heterogeneous Exemplar Selection* (H2ES) module that retrieves relevant and diverse exemplars based on dialogue context and associated knowledge, and (iii) a *Prompt Construction* (PC) module that formulates the final input prompt for the LLM.

We evaluate IMKDG using seven LLMs: BioGPT (Luo et al., 2022a), ChatGPT (OpenAI, 2024), MedAlpaca (Han et al., 2023), ChatDoctor (Yunxiang et al., 2023), PMC-LLaMA (Wu et al., 2023), MedPaLM 2 (Singhal et al., 2023), and GPT-4 (Achiam et al., 2023), on two widely used large-scale medical dialogue datasets, MedDialog-EN (Zeng et al., 2020) and CDialog (Varshney et al., 2022b). Both automatic and human evaluations demonstrate that IMKDG effectively reduces hallucinations, thereby improving fidelity in the MDG task. Notably, this study’s primary focus is not on selecting black-box LLMs and datasets; instead, we aim to explore the universality of IMKDG for the faithful MDG task.

To summarize, the key contributions are: (i) Investigate the knowledge-grounding and ICL-based approach for faithful MDG task. To the best of our knowledge, no prior research (Liu et al., 2022c; Ji et al., 2023; Kang et al., 2022) has investigated the potential of both knowledge-grounding and ICL for faithful MDG task; (ii) Devise IMKDG (In-

context Medical Knowledge-grounded Dialogue Generator), a novel plug-and-play retrieval-based framework to effectively integrate external medical knowledge using LLMs’ ICL capability to mitigate hallucinations during MDG; (iii) Present a novel Homo-Heterogeneous Exemplar Selection (H2ES) method for enriching the quality of selected exemplars to boost LLMs’ ICL performance; (iv) Conduct a thorough automatic and human evaluations and empirically show that IMKDG substantially reduces hallucinations in MDG task; (v) IMKDG applicability can be effortlessly extended to other LLMs and datasets<sup>1</sup>.

## 2. Related Work

There have been efforts to develop medical dialogue generation models (Li et al., 2021; Liu et al., 2022b; Priya et al., 2023; Mishra et al., 2023a,c,b). Recent advancements in LLMs have significantly transformed medical dialogue systems (Shi et al., 2024). Dou et al. (2023) explored ICL with LLMs for generating physician responses. Integration of knowledge graphs (KGs) into LLMs is a growing trend aimed at improving the factual accuracy of generated content (Pan et al., 2024). For instance, Gao et al. (2023) integrated KGs into LLMs for diagnosis prediction using a novel graph model that leverages path-based prompts. Wu et al. (2024) utilized medical KGs and ICL for the Chinese QA task, employing the BM25 method (Robertson and Walker, 1994) to retrieve relevant exemplars.

One of the significant challenges in utilizing LLMs for medical dialogue generation is the occurrence of hallucinations, where models generate inaccurate or unfaithful information. Researchers have devoted significant effort to mitigating hallucinations (i.e., improving faithfulness) in dialogue systems by integrating external knowledge. Xu et al. (2022) employed a lightweight adapter as a knowledge expert to boost GPT-2 (Radford et al., 2019). Wu et al. (2021) introduced an inductive attention mechanism for self-attention-based generation models. Ji et al. (2023) incorporated additional modules, namely knowledge grounding and response re-ranking modules, to mitigate hallucination. While these works integrate knowledge to reduce hallucination, these studies employ additional knowledge encoding modules, potentially leading to error propagation and the omission of critical details in generated responses.

Unlike existing studies, the present work uses external medical KGs and leverages LLMs’ ICL capabilities to achieve faithful MDG. We introduce a novel plug-and-play retrieval-based framework

---

<sup>1</sup>Resources are available at [https://github.com/priyanshu-profile/LREC26\\_IMKDG](https://github.com/priyanshu-profile/LREC26_IMKDG)

that integrates relevant knowledge directly into responses, thereby reducing error propagation. Further, we introduce a sophisticated exemplar selection method to enhance LLMs’ ICL performance for effectively incorporating knowledge and reducing factual inaccuracies. This approach represents a significant advancement in integrating structured medical knowledge with advanced LLMs for faithful dialogue generation in medical contexts.

### 3. Methodology

#### 3.1. Problem Formulation

In the present work, we formulate the medical knowledge-grounded dialogue generation task within an in-context learning paradigm. In this task setting, for a given data sample consisting of a dialogue context  $\mathcal{H} = \{p_1, d_1, p_2, d_2, \dots, p_{k-1}, d_{k-1}, p_k\}$  as an alternating sequence of utterances between the patient ( $p$ ) and doctor ( $d$ ) and the associated knowledge  $\mathcal{K}$ , the objective is to generate faithful and informative response  $d_k = \{w_1, w_2, \dots, w_n\}$  (with  $n$  words) based on  $\mathcal{H}$  and  $\mathcal{K}$ . Formally, it is defined as:

$$\mathbb{P}_{LLM}(d_k|\mathcal{H}, \mathcal{K}) = \prod_{i=0}^n \mathbb{P}_{LLM}(w_i|\mathcal{H}, \mathcal{K}, w_{<i}) \quad (1)$$

$$\mathbb{P}_{LLM}(w_i|\mathcal{H}, \mathcal{K}, w_{<i}) = \sum_{\mathcal{E}} \mathbb{P}_{LLM}(w_i|\mathcal{H}, \mathcal{K}, \mathcal{E}, w_{<i}) \mathbb{P}_{IES}(\mathcal{E}|\mathcal{H}, \mathcal{K}) \quad (2)$$

where,  $\mathcal{E}$  represents the retrieved in-context exemplars given  $\mathcal{H}$  and  $\mathcal{K}$ .  $\mathbb{P}_{IES}(\mathcal{E}|\mathcal{H}, \mathcal{K})$  represents the joint probability distribution of in-context exemplars, where  $IES$  denotes the in-context exemplar selection (IES) method. In this work, we characterize the inter-dependence among exemplars; hence, we define  $\mathbb{P}_{IES}(\mathcal{E}|\mathcal{H}, \mathcal{K})$  as follows:

$$\mathbb{P}_{IES}(\mathcal{E}|\mathcal{H}, \mathcal{K}) = \prod_{j=1}^t \mathbb{P}_{IES}(\mathcal{E}_j|\mathcal{H}, \mathcal{K}, \mathcal{E}_{j-1}) \quad (3)$$

By doing this, we ensure the dependencies inherent in the in-context exemplar selection method, indicating that the current exemplar selection depends on the target context and previously selected exemplar.

#### 3.2. In-context Medical Knowledge-grounded Dialogue Generator (IMKDG)

The overall architecture of the In-context Medical Knowledge-grounded Dialogue Generator (IMKDG) is shown in Figure 2. IMKDG comprises three main modules: Medical Knowledge Enrichment, In-context Exemplars Selection, and Prompt Construction. These modules are detailed as:

**Medical Knowledge Enrichment (MKE).** The MKE module is designed to construct a medical knowledge graph (MKG) for a given dialogue context to enrich the MDG task with reliable medical knowledge. We use the Quick-UMLS<sup>2</sup> tool to build the MKG by extracting necessary medical concepts from the UMLS database (Bodenreider, 2004). The Quick-UMLS identifies the biomedical entities and links them with unique identifiers (CUIs) and semantic types from the UMLS Metathesaurus. We feed the dialog context as input, then extract an approximate match in the set of strings from UMLS, and finally return the CUI, a list of semantic types for each concept in the dialogue context. We regard each distinct identifier (CUI) within the UMLS as a node within our MKG. We then utilize the Semantic Network (SN) module in the UMLS to determine the pertinent relationships between these nodes. Each MKG comprises a set of triplets and each triplet is of form  $(e_1, r, e_2)$ , where  $e_1$ , and  $e_2$  denote *head*, and *tail* entities, respectively and  $r$  is the relationship between them.

For instance, given an input “*I have anxiety and depression but no immunodeficiency disorders or chronic diseases. First I want to know if my immune system is weakened and how likely am I to do of Coronavirus?*”. A few identified concepts are “*anxiety*”, “*depression*”, “*coronavirus*”, etc. We then employ the SN module to ascertain the relationships among nodes to construct the graph corresponding to the medical concept “*coronavirus*”, resulting in the generation of the set of triplets as [*coronavirus, isa, infection*], [*coronavirus, related\_to, immune system*], [*coronavirus, affects, respiratory system*] and so on.

**Homo-Heterogeneous Exemplar Selection (H2ES).** Given that the meticulous selection of exemplars is imperative for improving the LLMs’ ICL capabilities (Brown et al., 2020; Liu et al., 2022c), this module is designed to retrieve the contextually appropriate in-context exemplars for the MDG task. In the dialogue generation, contextually relevant and diverse responses facilitate user engagement and retention, contributing to the conversation’s smooth progression (Liu et al., 2023). Therefore, it is crucial to ensure that the chosen exemplars are both relevant and diverse. Drawing inspiration from recommendation systems that involve multi-objective recommendation tasks, which typically entail ensuring both relevance and diversity in retrieval outcomes (Sá et al., 2022), we propose a novel Homo-Heterogeneous Exemplar Selection (H2ES) method. For a particular dialogue context, H2ES first retrieves  $L$  semantically relevant exemplars,  $\mathcal{E}_{rel} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_L\}$

<sup>2</sup><https://github.com/Georgetown-IR-Lab/QuickUMLS>

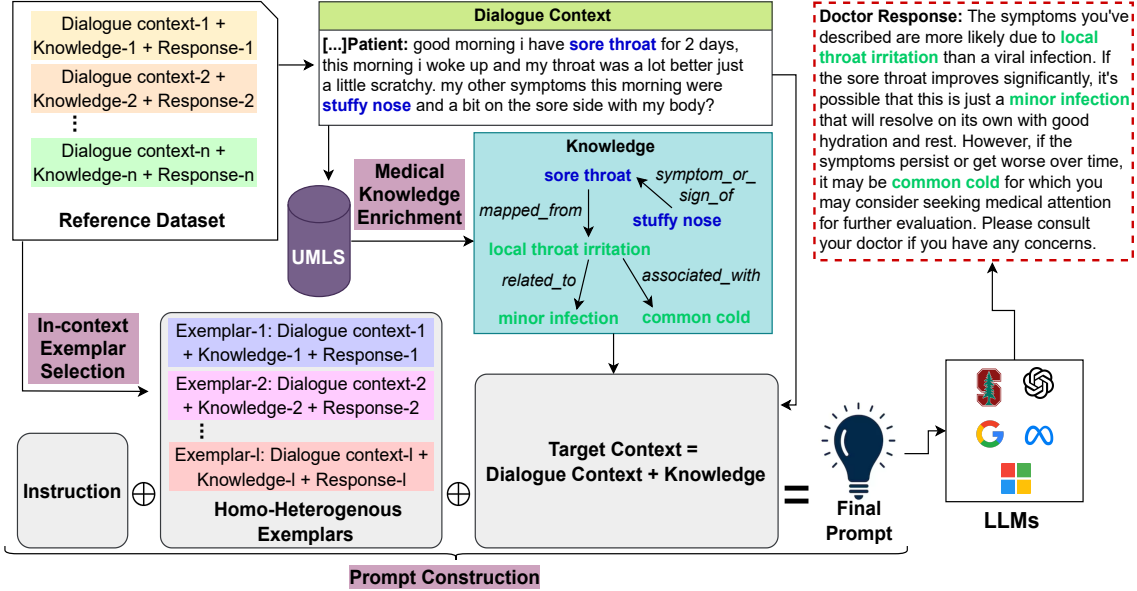


Figure 2: An overview of the proposed In-context Medical Knowledge-grounded Dialogue Generator (IMKDG) framework. It involves three components - (i) Medical Knowledge Enrichment, (ii) In-context Exemplar Selection using the Homo-Heterogeneous Exemplar Selection method, and (iii) Prompt Construction for designing the final prompt for guiding the LLM for faithful response generation.

and then find a diverse subset of exemplars,  $\mathcal{E}_{div} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_l (L \gg l)\}$ .

To select semantically relevant exemplars, we employ a query-based exemplar selection approach, which aims to search similar instances from a database of instances  $D$  based on the input query ( $q$ ). We concatenate the dialogue context  $\mathcal{H}$  and knowledge  $\mathcal{K}$  of each instance into a single sequence to form the query  $q$ , and utilize a pre-trained RoBERTa (Liu et al., 2019) as the sentence encoder (**Enc**) to obtain semantic representations for the query  $q$ , and each data instance  $d_i$  in  $D$ . Then, we compute cosine similarity between semantic representations of  $q$  and each ( $d_i \in D$ ):

$$Sim(q, d_i) = \frac{\mathbb{E}_q \cdot \mathbb{E}_{d_i}}{\|\mathbb{E}_q\| \|\mathbb{E}_{d_i}\|} \quad (4)$$

where,  $\mathbb{E}_q$  and  $\mathbb{E}_{d_i}$  are the semantic representations of  $q$  and  $d_i$ , respectively. Finally, we select  $L$  instances with the highest similarity scores to  $q$  to obtain relevant exemplars,  $\mathcal{E}_{rel} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_L\}$ . Afterward, to select a diverse subset of exemplars, we filter out the exemplar with the highest similarity score and add it to the queue. We then determine the distance ( $S$ ) between the remaining exemplars and the last exemplar in the current queue, yielding the diversity score ( $Div_{score}$ ). The final score,  $\mathcal{F}_{score}$  is a linear combination of distance score and similarity score as:

$$\begin{aligned} \mathcal{F}_{score} &= (1 - \lambda) * Div_{score} + \lambda * Sim(q, \mathcal{E}_i) \\ &= (1 - \lambda) * S(\mathcal{E}_i, \mathcal{E}_{last}) + \lambda * Sim(q, \mathcal{E}_i) \end{aligned} \quad (5)$$

where  $\mathcal{E}_{last}$  denotes the last element in the queue. Finally, we place the exemplar with the highest final score at the end of the queue. We iterate this

procedure until there are  $l$  exemplars in the queue. This procedure ensures that the selected exemplars are not only relevant to the input dialogue context and associated knowledge but also distinct from those already chosen. A formal description of the **H2ES** method is given in Algorithm 1.

**Prompt Construction (PC).** Inspired by the few-shot learning approach (Brown et al., 2020), providing LLMs with suitable and intuitive prompts enables them to generate relevant content. Hence, a PC module is devised to formulate the final prompt template for the MDG task. The prompt comprises task-specific instruction ( $I$ ), the selected in-context exemplars ( $\{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_l\}$ ), and the target context ( $input$ ). Concretely, the task-specific instructions that guide the LLM for response generation grounded in the medical knowledge are concatenated with the retrieved  $l$  exemplars (placed in descending order of their final score) and the target context to form the final prompt to be fed to LLMs. For prompts for the selected exemplars ( $\mathbf{P}_{\mathcal{E}_j}, j \in [1, l]$ ), we use ‘Doctor:’ and ‘Patient:’ to connect different turns in the dialogue context, and ‘We know that:’ and ‘Doctor Response:’ to introduce the knowledge and doctor’s response, respectively. To design a prompt for the target context, we follow the same template except that we keep the ‘Doctor Response:’ empty for the LLM to generate. To create the final prompt, we concatenate the prompts for selected exemplars and the target using “\n” and then feed them into the LLM to generate the response as follows:  $R = LLM(I \setminus \mathbf{P}_{\mathcal{E}_1} \setminus \dots \setminus \mathbf{P}_{\mathcal{E}_l} \setminus \mathbf{P}_{input})$ . A sam-

ple prompt example is depicted in Figure 3.

Instruction

Generate the Doctor's response for the following patient doctor conversation based on the given examples:

ICL Exemplars

**Dialogue Context**  
 Patient: What extra additional precautions do they need to prevent catching novel coronavirus?  
 Doctor: COVID-19 is now the official name for the illness caused by the newly discovered coronavirus (COronaVirus Infectious Disease - 2019). So far, it is extremely rare in the US (2/12/20). Until and unless COVID-19 becomes common no special precautions are necessary.  
 Patient: For students in dorms living with students from many countries?  
**We know that:**  
 [Infectious Disease, focus\_of, Respiratory secretion precautions], [Infectious Disease, possibly\_equivalent\_to, Non-human infectious disorder], [Infectious Disease, focus\_of, Handwashing precautions], [Infectious Disease, associated\_with, infectious bronchiolitis], [Infectious Disease, focus\_of, Standard precautions]  
**Doctor Response:** In any dormitory or group living situation people with respiratory symptoms (colds, flu, etc.) should cover their coughs and wash their hands frequently.

Target Context

**Dialogue Context**  
 Patient: Am I at risk for COVID-19 in the United States?  
 Doctor: COVID infections will likely pick up over the next few weeks in the US.  
 Patient: What should I do in order to avoid infection?  
**We know that:**  
 [Infection, focus\_of, Handwashing precautions], [Infection, focus\_of, Standard precautions], [Infections, affects, Traveler's Health], [Infection, focus\_of, Contact precautions], [Infection, related\_to, Traveler's Health]  
**Doctor Response:**

Figure 3: Sample prompt used to prompt IMKDG.

---

**Algorithm 1** Homo-Heterogeneous Exemplar Selection (H2ES)

---

**Input:** query  $q$ , database of instances  $D, L, l$   
**Output:** Homo-Heterogeneous Exemplars -  $\mathcal{E}_{div} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_l (L \gg l)\}$   
**Enc:** Sentence Encoder  
**Sim:** Cosine Similarity  
**Q:** Queue

1. BEGIN
2.  $\mathcal{E} = \{\}$
3.  $\mathbb{E}_q = \mathbf{Enc}(q)$
4. **for each**  $d_i$  in  $D$  **do**
5.    $\mathbb{E}_{d_i} = \mathbf{Enc}(d_i)$
6.    $Sim_{score} = \mathbf{Sim}(\mathbb{E}_q, \mathbb{E}_{d_i})$
7.    $\mathcal{E}.add(d_i)$
8.  $\mathcal{E}_{rel} \leftarrow$  Select  $L$  exemplars from  $\mathcal{E}$  with highest  $Sim_{score}$
9. Select exemplar  $\mathcal{E}_1$  with the highest similarity score:  
 $Q.insert(\mathcal{E}_1), \mathcal{E} = \mathcal{E}/\mathcal{E}_1$
10. **while**  $length(Q) < l$  **do**
11.    $\mathcal{E}_{temp} \leftarrow Q[-1]$
12.    $Div_{score} \leftarrow S(\mathcal{E}_{temp}, \mathcal{E}/\mathcal{E}_{temp})$
13.    $Sim_{score} \leftarrow \mathbf{Sim}(q, \mathcal{E}/\mathcal{E}_{temp})$
14.    $\mathcal{F}_{score} \leftarrow (1 - \lambda) * Div_{score} + \lambda * Sim_{score}$
15.    $u \leftarrow \arg \max \mathcal{F}_{score}$
16.    $Q.insert(\mathcal{E}_u), \mathcal{E} = \mathcal{E}/\mathcal{E}_u$
17. **end**
18. **return**  $Q$
19. END

---

## 4. Experimental Setup

**Large Language Models and Datasets.** We evaluate the IMKDG framework using various open-source LLMs: BioGPT (Luo et al., 2022b), MedAlpaca-7B (Han et al., 2023), ChatDoctor (Yunxiang et al., 2023), and PMC-LLaMA-7B (Wu

et al., 2023), alongside commercial LLMs including ChatGPT (gpt-3.5-turbo) (OpenAI, 2024), MedPaLM 2 (Singhal et al., 2023), and GPT-4 (gpt-4-turbo) (Achiam et al., 2023). Our experiments utilize two publicly available medical dialogue datasets: MedDialog-EN (Zeng et al., 2020) and CDialog (Varshney et al., 2022a).

**Compared Models.** We employ three categories of baselines: Fine-tuning-based Baseline Models (FBM), viz. BERT (Devlin et al., 2019), DialoGPT (Zhang et al., 2020), BART (Lewis et al., 2019), BioBERT (Lee et al., 2019) and RHO (Ji et al., 2023), Prompting-based Baseline Models (PBM), viz. Plug-and-play Language Model (PPLM) (Dathathri et al., 2020), and Multi-Stage Dialogue Prompting for Knowledge-grounded Generation (MSDP-KG) (Liu et al., 2022d), and vanilla LLMs, including all seven LLMs without knowledge grounding and in-context exemplars to assess the efficacy of the proposed IMKDG framework.

**Implementation Details.** For experiments, we split the datasets into a 7:3 ratio for training and testing. The training set serves as database  $D$  for exemplar selection and the model is evaluated on the test set. For FBM, we further split the training set into training and dev sets in an 8:2 ratio. We use RoBERTa (Liu et al., 2019) as the text encoder in the H2ES method to generate semantic representations of exemplars and compute the Manhattan distance between them. We experiment with distinct values of  $\lambda = \{0.1, 0.2, 0.4, 0.5\}$  and empirically set  $\lambda = 0.2$ . For all experiments, the temperature is set to 0 to eliminate randomness.

**Evaluation Metrics.** We perform automatic and human evaluations to gauge the general quality of generation and the faithfulness of generated responses. For automatic evaluation of general generation quality, we employ Perplexity (PPL) (Brown et al., 1992), traditional word-overlap-based metrics, viz. BLEU4 (B-4) (Papineni et al., 2002), METEOR (MET.) (Banerjee and Lavie, 2005), and ROUGE-L (R-L) (Lin, 2004), and embedding-based metrics, viz. BERTScore F1 (BS-F1) (Zhang et al., 2019), Embedding Average (EA), Vector Extrema (VE), and Greedy Matching (GM) (Liu et al., 2016). For assessing faithfulness of responses, we employ Entity Coverage F1 (E-F1) (Ji et al., 2023), FEQA (Durmus et al., 2020), and QuestEval (QE) (Scialom et al., 2021). E-F1 measures whether entities in the generated responses align with those in the provided knowledge and dialogue context. To compute E-F1, we utilize a ScispaCy Named Entity Recognition model (AI, 2025) to extract medical named entities in generated responses and the dialogue context, and then compute the exact match E-F1 score between entities in the generated responses and entities in KG and dialogue context. FEQA and QE

are question-answering (QA) metrics that take the source (e.g., a document) and the corresponding output to be evaluated (e.g., a summary or golden answers) as input. We concatenate the dialog context and knowledge into the document, while the response serves as the golden answer for evaluating the response’s faithfulness to the original context. The higher the E-F1, FEQA, and QE scores, the lower the hallucination.

For human evaluation, we employ Fluency (F), Adequacy (A), Contextual Relevance (CR) as the general generation evaluation metrics; and Faithful and Hallucinated (Intrinsic (I), Extrinsic (E), or Both (B)) (Ji et al., 2023) as the faithful generation evaluation metrics. Besides, we assess whether the generated responses are helpful and safe, utilizing Helpfulness (H) and Safety (S), respectively. To assess the quality of generated responses from a human perspective, we randomly select 60 samples from the test set and analyze the predicted responses. To streamline the evaluation process and minimize costs, we divide the human evaluation into two distinct groups of evaluators.

For the general generation evaluation, we employ three human evaluators (hereafter, general evaluators), two with a Ph.D. in Linguistics and one with a Master’s in Computer Science. These evaluators are regular employees working with our research team and have sufficient expertise in similar projects over the past 4 years. Given that faithful, helpful, and safe response generation evaluation requires medical knowledge and expertise, and given the associated high evaluation costs, we engage two medical professionals from our institute for this aspect of the evaluation. These medical experts (hereinafter, expert evaluators) hold postgraduate degrees in medicine. All evaluators are compensated in accordance with the institute’s guidelines.

Before beginning the evaluation process, we brief the evaluators about the different metrics and suitable examples. For each sample, we then provide the dialogue context, associated knowledge, the ground-truth response, and the model-generated response. We then instruct the general evaluators to rate the responses on the scales of 1-5 for F, A, and CR, and 0-2 for H, and 0-1 for S. Further, for faithful evaluation, we request the expert evaluators to identify whether the generated response is faithful or hallucinated. If hallucinated, we request them to further annotate those samples, with the type of hallucination (Intrinsic, Extrinsic, or Both).

## 5. Evaluation Results

**Automatic Evaluation.** Table 1 and Table 2 report automatic evaluation results on MedDialog-

EN and CDialog datasets, respectively. The proposed IMKDG framework, paired with various LLMs, outperforms the FBM, PBM, and vanilla LLM baselines on both datasets, achieving superior scores across various evaluation metrics. Specifically, IMKDG+GPT-4 achieves the best scores across various evaluation metrics on both datasets. Notably, there is a significant drop in the PPL score. IMKDG+GPT-4 attains a PPL of 7.23 on MedDialog-EN and 1.13 on CDialog, indicating drops of 12.01 and 14.91 points, respectively, as compared to the state-of-the-art (SOTA) method: RHO. Besides, IMKDG+GPT-4 achieves better scores than RHO in terms of B-4, MET., R-L, BS-F1, EA, VE, and GM with a significant increase of 0.033, 0.282, 0.169, 14.67, 0.061, 0.069, and 0.075 points, respectively, on MedDialog-EN. Likewise, on the CDialog dataset, the IMKDG+GPT-4 achieves improvements of 0.040, 0.275, 0.212, 8.88, 0.114, 0.105, and 0.104 points over RHO on B-4, MET., R-L, BS-F1, EA, VE, and GM, respectively. The lower PPL suggests a preference for context-relevant responses, while the higher B-4, MET., and R-L scores indicate alignment with gold responses. IMKDG+GPT-4 also excels at embedding-based metrics (BS-F1, EA, VE, and GM), demonstrating its effectiveness at generating contextually relevant, high-quality responses. The superior performance of IMKDG+GPT-4 can be attributed to the use of relevant and diverse in-context exemplars that enhance in-context learning, along with GPT-4’s inherent generation capabilities.

Moreover, the proposed IMKDG+GPT-4 achieves higher scores than RHO in faithful generation metrics. Notably, it achieves E-F1 scores of 85.00% on MedDialog-EN and 85.45% on CDialog, representing increases of 19.61 and 12.94 points, respectively. On the FEQA and QE scores, it also achieves better performance than RHO. These findings suggest that the proposed framework effectively incorporates knowledge and leverages ICL to facilitate faithful medical dialogue generation. Similar performance improvements are observed across both datasets with other LLMs, except for BioGPT (a GPT-based model). One potential explanation could be the absence of ICL ability of smaller LMs, such as GPT-2. Further, since the prompting-based baselines (PPLM and MSDP-KG) also rely on GPT-based methods, their comparatively lower performance may stem from the lack of ICL capability inherent in smaller LMs. It is also evident that vanilla LLMs achieve significantly inferior performance compared to LLMs with the proposed IMKDG framework. These results further underscore the relevance of medical knowledge grounding and of in-context exemplars retrieved using the novel

	Models	PPL	B-4	MET.	R-L	BS-F1	EA	VE	GM	E-F1	FEQA	QE
Fine-tuned Baseline Models	BERT	36.28	0.031	0.346	0.332	61.29%	0.863	0.319	0.683	48.79%	24.65	34.33
	DialoGPT	68.19	0.013	0.291	0.311	58.67%	0.715	0.254	0.561	42.25%	15.55	28.89
	BART	24.74	0.047	0.369	0.365	67.31%	0.869	0.355	0.665	59.18%	32.6	42.55
	BioBERT	22.58	0.049	0.375	0.388	68.24%	0.874	0.362	0.671	59.97%	33.14	43.78
	RHO	19.24	0.052	0.397	0.428	72.11%	0.917	0.378	0.688	65.39%	43.89	48.61
Prompting-based Baseline Models	PPLM	25.78	0.037	0.323	0.346	65.12%	0.761	0.284	0.524	53.28%	20.92	28.29
	MSDP-KG	23.19	0.042	0.378	0.362	67.84%	0.782	0.306	0.558	56.19%	23.44	30.72
Vanilla LLMs	vanilla BioGPT	58.13	0.013	0.012	0.014	42.98%	0.531	0.192	0.382	21.238	11.88	12.47
	vanilla MedAlpaca	36.44	0.021	0.085	0.068	63.52%	0.654	0.088	0.527	54.21%	18.65	23.61
	vanilla ChatDoctor	29.18	0.025	0.182	0.098	66.75%	0.675	0.199	0.541	55.30%	17.56	25.46
	vanilla ChatGPT	27.35	0.037	0.195	0.148	72.56%	0.742	0.212	0.583	61.45%	23.97	31.45
	vanilla PMC-LLaMA	21.53	0.042	0.353	0.453	74.28%	0.827	0.273	0.608	64.88%	24.83	34.18
	vanilla MedPaLM 2	13.89	0.065	0.567	0.498	79.18%	0.899	0.392	0.693	74.31%	31.94	39.2
	vanilla GPT-4	16.30	0.061	0.527	0.425	77.50%	0.875	0.386	0.662	72.60%	27.42	37.94
Proposed Approach	IMKDG + BioGPT	46.30	0.065	0.035	0.056	80.00%	0.817	0.373	0.663	31.86%	19.45	25.82
	IMKDG + MedAlpaca	14.06	0.216	0.188	0.146	83.80%	0.92	0.403	0.737	72.59%	32.08	40.29
	IMKDG + ChatDoctor	13.25	0.046	0.254	0.168	83.64%	0.928	0.385	0.726	75.00%	35.13	43.01
	IMKDG + ChatGPT	12.81	0.052	0.343	0.239	84.17%	0.931	0.388	0.728	77.00%	38.64	44.2
	IMKDG + PMC-LLaMA	11.8	0.062	0.512	0.54	83.53%	0.941	0.39	0.729	80.27%	43.27	52.61
	IMKDG+MedPaLM 2	9.58	0.081	0.657	0.561	85.10%	0.965	0.415	0.746	82.00%	46.97	57.28
	IMKDG+GPT-4	<b>7.23</b>	<b>0.085</b>	<b>0.679</b>	<b>0.597</b>	<b>86.78%</b>	<b>0.978</b>	<b>0.447</b>	<b>0.763</b>	<b>85.00%</b>	<b>50.29</b>	<b>59.94</b>

Table 1: Automatic evaluation results on MedDialog-EN. Bolded values represent the best scores. Results are statistically significant based on Welch’s t-test (Welch, 1947) conducted at a 5% significance level.

	Models	PPL	B-4	MET.	R-L	BS-F1	EA	VE	GM	E-F1	FEQA	QE
Fine-tuned Baseline Models	BERT	39.52	0.011	0.088	0.115	61.16%	0.854	0.351	0.638	51.23%	18.73	27.68
	DialoGPT	51.25	0.006	0.084	0.108	59.82%	0.827	0.342	0.633	48.05%	16.21	24.56
	BART	26.09	0.016	0.142	0.138	64.58%	0.851	0.359	0.682	54.89%	21.15	29.87
	BioBERT	24.74	0.052	0.175	0.170	67.32%	0.870	0.412	0.647	59.72%	24.89	32.91
	RHO	16.04	0.054	0.173	0.266	76.20%	0.843	0.380	0.719	72.51%	38.42	47.65
Prompting-based Baseline Models	PPLM	21.06	0.046	0.098	0.192	71.45%	0.737	0.264	0.594	64.10%	12.83	21.76
	MSDP-KG	19.63	0.041	0.077	0.163	67.21%	0.695	0.207	0.538	69.23%	15.33	26.87
Vanilla LLMs	vanilla BioGPT	39.40	0.014	0.018	0.008	66.50%	0.514	0.246	0.479	13.45%	5.18	8.29
	vanilla MedAlpaca	26.39	0.013	0.081	0.102	68.14%	0.703	0.312	0.651	50.35%	24.14	31.95
	vanilla ChatDoctor	24.53	0.017	0.124	0.129	68.19%	0.714	0.331	0.627	53.97%	22.67	35.92
	vanilla ChatGPT	17.92	0.029	0.129	0.198	70.13%	0.721	0.344	0.733	56.82%	27.22	37.55
	vanilla PMC-LLaMA	18.10	0.034	0.126	0.168	68.50%	0.729	0.327	0.65	61.39%	25.36	33.89
	vanilla MedPaLM 2	9.09	0.076	0.562	0.491	73.93%	0.842	0.393	0.735	71.92%	39.11	53.85
	vanilla GPT-4	10.43	0.083	0.597	0.416	72.01%	0.854	0.395	0.667	69.98%	35.12	44.78
Proposed Approach	IMKDG + BioGPT	28.50	0.026	0.032	0.022	79.94%	0.636	0.358	0.594	25.36%	10.27	18.35
	IMKDG + MedAlpaca	11.43	0.021	0.098	0.139	78.69%	0.821	0.456	0.734	64.37%	32.99	40.84
	IMKDG + ChatDoctor	10.25	0.033	0.142	0.147	79.23%	0.829	0.451	0.739	65.41%	34.08	42.15
	IMKDG + ChatGPT	8.71	0.049	0.236	0.219	81.76%	0.845	0.467	0.742	69.58%	37.59	46.67
	IMKDG + PMC-LLaMA	2.38	0.059	0.287	0.329	82.87%	0.891	0.436	0.777	73.59%	39.73	49.01
	IMKDG + MedPaLM 2	1.58	0.091	0.637	0.561	84.51%	0.945	0.519	0.806	82.17%	47.29	58.28
	IMKDG + GPT-4	<b>1.13</b>	<b>0.094</b>	<b>0.448</b>	<b>0.478</b>	<b>85.08%</b>	<b>0.957</b>	<b>0.485</b>	<b>0.823</b>	<b>85.45%</b>	<b>51.29</b>	<b>61.94</b>

Table 2: Automatic evaluation results on CDialog. Bolded values represent the best scores. Results are statistically significant based on Welch’s t-test (Welch, 1947) conducted at a 5% significance level.

H2ES approach for faithful response generation in medical dialogues.

**Human Evaluation.** Table 3 depicts the human evaluation results of the proposed IMKDG framework with different LLMs on MedDialog-EN and CDialog datasets. We compare our proposed approach against RHO, PPLM, and MSDP-KG only, as human evaluation is expensive. The proposed IMKDG with GPT-4 outperforms the baselines on various evaluation metrics. Specifically, on MedDialog-EN, we observe an increase of 45.60%, 35.17%, and 30.40% in F, A, and CR, respectively, compared to RHO. Likewise, on CDialog, we observe improvements of 16.55%, 25.13%, and 15.19% in F, A, and CR, respectively, compared to RHO. These results signify that the proposed approach consistently generates fluent, adequate, and contextually coherent responses. The high F scores align with the automatic evalua-

tion results, where IMKDG+GPT-4 shows lower PPL and higher B-4, MET., and R-L scores. The high A and CR scores correspond with higher BS-F1, EA, VE, and GM scores. This could be because IMKDG learned much context-relevant knowledge through the knowledge incorporation and in-context learning process, which can provide effective guidance for faithful MDG.

IMKDG+GPT-4 outperforms RHO in faithful evaluation metrics, demonstrating a notable reduction in both extrinsic and intrinsic hallucinations. Specifically, it reduces extrinsic hallucinations by 48.40% on MedDialog-EN and 60.71% on CDialog, and intrinsic hallucinations by 41.40% and 28.78%, respectively. These results suggest that in-context exemplars grounded in both knowledge and dialogue context effectively incorporate knowledge to mitigate hallucinations and ensure faithfulness in the generated responses. Further, the proposed framework demonstrates efficacy in gener-

Models	MedDialog-EN										CDialog											
	F	A	CR	Faithful			Hallucinated			H	S	F	A	CR	Faithful			Hallucinated			H	S
	(1-5)	(1-5)	(1-5)	E	I	B	(0-2)	(0-1)	(1-5)	(1-5)	(1-5)	E	I	B	(0-2)	(0-1)						
RHO	3.18	3.27	3.29	79.86	10.00	6.86	<b>3.28</b>	1.73	1.00	4.11	3.74	3.95	80.82	9.34	4.83	5.01	1.84	1.00				
PPLM	2.32	2.79	2.69	64.58	14.09	11.32	10.01	1.24	1.00	2.93	2.73	2.72	68.53	15.24	9.64	6.59	1.62	1.00				
MSDP-KG	2.86	2.72	2.49	67.41	12.41	10.78	9.40	1.37	1.00	3.91	3.29	3.72	69.75	15.63	7.81	6.81	1.65	1.00				
IMKDG+BioGPT	2.83	2.91	2.74	62.83	20.59	9.27	7.31	1.68	1.00	3.15	2.84	2.98	63.21	18.27	10.86	7.66	1.39	1.00				
IMKDG+MedAlpaca	3.53	3.47	3.32	73.59	13.36	7.71	5.34	1.78	1.00	3.41	3.14	3.22	79.97	9.10	7.92	3.01	1.78	1.00				
IMKDG + ChatDoctor	3.67	3.59	3.45	74.84	11.28	8.13	5.75	1.65	1.00	3.65	3.33	3.54	74.75	11.39	8.12	5.74	1.72	1.00				
IMKDG + ChatGPT	3.96	3.82	3.67	76.25	9.87	8.14	5.74	1.42	1.00	3.87	3.67	3.72	76.63	8.36	8.60	6.41	1.53	1.00				
IMKDG+PMC-LLaMA	4.02	3.63	3.52	78.99	8.00	6.00	7.01	1.82	1.00	4.32	3.96	4.15	82.84	8.76	4.43	<b>3.97</b>	<b>1.90</b>	1.00				
IMKDG + MedPaLM 2	4.29	4.18	4.01	82.33	6.28	4.81	6.58	1.38	1.00	4.56	4.21	4.33	84.90	5.67	4.01	5.42	1.65	1.00				
IMKDG + GPT-4	<b>4.63</b>	<b>4.42</b>	<b>4.29</b>	<b>85.15</b>	<b>5.16</b>	<b>4.02</b>	5.67	<b>1.75</b>	<b>1.00</b>	<b>4.79</b>	<b>4.68</b>	<b>4.55</b>	<b>88.07</b>	<b>3.67</b>	<b>3.44</b>	4.82	1.28	<b>1.00</b>				

Table 3: Results of human evaluation. The bolded values represent the best scores. The scales for the metrics are indicated in parentheses.

Models	Automatic Evaluation											Manual Evaluation								
	PPL	B-4	MET	R-L	BS-F1	EA	VE	GM	E-F1	FEQA	QE	F	A	CR	Faithful	Hallucinated			H	S
	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	(1-5)	E	I	B	(0-2)	(0-1)
<b>MedDialog-EN</b>																				
IMKDG	7.23	0.085	0.679	0.597	86.78%	0.978	0.447	0.763	85.00%	50.29	59.94	4.63	4.42	4.29	85.15	5.16	4.02	5.67	1.75	1.00
IMKDG w/o $\mathcal{K}$	12.71	0.082	0.594	0.511	82.11%	0.923	0.423	0.751	79.60%	42.68	52.89	4.32	4.17	4.02	81.20	7.32	5.16	6.32	1.66	1.00
IMKDG w/o $\mathcal{E}$	10.53	0.088	0.605	0.568	83.19%	0.956	0.431	0.759	81.50%	46.05	54.72	4.44	4.26	4.11	83.65	6.13	5.07	5.15	1.71	1.00
IMKDG w/o ( $\mathcal{K}+\mathcal{E}$ )	16.3	0.061	0.527	0.425	77.50%	0.875	0.386	0.662	72.60%	38.42	37.94	3.16	3.98	3.84	76.43	8.79	7.21	7.57	1.14	1.00
<b>CDialog</b>																				
IMKDG	1.13	0.094	0.448	0.478	85.08%	0.957	0.485	0.823	85.45%	51.29	61.94	4.79	4.68	4.55	88.07	3.67	3.44	4.82	1.28	1.00
IMKDG w/o $\mathcal{K}$	8.87	0.081	0.645	0.579	78.19%	0.946	0.468	0.789	81.30%	45.85	56.78	4.47	4.35	4.22	83.67	5.27	4.54	6.52	1.19	1.00
IMKDG w/o $\mathcal{E}$	7.74	0.087	0.652	0.594	81.12%	0.952	0.475	0.801	82.10%	47.02	58.13	4.59	4.47	4.31	85.12	4.88	4.61	5.39	1.25	1.00
IMKDG w/o ( $\mathcal{K}+\mathcal{E}$ )	10.43	0.083	0.597	0.416	72.01%	0.854	0.395	0.667	69.98%	35.12	44.78	3.31	3.13	3.09	80.89	7.02	6.02	6.07	1.09	1.00

Table 4: Ablation study for knowledge ( $\mathcal{K}$ ) and in-context exemplars ( $\mathcal{E}$ ) for the IMKDG + GPT-4.

ating helpful and safe responses, as indicated by H and S scores. Fleiss’ kappa (Fleiss, 1971) is used to calculate agreement among evaluators. For general generation evaluation, we observe that all kappa values are  $> 0.82$ , indicating substantial agreement among the evaluators. Similarly, for faithful and helpful generation evaluation, we note all kappa values surpassing 0.76, suggesting substantial agreement among evaluators. For safe generation evaluation, kappa is 1, indicating perfect agreement among evaluators.

**Ablation Analysis.** In Table 4, we present an ablation study of IMKDG with GPT-4 to examine the impact of medical knowledge ( $\mathcal{K}$ ) and in-context exemplars ( $\mathcal{E}$ ) for faithful MDG. The removal of in-context exemplars, medical knowledge, or both results in a noticeable decrease in E-F1, FEQA, and QE. This indicates that removing in-context exemplars hampers LLMs’ contextual understanding, potentially resulting in irrelevant or incorrect responses. Moreover, omitting medical knowledge results in significant performance drops, notably affecting FEQA scores (-15.13% for MedDialog-EN and -10.61% for CDialog). Similar trends are observed in E-F1 and QE metrics, highlighting the model’s tendency to hallucinate without adequate knowledge. The largest performance drop occurs when both knowledge and in-context exemplars are ablated. Human evaluation also shows reduced performance under these conditions, underscoring the importance of knowledge and in-context exemplars for enhancing faithful MDG.

**Qualitative and Error Analysis.** We conduct a qualitative and error analysis to further examine the quality and limitations of the IMKDG framework. Representative examples are presented in Table 5. Overall, the responses generated by IMKDG demonstrate strong alignment with the provided medical knowledge and dialogue context. IMKDG+GPT-4 consistently generates more accurate, knowledge-grounded responses than IMKDG+MedPaLM 2, corroborating both automatic and human evaluation results. For instance, in Table 5, for Sample (1), IMKDG+GPT-4 offers a detailed diagnosis and explicitly recommends antibiotics as the primary treatment for pneumonia by effectively leveraging knowledge triplets, such as [*pneumonia, may\_be\_treated\_by, antibiotics*] and [*pneumonia, associated\_with, pregnancy\_complications\_pneumonia*]. In contrast, IMKDG+MedPaLM 2 provides a more general overview, omitting critical treatment details. Similarly, for sample (2), proposed model correctly infers that handwashing is a necessary precaution for infectious diseases based on the triplet [*infectious\_disease, focus\_of, handwashing\_precautions*].

Despite these strengths, some failure cases persist [Sample (3)]. We observe both intrinsic and extrinsic hallucinations. In one instance, IMKDG+MedPaLM 2 and IMKDG+GPT-4 incorrectly advise the application of hot water instead of ice, contradicting the relevant knowledge triplet [*arm, procedure\_site\_of, massage\_of\_left\_arm\_with\_ice\_or\_coolant\_spray*]. In another case, the gener-

<p>(1) Context</p> <p>Gold Response</p> <p>Triplets</p> <p>IMKDG + MedPaLM 2</p> <p>IMKDG + GPT-4</p>	<p><b>Patient:</b> what do scabs in nose and mouth of an infant indicate? my lo is 1 year old and 3 days ago came out in some scabs around her nose and mouth she also has a blister which has peeled on her thumb. i worry she may have impetigo passed from my cousin who currently has it to! anyways i was just wondering if putting savlon cream on would be safe? and if it would help? thanks in advance!</p> <p><b>Doctor:</b> yes, savlon is a good disinfectant and it can be used to prevent the spread of infection from others to one's own skin subcutaneous tissues (area of skin under the thin brown skin that we see)</p> <p>[nose, contains, skin] [impetigo, contains, skin redness], [nose, contains, dry skin], [impetigo, isa, bacterial infection], [nose, contains, skin rashes], [impetigo affects, infants], [nose, contains, prednisone 20 mg], [impetigo, contains, skin dry], [nose, contains, breathing pain]</p> <p>It's understandable to be concerned about your little one's scabs. Based on your description, the scabs around her nose and mouth, along with the peeled blister on her thumb, could be a sign of <b>impetigo</b>. This is a common <b>skin infection in children, often caused by bacteria</b>. It's crucial to consult a doctor for proper diagnosis and treatment. Savlon cream is an antiseptic, and while it may help with minor cuts and abrasions, it's not a treatment for impetigo. A doctor can prescribe the appropriate medication, likely an antibiotic cream.</p> <p>It sounds like your little one may indeed have <b>impetigo</b>, especially considering the close contact with your cousin who currently has it. Impetigo is a common and contagious <b>skin infection</b> that often presents with red sores, scabs around the nose and mouth, and sometimes blistering as you described. However, it might be due to common <b>skin rashes</b>. Regarding the use of Savlon cream, it is best to consult with your pediatrician before applying any antiseptic creams like Savlon on an infant's sensitive skin.</p>
<p>(2) Context</p> <p>Gold Response</p> <p>Triplets</p> <p>IMKDG + MedPaLM 2</p> <p>IMKDG + GPT-4</p>	<p><b>Patient:</b> What extra additional precautions do they need to prevent catching novel coronavirus?</p> <p><b>Doctor:</b> COVID-19 is now the official name for the illness caused by the newly discovered coronavirus (CoRoNaVirus Infectious Disease - 2019). So far it is extremely rare in the US (2/12/20). Until and unless COVID-19 becomes common no special precautions are necessary.</p> <p><b>Patient:</b> For students in dorms living with students from many countries.</p> <p><b>Doctor:</b> In any dormitory or group living situation people with respiratory symptoms (colds, flu, etc.) should follow the respiratory secretion precautions, cover their coughs and wash hands frequently.</p> <p>[covid-19, caused_by, novel coronavirus], [infectious disease, focus_of, handwashing precautions], [infectious disease, focus_of, respiratory secretion precautions], [infectious disease, focus_of, standard precautions], [respiratory secretion precautions, include, covering coughs]</p> <p>Standard precautions like <b>handwashing and avoiding contact with respiratory secretions</b> are always a good idea. For students in dorms living with individuals from many countries, it is essential to take additional precautions to prevent catching COVID-19. This includes rigorously <b>following respiratory secretion precautions</b>, such as wearing masks, especially in common areas, and <b>practicing good hand hygiene by frequently washing hands with soap and water</b> for at least 20 seconds. Additionally, maintaining standard precautions such as avoiding close contact with sick individuals and regularly disinfecting frequently touched surfaces can further help reduce the risk of transmission.</p>
<p>(3) Context</p> <p>Gold Response</p> <p>Triplets</p> <p>IMKDG + MedPaLM 2</p> <p>IMKDG + GPT-4</p>	<p><b>Patient:</b> Last tuesday I had a flu shot and a pneumonia shot. I am 68 and male. I had a very sore arm and felt bad the next day in which I had blood drawn for a test. The results showed a high white cell blood count as well as high ANC and AMC. Could these high readings be because of the shots the day before?</p> <p><b>Doctor:</b> Hello, The pneumonia shot per se is not responsible for high WBC count, but as you are telling there are soreness and bad feeling at the injection site over arm, there are all possibilities that you have cellulitis or possible abscess formation at the injection site; which will recruit neutrophils at the site and cause secondary rise in WBC count.</p> <p><b>Patient:</b> Okay Doctor, So what would you recommend ?</p> <p><b>Doctor:</b> Primary relief with ice application and secondary management with systemic antibiotics and anti-inflammatory medicines with your doc. Hope I have answered your query. Let me know if I can assist you further.</p> <p>[pneumonia, related_to, inflammation], [abscess, use, inflammation], [arm, procedure_site_of, massage of left arm with ice or coolant spray]</p> <p>Okay, I recommend we take a closer look at the injection site. I'd like to examine it and possibly take a culture to determine if it's cellulitis or an abscess. We can then discuss treatment options based on the findings. In the meantime, I recommend applying a <b>warm compress to the area for 15-20 minutes, 3-4 times a day</b>. If you experience any increased redness, swelling, or pain, please don't hesitate to contact the office.</p> <p>I recommend visiting your healthcare provider to examine the injection site. You may need antibiotics if there's an infection. Keep the area clean by and <b>apply hot water</b> and monitor for further symptoms like increased redness or swelling.</p>

Table 5: Sample predictions from the IMKDG framework on (1) MedDialog-EN and (2), (3) CDIALOG.

ated response introduces extraneous content (e.g., “sinusitis”) not supported by the input knowledge or dialogue context. Occasionally, the model fails to fully integrate the provided knowledge triplets, resulting in generic or clinically less relevant responses. These observations highlight the need for more robust grounding and context comprehension mechanisms to minimize hallucination and improve clinical reliability.

## 6. Conclusion

In this work, we empirically assess medical and general LLMs for faithful response generation in medical dialogues, aiming to mitigate hallucination. Our approach integrates external medical knowledge and employs ICL to enhance faithfulness in MDG tasks. We propose IMKDG (In-context Knowledge-grounded Medical Dialogue Genera-

tor), a plug-and-play framework that leverages structured medical knowledge and the ICL strategy. We introduce the Homo-Heterogeneous Exemplar Selection (H2ES) method to ensure relevant and diverse exemplar selection, thereby enhancing LLMs’ ICL performance. Automatic and human evaluations empirically show that IMKDG substantially reduces hallucinations and ensures faithful medical dialogue generation, while demonstrating efficacy and generalizability across distinct LLMs and datasets. For future work, we aim to further enhance explainability and transparency by exposing intermediate reasoning steps for each model decision. Moreover, we plan to incorporate multi-agent and collaborative schemes, in which specialized agents generate, verify, and cross-check responses to further reduce hallucinations and improve the factual reliability of generated medical dialogues.

## Ethics Statement

The significant implications of developing computational models for medical applications underscore the imperative need for meticulous consideration of ethical concerns. To ensure ethical compliance, this study has been evaluated and approved by our Institutional Review Board (IRB). All datasets used in this research are publicly accessible and sourced from public websites. For medical dialogue generation systems, safeguarding patient confidentiality is of utmost importance. These systems aim to provide support and guidance while upholding patient privacy. We strictly adhere to data usage policies and copyright regulations, maintaining the ethical integrity of our research. Despite using publicly available medical dialogue datasets, the evaluators pledged not to contact or deanonymize any users or to share the data with others. It is important to note that the dataset used in our study is intended solely for academic research. Furthermore, in practical applications, medical dialogue systems hold the potential for patient counseling and diagnostic data collection. In instances where minor errors occur during the dialogue process, medical professionals intervene to rectify them, ensuring the integrity of patient information remains intact.

## Limitations

In this study, we do not investigate the process of selecting relevant knowledge, which is also crucial for faithful response generation in dialogues. How to effectively integrate external knowledge into LLMs is a promising research area that has received limited attention. Due to limitations imposed by text window size, we are unable to provide results for additional examples. Due to resource constraints, we empirically set  $\lambda$ . A hyperparameter search for coefficients may further enhance the performance of H2ES method. Further, the sensitivity of LLMs to decoding parameters is a significant concern, as altering these parameters can result in models that produce correct examples while hallucinating wrong answers (Zhou et al., 2023). This issue highlights the need for more research to make LLMs more robust to these settings. In the present study, we refrain from conducting a parameter sensitivity analysis to determine optimal settings that can further reduce hallucination and improve the overall performance of the model. Additionally, we refrain from conducting analysis pertaining to the performance improvement of the proposed IMKDG framework with stronger LLMs compared to the smaller ones. These aspects remain unexplored and offer avenues for future investigation.

## Acknowledgements

Priyanshu Priya acknowledges the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship implemented by the Department of Science and Technology, Ministry of Science and Technology, Government of India for financial support.

## 7. Bibliographical References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.
- Explosion AI. 2025. Pipeline · spacy universe. <https://spacy.io/universe/category/pipeline>. Accessed: 2025-10-25.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.
- BSI. 1973b. *Natural fibre twines*. BS 2570, British Standards Institution, London. 3rd. edn.

- A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.
- J.L. Chercœur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.
- N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and play language models: A simple approach to controlled text generation](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2023. Plugmed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5050–5066.
- Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.
- Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.
- Global Market Estimates. 2020. Global online doctor consultation market. <https://www.globalmarketestimates.com/market-report/global-online-doctor-consultation-market-2172>. Accessed: 2025-10-25.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- YanJun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.
- Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.
- Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.
- Abid Hossain, Priyanshu Priya, Armita Mani Tripathi, Pradeepika Verma, and Asif Ekbal. 2025. Mender: Multi-hop commonsense and domain-specific cot reasoning for knowledge-grounded empathetic counseling of crime victims. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 501–516.
- Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.
- Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. Rho: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522.
- Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*.
- Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2022. Knowledge-consistent dialogue generation with knowledge graphs. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- 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.
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. 2022a. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737.
- Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022b. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023. Pvgru: Generating diverse and relevant dialogue responses via pseudo-variational mechanism. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3295–3310.
- Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022c. Multi-stage prompting for knowledgeable dialogue generation. *arXiv preprint arXiv:2203.08745*.
- Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022d. Multi-stage prompting for knowledgeable dialogue generation.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).
- Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023a. e-therapist: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967.
- Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023b. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14408–14416.
- Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023c. Pal to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271.
- OpenAI. 2024. *Chatgpt*.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.
- 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.
- Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. 2023. Partner: A persuasive mental health and legal counselling dialogue system for women and children crime victims. In *IJCAI*, pages 6183–6191.

- Priyanshu Priya, Armita Mani Tripathi, Deeksha Varshney, Mauajama Firdaus, and Asif Ekbal. 2025. Empower-kare: Deep prompt learning for knowledge-aware response generation in clinical counseling and legal support conversations. *IEEE Transactions on Artificial Intelligence*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.
- Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.
- João Sá, Vanessa Queiroz Marinho, Ana Rita Magalhães, Tiago Lacerda, and Diogo Goncalves. 2022. Diversity vs relevance: A practical multi-objective study in luxury fashion recommendations. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*, pages 2405–2409.
- Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.
- Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, and Shaoting Zhang. 2024. Medical dialogue: A survey of categories, methods, evaluation and challenges. *arXiv preprint arXiv:2405.10630*.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.
- 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. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).
- S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Deeksha Varshney, Aizan Zafar, Niranshu Behra, and Asif Ekbal. 2022a. *CDialog: A multi-turn covid-19 conversation dataset for entity-aware dialog generation*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11373–11385, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behra, and Asif Ekbal. 2022b. *Cdialog: A multi-turn covid-19 conversation dataset for entity-aware dialog generation*. *arXiv preprint arXiv:2212.06049*.
- Bernard L Welch. 1947. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. *Pmc-llama: Towards building open-source language models for medicine*. *arXiv preprint arXiv:2305.10415*, 6.
- Jiageng Wu, Xian Wu, Zhaopeng Qiu, Minghui Li, Shixu Lin, Yingying Zhang, Yefeng Zheng, Changzheng Yuan, and Jie Yang. 2024. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. *Journal of the American Medical Informatics Association*, page ocae079.

- Zequi Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14085–14093.
- Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107.
- Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.
- 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. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.