Listening to Patients: A Framework of Detecting and Mitigating Patient Misreport for Medical Dialogue Generation

Anonymous ACL submission

Abstract

Medical Dialogue Systems aim to provide automated healthcare support through patient-agent conversations. Previous efforts typically regard patients as ideal users—one who accurately and consistently reports their health conditions. However, in reality, patients often misreport their symptoms, leading to discrepancies between their reports and actual health conditions. Overlooking patient misreport will affect the quality of healthcare consultations provided by MDS. To address this issue, we argue that MDS should "listen to patients" and tackle two key challenges: how to detect and mitigate patient misreport effectively. In this work, we propose **PaMis**, a framework of detecting and mitigating Patient Misreport for medical dialogue generation. PaMis first constructs dialogue entity graphs, then detects patient misreport based on graph entropy, and mitigates patient misreport by formulating clarifying questions. Experiments indicate that PaMis effectively enhances medical response generation, enabling models like GPT-4 to detect and mitigate patient misreports, and provide high-quality healthcare consultations.

1 Introduction

Medical Dialogue Systems (MDSs) aim to provide automated healthcare support through natural language interactions between patients and agents (Li et al., 2021; Liu et al., 2022b; Xu et al., 2024). The patient describes symptoms or health concerns, while the agent processes the patient self-report and responds with appropriate medical guidance and follow-up questions, mimicking the strategies employed by real doctors. Taking Figure 1 as an example, when a patient reports that he/she *feels dizzy (P1)*, the agent will inquire about possessing more related symptoms, such as *vomiting (A1)* and *cold (A2)*. In order to better provide medical support, previous works on MDSs has devoted significant effort to leveraging advanced frameworks

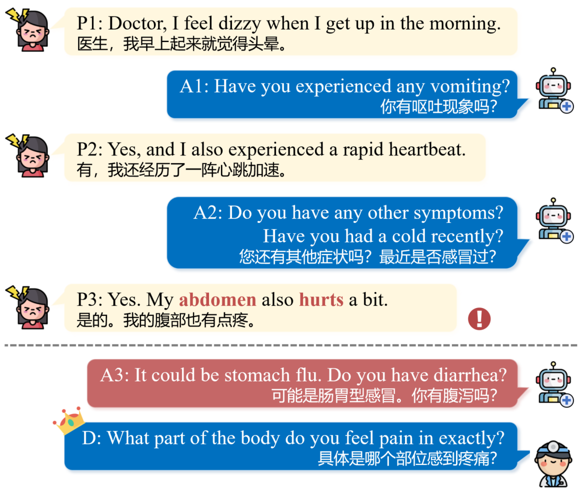


Figure 1: Example of patient misreport in patient (P)-agent (A) conversations and a response generated by the real experienced doctor (D). When the patient with myocardial infarction misreports angina as abdominal pain, the doctor remains vigilant and asks more detailed questions to discern the patient's actual symptoms. However, the agent can easily be influenced by the patient's misreport and arbitrarily shift the focus to stomach flu.

(e.g., reinforcement learning (Wei et al., 2018) and graph-based structures (Lin et al., 2019)) and integrating external medical knowledge (Xu et al., 2023).

Despite extensive research, current efforts still operate under the assumption of an ideal patient—one who accurately and consistently reports their health conditions. In reality, patients often misreport their symptoms due to cognitive limitations or emotional factors, etc., leading to discrepancies between their reports and actual health conditions (Berkman et al., 2011; Prior et al., 2011). Some research indicates that patients misreport their symptoms in approximately 15–20% of cases (Fleischer et al., 2015; Merckelbach et al., 2019). Meanwhile, we observed that patient misreports occurred in 16.9% of the dialogues within the public corpus (Liu et al., 2022c). Still taking Figure 1 as an example, a patient experiencing a myocardial infarction might inaccurately describe angina as *abdomen pain (P3)*. When by default the patient is able to accurately report his or her symptoms, the agent will arbitrarily infer that the patient may have a *stomach flu (A3)*. This would potentially delay appropriate treatment and adversely impact the patient's health. In contrast, an experienced doctor would remain vigilant and ask *more details about the location of pain (D)* to discern the patient's actual symptoms.

Therefore, it is essential to move beyond the assumption of idealized patients, enabling dialogue systems to detect potential misreports in patient narratives and ask clarifying questions, much like an experienced doctor, to discern the actual health conditions. When confronting the issue of patient misreport, we propose that an effective MDS should tackle two key challenges:

- ***Misreport Detection***: Due to the complexity and subtlety of patient narratives, misreport often emerges as implicit contradictions in patients' self-reported information. Detecting patient misreports may span multiple dialogue rounds and vary across individuals, as well as requiring deep medical domain knowledge.
- ***Misreport Mitigation***: If detecting a misreport, the agent needs to generate targeted clarifying questions to help the patient calibrate the self-report to ensure high-quality healthcare. The generated questions need to integrate the detected misreport with medical knowledge to not only mitigate misreporting, but also maintain natural dialogue flow.

Inspired by the aforementioned analyses, we propose **PaMis**, a framework of detecting and mitigating Patient Misreport for medical dialogue generation. PaMis utilizes dialogue context and external medical knowledge to (1) construct entity graph, (2) detect patient misreport based on dialogue entity graph and (3) mitigate patient misreport by formulating clarifying questions. Specifically, the misreport detection module calculate the structural entropy of the dialogue entity graph for detection, while the misreport mitigation module generates controlled clarifying questions based on the detected misreport information. Experimental results on two medical dialogue datasets, MedDG and KaMed, demonstrate PaMis's superior performance in medical response generation. Furthermore, when integrated with state-of-the-art language models like GPT-4, PaMis significantly enhances their ability to detect and mitigate patient misreports. On the strength of the encouraging performance, we are confident that PaMis can effectively contribute to the MDS in providing high-quality healthcare consultations. In conclusion, the key contributions of this research are outlined below.

- We call attention to the underexplored phenomenon of patient misreport that occurs in patient-agent conversations.
- We propose PaMis, a framework for detecting patient misreport based on graph entropy, and mitigating patient misreport by formulating clarifying questions.
- Experiments indicate that PaMis effectively enhances medical response generation, enabling models like GPT-4 to detect and mitigate patient misreports, and provide high-quality healthcare consultations.

2 Related Work

2.1 Medical Dialogue Systems

Medical dialogue systems aim to collect symptoms and automate diagnosis by obtaining information about patients' health conditions through conversation. **(1) Symptom Collection:** Given the critical role of entities in medical dialogues (Liu et al., 2022c), previous studies have developed entity-aware models for symptom collection. Lin et al. (2019) utilized a symptom graph with global attention to identify symptoms. Xu et al. (2023) introduced a framework that combines medical entity flows with dialogue action flows. **(2) Automatic Diagnosis:** Early research focused on developing agents that mimic physicians' diagnostic dialogue behavior. Wei et al. (2018) applied DQN to refine strategies for extracting symptoms from dialogues, aiding in diagnosis. Xu et al. (2019b) incorporated knowledge graphs to optimize end-to-end automated diagnosis. **(3) MDS Challenges:** Previous studies have highlighted the effectiveness of guiding agents to emulate physician behavior (Li et al., 2021; Liu et al., 2022b) and have addressed challenges such as vague patient statements and limited medical data (Zhao et al., 2022; Xu et al., 2023; Tang et al., 2023; Lin et al., 2021; Hou et al., 2023). However, they have not adequately considered the issue of patient misreporting, and recent research continues to assume that patients will provide ac-

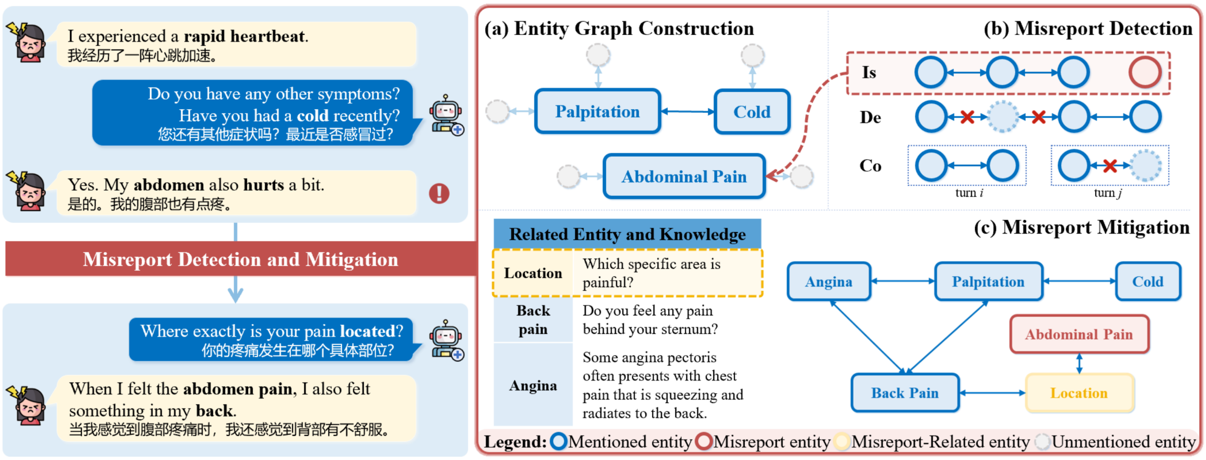


Figure 2: An illustration of PaMis, using the dialogue in Figure 1 as an example. PaMis first constructs the entity graph, and then detects and mitigates the patient misreport based on the entity graph.

curate answers based on correct facts (Li et al., 2024).

This study addresses patient misreporting by introducing a medical dialogue management framework that detects and mitigates inaccuracies.

2.2 Misreport in Task-oriented Dialogue

To address vague or incorrect information provided by users, task-oriented dialogue systems often employ proactive questioning to clarify issues. These systems construct clarifying questions to maximize the expected information gain based on the principle that "a good question is one whose expected answer is the most useful (Rao, 2017)." Specifically, when confronted with unclear user intent, ambiguous expressions, or statements containing errors, prior research suggests several approaches: Aliannejadi et al. (2019) retrieve related questions based on interaction history to identify user information needs in open-domain information-seeking tasks; Xu et al. (2019a) employ a "judge-then-generate" method in knowledge-based QA systems to resolve ambiguities in knowledge items; Feng et al. (2023) utilize clarifying questions to gather necessary user-specific information in task-oriented dialogues; Zhao and Dou (2024) address ambiguous or multifaceted user intents in web search through multi-turn questions. Additionally, Oshima et al. (2023) investigated challenges arising from human errors that lead to agent failures in goal-oriented visual QA tasks, emphasizing the importance of agents detecting and pointing out these errors.

In this paper, we propose a misreport detection mechanism for medical dialogue systems and utilize agent responses to disambiguate information and accurately capture the patient's health status.

3 Methodology

3.1 Overview

We formulate the doctor-patient conversation as $\mathcal{C} = \{(\mathcal{P}_i, \mathcal{D}_i)\}_{i=1}^{T}$, where $\mathcal{P}_i$ denotes the patient's statement and $\mathcal{D}_i$ represents the doctor's response. The primary objective of MDS is to develop a physician agent that generates an appropriate response $\mathcal{D}_i$ based on the dialogue history $\mathcal{C}_{1:i-1}$ and the patient's current statement $\mathcal{P}_i$.

The entities $\{e_i\}$ appearing in the conversation $\mathcal{C}$ can be organized into a graph G_e to represent the patient's health condition and the doctor's logic of inquiry. In addition to generating responses $\mathcal{D}_i$, a MDS that addresses patient misreports must detect misreported entities e_m and ask clarifying questions to reduce inaccuracies. The system leverages both the dialogue content and the entity graph to maximize the probability $P\left(\mathcal{D}_i \mid \mathcal{C}_{1:i-1}, \mathcal{P}_i, G_e, e_m\right)$.

3.2 Entity Graph Construction

We introduce a modeling approach that integrates a static knowledge graph G with a dynamic dialogue entity graph $G_e \subseteq G$ to improve detection effectiveness. The graph G is constructed from the corpus, with nodes representing entities and edge weights representing co-occurrence relationships. The likelihood of mentioning entity e_t at turn t is modeled as the conditional probability given the entities $\{e_1, e_2, \ldots, e_{t-1}\}$ in the dialogue history:

$$P\left(e_t \mid \mathcal{C}_{1:t-1}\right) \propto P\left(e_t \mid e_1, e_2, \ldots, e_{t-1}\right). \quad (1)$$

Thus, we use the co-occurrence frequency in the corpus as the directed[1] weight w_{ij} from e_i to e_j:

$$w_{ij} = \frac{\text{freq}\,(e_i, e_j)}{\text{freq}\,(e_i)} \propto P\,(e_j \mid e_i)\,. \qquad (2)$$

The edge weights in graph G serve as the basis for ranking relevant entities during subsequent retrieval for response generation.

The dialogue history references or denies medical entities, which are extracted through medical slot-filling (Hu et al., 2023). The extracted entities and relationships $r = (e_i, e_j, w_{ij})$ must align with established background knowledge G, rather than transient associations. Entities identified in each utterance form a separate dialogue entity graph G_e for that utterance. We track changes in these graphs to generate a sequence of dynamic graphs, $\{G_{e_1}, G_{e_2}, \ldots\}$. Misreport detection and classification are then performed using graph entropy theory.

3.3 Misreport Detection

Patient misreporting occurs when discrepancies arise between the health conditions described in the dialogue and the patient's actual health conditions. Health conditions often follow inherent medical co-occurrence relationships (Liu et al., 2022a; Bhoi et al., 2023), which can be disrupted by misreports, leading to anomalies in the dialogue's entity graph. We analyze misreport patterns using real-world data (Liu et al., 2022c; Li et al., 2021) and apply graph entropy to detect these misreports.

(1) Misreport Feature Classification

Patient misreports often lack clear linguistic indicators, but they manifest as anomalies on the entity graph due to inconsistencies with established medical knowledge. Based on observed patient behavior, we categorize misreports into three types.

- *Introducing isolated entities* (Figure 2(b)-(Is)). This type is typically initiated by patients because of anxiety or vague descriptions.
- *Denying crucial entities* (Figure 2(b)-(De)). This type involves deleting entities and transferring graphs to disconnect components, which occurs when patients deny entities in doctors' inquiries.
- *Presenting self-contradictions* (Figure 2(b)-(Co)). This type involves entities appearing and disappearing without disrupting the graph's structure, typically reflecting patients misreporting the duration of symptoms or medical history[2].

(2) Graph Entropy-based Detection

Graph entropy characterizes the structural information of graphs. Given that misreporting disrupts graph structure, we use graph entropy for detection. Specifically, to address the potential disconnected features of entity graphs, we apply one-dimensional structural entropy (Li and Pan, 2016) to detect misreporting phenomena.

The definition of graph entropy aligns with Shannon's information entropy (Shannon, 1953), aiming to represent structure information and complexity through the degree distribution of nodes. In the following formula, d_i denotes the degree of node i, and $\text{Vol}(G)$ is the degree sum of G:

$$H(G) = -\sum_{i=1}^{n} \frac{d_i}{\text{Vol}(G)} \log_2 \frac{d_i}{\text{Vol}(G)}. \qquad (3)$$

For the entity graph G_e, the one-dimensional structural entropy is calculated as the weighted average of the entropy of each connected component, defined as follows:

$$H^1(G) = \frac{1}{\text{Vol}(G)} \sum_{j=1}^{L} \text{Vol}\,(G_j) \cdot H^1\,(G_j)\,, \qquad (4)$$

where G_j represents a connected subgraph of G, L represents the number of connected components, and the entropy is considered as 0 if has no edges.

An effective inquiry process should enhance the co-occurrence relationships between entities, thereby expanding information pathways and increasing graph entropy. As illustrated in Figure 2, misreports can disrupt this pattern. A single calculation of graph entropy can be used to detect and classify such misreports:

- *(Is)* occurs when the number of nodes increases while the graph entropy remains unchanged because an isolated node has zero entropy.
- *(De)* and *(Co)* reduce information pathways in the graph, leading to a decrease in entropy. When nodes are lost, the entropy lower bound for contradictions is higher than the upper bound for

[1]Based on the sequence of entity occurrences in the corpus, it reflects the logic of doctors' inquiries.

[2]Doctors may focus on recent symptoms for acute conditions or long-term medication history for chronic illnesses. Patients might provide contradictory information, such as first claiming "never had a stomach problem" and later mentioning "occasional stomach pain."

denials. This boundary helps detect and classify misreports, with the formulations provided below and detailed proofs available in Appendix A:

$$H(G_{Co}) \geq -\sum_{i=1}^{n} \frac{d_i - 1}{2(n-1)} \log_2 \frac{d_i - 1}{2(n-1)}, \quad (5)$$

$$H(G_{De}) \leq -\left(\sum_{i=1}^{n-2} F(d_i) + F(d_{n-1} - 1)\right), \quad (6)$$

$$F(d) = \frac{d}{\mathrm{Vol}(G) - 4} \log_2 \frac{d}{\mathrm{Vol}(G) - 4}. \quad (7)$$

In summary, the sequence of graph entropy values enables the straightforward detection and classification of misreports.

3.4 Misreport Mitigation

The mitigation module generates clarifying questions based on detected patient misreports to obtain accurate information about patients' health conditions and mitigate misreporting.

(1) Response Generation

Misreport mitigation can be perceived as an instance of response generation as it involves the agent asking questions. Therefore, we first introduce the standard process of PaMis in utilizing medical knowledge to generate responses. As illustrated in Figure 2, we establish a connection between the medical knowledge graph and real physician responses based on entities. Next, we identify the most valuable responses to serve as guiding information. A set of real responses $\{S\}$ containing the entity e can be extracted from the corpus. We calculate the cosine similarity of word vectors $\mathbf{v}$ for the sentences in $\{S\}$ and select the top-k sentences that exhibit the highest average similarity to others as the knowledge $\{S_e\}$related to entity e:

$$\text{avg\_sim}(e) = \frac{1}{|S|} \sum_{s_i, s_j \in S, i \neq j} \text{sim}(s_i, s_j), \quad (8)$$

$$\text{sim}(s_i, s_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \, \mathbf{v}_j\|}. \quad (9)$$

The agent leverages the dialogue context and the dialogue entity graph to retrieve knowledge pairs $\{(e, S_e)\}$ from the one-hop neighbors of existing nodes. The candidate ranking process uses the weights described in Section 3.2, where the edge weights to adjacent nodes serve as scores for neighboring entities. This step does not mean performing the entity prediction task in MDS but provides a preliminary enhancement to the proposed framework. This explicit approach gives the framework flexibility to incorporate generative models, including large language models.

Dataset	MedDG	KaMed
Dialogues	17,864	63,754
Avg. # of utterances	19.85	23.25
Avg. # of entities	12.11	14.94

Table 1: The statistics of datasets.

Overall, the aim of response generation is to maximize the conditional probability $P\left(\mathcal{D}_i \mid \mathcal{C}_{1:i-1}, \mathcal{P}_i, \{(e, S_e)\}\right)$.

(2) Clarifying Process

Based on the approach described above, the clarifying process is triggered when misreports are detected. This process involves re-retrieving entity-response pairs $\{(e, S_e)\}$ related to the misreported entity e_m as guiding information. The implications vary depending on the scenario:

- (Is) PaMis retrieves bridging nodes from the static graph G to complete the dialogue entity graph, aiming to generate the next inquiry that restores the graph to a connected state.
- (De) PaMis uses related knowledge of denied entities to generate clarifying questions, reconfirming the existence or absence of the relevant health condition.
- (Co) PaMis selects broader attribute entities (such as duration, location, etc.) to generate inquiries, obtaining supplementary information to assist in judgment.

In summary, related knowledge is integrated to augment clarifying question generations by retrieving information related to misreports.

4 Experiments

We conduct experiments focusing on three core research questions:

- **RQ1**: Does PaMis outperform current methods in terms of overall performance of medical response generation?
- **RQ2**: Does PaMis perform better in meeting the fundamental requirements of the medical dialogue system?
- **RQ3**: Can PaMis effectively reduce misreports in doctor-patient interactions?

Method	MedDG						KaMed					
	P	A	H	LQ	DV	Overall	P	A	H	LQ	DV	Overall
HuatuoGPT2	3.46 ±0.03	3.99 ±0.02	3.32 ±0.03	4.27 ±0.07	3.08 ±0.03	18.11 ±0.08	3.35 ±0.05	3.95 ±0.05	3.26 ±0.04	4.18 ±0.03	2.96 ±0.04	17.70 ±0.19
DISC-MedLLM	3.20 ±0.05	3.95 ±0.05	3.19 ±0.06	4.16 ±0.09	2.87 ±0.07	17.37 ±0.28	3.12 ±0.01	3.93 ±0.02	3.10 ±0.01	4.03 ±0.05	2.75 ±0.04	16.93 ±0.08
ChatGPT	3.70 ±0.05	4.15 ±0.03	3.61 ±0.05	4.23 ±0.01	3.43 ±0.04	19.13 ±0.17	3.72 ±0.03	4.18 ±0.04	3.61 ±0.05	4.19 ±0.00	3.44 ±0.04	19.14 ±0.14
ChatGPT (w/ Gold Know.)	3.90 ±0.02	4.23 ±0.02	3.75 ±0.06	4.28 ±0.02	3.70 ±0.04	19.87 ±0.12	3.89 ±0.06	4.23 ±0.07	3.77 ±0.06	4.24 ±0.03	3.65 ±0.05	19.78 ±0.27
GPT-4	3.93 ±0.03	4.21 ±0.01	3.82 ±0.01	4.36 ±0.03	3.61 ±0.02	19.93 ±0.04	3.98 ±0.03	4.27 ±0.03	3.89 ±0.05	4.37 ±0.02	3.65 ±0.04	20.16 ±0.14
GPT-4 (w/ Gold Know.)	3.96 ±0.02	4.23 ±0.02	3.87 ±0.03	**4.37** ±0.02	3.67 ±0.03	20.10 ±0.09	4.03 ±0.04	4.29 ±0.02	3.96 ±0.04	**4.38** ±0.02	3.72 ±0.05	20.38 ±0.14
PaMis	**4.00**† ±0.01	**4.26**† ±0.02	**3.93**† ±0.05	4.36 ±0.01	**3.78**† ±0.02	**20.33**† ±0.03	**4.05** ±0.02	**4.32** ±0.01	**4.00** ±0.02	4.37 ±0.03	**3.84**† ±0.01	**20.58**† ±0.06

Table 2: Evaluation results of responses generated from LLMs and PaMis. "Gold" indicates that the entities in the actual doctor's responses are included in the input. † denotes statistically significant differences ($p < 0.05$).

Method	B-1	B-4	R-1	R-2	D-1	D-2	E-F
MedDG							
GPT-2	35.27	19.16	28.74	13.61	0.87	11.20	16.14
VRBot	29.69	16.34	24.69	11.23	-	-	12.78
DFMED	42.56	22.53	29.31	14.21	0.85	8.28	22.66
EMULATION	42.35	22.76	**30.91**	**15.17**	-	-	24.03
PaMis	**44.28**	**24.88**	28.12	13.80	**1.23**	**11.58**	**25.13**
KaMed							
GPT-2	33.76	17.82	26.80	10.56	-	-	17.26
VRBot	30.04	16.36	18.71	7.28	**1.61**	9.30	12.08
DFMED	40.50	20.92	28.33	11.73	-	-	22.31
EMULATION	39.87	19.79	**28.54**	**12.33**	-	-	24.27
PaMis	**41.02**	**21.30**	28.42	12.01	1.30	**11.39**	**28.18**

Table 3: Evaluation results of responses generated from fine-tuned baselines and PaMis under the N-gram-based metrics. Baseline results are reported from Li et al. (2021) and Xu et al. (2024).

4.1 Settings

Datasets We conducted experiments using the MedDG dataset (Liu et al., 2022c) and the KaMed dataset (Li et al., 2021). Medical entities mentioned in the dialogues were annotated by domain experts. Detailed dataset statistics are provided in Table 1. To enrich the medical background knowledge, we integrated multiple knowledge graphs from the datasets, focusing on the co-occurrence relationships between medical entities.

Baselines We selected two categories of baselines: LLMs with demonstrated medical capabilities and fine-tuned models. Detailed information about the baselines is provided in Appendix B.

(1) *LLMs*: DISC-MedLLM (Bao et al., 2023), HuatuoGPT2 (Chen et al., 2024), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI, 2024).

(2) *Fine-tuned models*: GPT-2 (Radford et al., 2019), VRBot (Li et al., 2021), DFMED (Xu et al., 2023), and EMULATION (Xu et al., 2024).

Metrics We employed three categories of evaluation metrics:

(1) *LLM-based*: We follow previous studies (Bao et al., 2023; Xu et al., 2024) that employ LLMs (e.g., GPT-4) as evaluators to assess MDS on the dimensions of *Proactivity, Accuracy, Helpfulness,* and *Linguistic Quality*. Additionally, we introduced a dimension called *Diagnostic Vigilance*[3] to measure the model's ability to detect and respond to misreporting phenomena.

(2) *N-gram-based*: We follow prior works (Liu et al., 2022c; Xu et al., 2023) that utilize BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Distinct (Li et al., 2016), and Entity-F1 (Liu et al., 2022c) for the response generation task.

(3) *Misreport-aware metrics*: To evaluate the effectiveness of misreport mitigation, we introduce two metrics: ΔGE, which quantifies the average change in graph entropy before and after interactions, and *MR*, the mitigation rate of misreports.

Implementation Details Building on previous studies (Chen et al., 2023; Bao et al., 2023), we construct prompts for generating responses and evaluations in comparison with LLMs. To compare with fine-tuned models, we employ a backbone model[4] of similar scale to the baselines. The code will be released upon acceptance of this paper. De-

[3]DV: The doctor skillfully identifies inconsistencies or errors in the patient's statement and asks clarifying questions to ensure an accurate diagnosis.

[4]huggingface.co/fnlp/bart-base-chinese

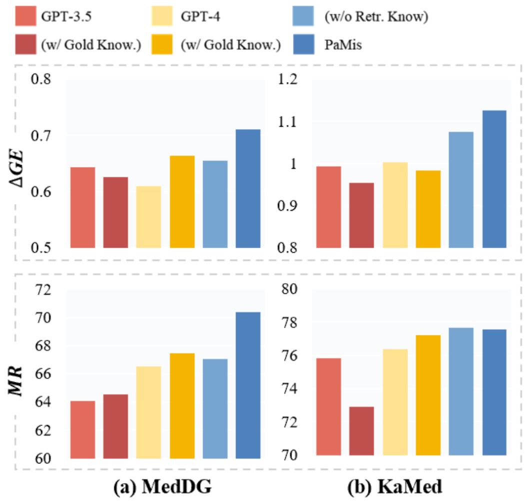


Figure 3: Evaluation results of interactive experiment under two misreport-aware metrics: ΔGE and MR.

tails of the prompts and additional implementation information are provided in Appendix C and D.

4.2 Results and Observations

Overall Performance We conducted experiments on the MedDG and KaMed datasets with the response generation task, which involves the agent acting as a doctor and responding to the patient. Results are presented in Tables 2 and 3.

For **RQ1**, Table 2 demonstrates that while PaMis is specifically designed to address patient misreports, it outperforms advanced LLMs on general metrics. We provided a robust baseline setting (i.e., w/ Gold Know.) for LLMs with medical capabilities. Nevertheless, introducing the misreporting mechanism and related entities in the input content via PaMis significantly improved response quality. Given that GPT-4 served as the generation model, this suggests that the PaMis management framework can enhance even highly advanced methods. Furthermore, the improvements are concentrated in areas beyond linguistic quality, suggesting that the enhancements arise from medical-related capabilities rather than language tricks.

For **RQ2**, The n-gram-based results presented in Table 3 illustrate that the proposed misreport mitigation module can remain competitive with the state-of-the-art approach in the response generation task of medical dialogue systems. Specifically, when relevant knowledge from the entity graph is introduced into the input of the generation model, it tends to generate inquiries that comprehensively incorporate related entities. This approach improves the E-F score but reduces the ROUGE score compared to ground-truth responses, as some doctors

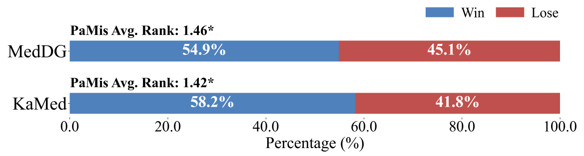


Figure 4: The human evaluation results of PaMis vs. GPT-4 (w/ Gold Know.) on two datasets.

Method	B-1	B-4	R-1	R-2	D-1	D-2	E-F
			MedDG				
PaMis	**44.28**	**24.88**	**28.12**	**13.80**	1.23	11.58	**25.13**
w/o Detection	37.06	20.28	27.05	12.55	**1.36**	**13.28**	24.02
w/o Mitigation	42.17	23.06	28.05	13.55	1.26	12.34	24.61
w/o D&M	35.83	19.57	26.34	11.83	1.29	12.48	24.09
			KaMed				
PaMis	**41.02**	**21.30**	**28.42**	12.01	1.30	11.39	**28.18**
w/o Detection	35.34	18.36	27.22	11.95	**1.34**	**12.70**	24.12
w/o Mitigation	39.01	19.74	28.35	**12.12**	1.31	12.44	24.89
w/o D&M	34.23	17.99	27.01	11.37	1.33	12.36	24.17

Table 4: Ablation study. "D" and "M" denote Detection and Mitigation modules respectively.

prefer shorter, more conversational questions.

The aforementioned results indicate that PaMis is able to simulate real doctors who possess professional skills to manage patient misreports while not compromising the fundamental abilities of MDS.

Interactive Experiment For **RQ3**, mitigating misreports requires continuous dialogue to guide the patient in confirming or modifying the mentioned information. We conducted interactive experiments utilizing a simulator-agent format. Implementation details are provided in Appendix D.

The results of the interactive experiment are presented in Figure 3. It demonstrates that in both quantitative and qualitative analyses, the responses generated by models guided by PaMis exhibit the ability to mitigate misreporting. This is reflected in an increase in the entropy of the entity graph, representing the completion of co-occurrence relationships, which leads to a more reasonable collection of symptoms. Notably, after removing the knowledge related to misreported entities retrieved by PaMis, the results on MedDG perform worse than the baseline, suggesting that the naive retrieval method described in Section 3.4 is effective for mitigating misreporting. In contrast, the MR on KaMed slightly increased, possibly due to the longer average dialogue length in KaMed, which provides sufficient entities to be confirmed, thereby reducing the need for redundant retrieval.

Experimental results above indicate that PaMis has the potential to serve as an aid tool to alleviate the burden of doctor inquiries and mitigate subsequent risks of diagnostic mistakes.

Context	**P**: After drinking a bit of liquor, I woke up in the middle of the night to use the restroom and noticed bleeding. The blood was bright red, and my stomach hurt a little. **D**: Hello, have you had any gastrointestinal issues in the past? **Do you have hemorrhoids?** How many times have you had bloody stools? Have you experienced abdominal pain, nausea, vomiting, fever, or any other discomfort? **P**: I have a slight **abdominal pain and a burning sensation**. I've **had discomfort from hemorrhoids before** but never treated it. I only had bloody stool once, and I haven't had a bowel movement since. No other discomfort. **D**: Could you describe the location of your pain? If your gastrointestinal health has been good in the past, the most likely cause is an **acute hemorrhoid flare-up**, followed by **acute damage to the intestinal mucosa**. **P**: **The burning sensation is in the stomach area**, but there's almost no pain.
Gold	**D**: Alright, it's quite normal to experience a burning sensation in the stomach after drinking. How do you feel compared to yesterday?
Misreported	Heartburn: Isolated
GPT-4	**D**: Your condition could be caused by acute hemorrhoids or mild gastritis. I suggest monitoring it for now, drinking plenty of water, and avoiding spicy foods. If the bleeding persists or the symptoms worsen, you may want to consider seeking medical attention.
PaMis	**D**: Based on your description, the symptoms may be caused by gastritis or hemorrhoids. I recommend that you seek medical attention as soon as possible to determine the appropriate treatment plan. Have you experienced acid reflux recently?

Table 5: Case study on the misreported entities and the responses by LLMs.

Human evaluation We selected GPT-4 (w/ Gold Know.) as a baseline for comparison with the proposed method and conducted human evaluations using sample-wise comparisons. We employed three doctors to assess the overall quality of the responses, with prior knowledge of the patient misreporting phenomenon. As shown in Figure 4, PaMis significantly outperforms GPT-4 in ranking and achieves a higher sample win rate.

4.3 Ablation Study

We conducted ablation experiments on the misreport-related modules to analyze their effects on response generation, as shown in Table 4. It observed a decrease in E-F scores after removing the misreport detection module and employing a naive prediction method based on entity co-occurrence relationships (i.e., "w/o Detection"). Furthermore, after removing the detection module, the Distinct-1/2 scores of the generated results on both datasets increased. This may be because the generated content was not constrained by specific entities. Removing the misreport mitigation module (i.e., "w/o Mitigation") resulted in fluctuations in BLEU and ROUGE scores. An increase in R-2 can be observed on the KaMed dataset, which may be because the model no longer receives an excessive number of entities for constructing clarifying questions, resulting in outputs that more closely resemble the conversational style of doctors. Further ablation results are provided in Appendix E.

4.4 Case Study

As shown in Table 5, when the patient mentions "burning in the stomach", given that the patient's description may not be accurate, it could indicate either gastric mucosal damage or heartburn. Since there is no additional information supporting the heartburn symptom, it is necessary to ask the patient again about the presence of acid reflux to determine whether he is experiencing esophageal damage. Existing models often ignore the potential inaccuracy, failing to confirm the patient's actual condition. PaMis retrieves related entities [Heartburn, Hemorrhoids, Gastritis] for response and probes the intermediate node "reflux", which could link heartburn to gastritis.

5 Conclusion

This paper focuses on the differences between real patients and the typically assumed ideal users in the field of medical dialogue systems. We defined the phenomenon that the content mentioned by the patient does not align with known health conditions as *Patient Misreports* and we highlighted the importance of addressing this issue. We then proposed PaMis to detect and mitigate patient misreports. Experimental results indicate the high effectiveness of PaMis in response generation. Interactive experiments further show its effectiveness in mitigating misreports. We believe the proposed approach can serve as a diagnostic aid tool to alleviate the burden on doctors during consultations.

Limitations

Although experimental results demonstrate that our method can enhance various generative models, including large language models, it cannot function independently as a standalone medical dialogue system. The reliability of this method derives from analyzing co-occurrence relationships among entities within authentic corpora. Although suspicion detection and questioning strategies help mitigate misreporting and prevent arbitrary decisions, they do not account for extreme cases absent from the corpus. Given the critical importance of reliability in the medical field, this method should be used solely as an auxiliary tool to support doctors during consultations. To prevent potential harm to patients, it should not be employed as a direct diagnostic tool.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of SIGIR*, page 475–484.

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*.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Suman Bhoi, Mong Li Lee, Wynne Hsu, and Ngiap Chuan Tan. 2023. Refine: A fine-grained medication recommendation system using deep learning and personalized drug interaction modeling. In *Proceedings of NeurIPS*, volume 36, pages 24013–24024.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Song Dingjie, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *Proceedings of the 1st Conference on Language Modeling*.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Yue Feng, Hossein A Rahmani, Aldo Lipani, and Emine Yilmaz. 2023. Towards asking clarification questions for information seeking on task-oriented dialogues. *arXiv preprint arXiv:2305.13690*.

Avi Fleischer, Alan D Mead, and Jialin Huang. 2015. Inattentive responding in mturk and other online samples. *Industrial and Organizational Psychology*, 8(2):196–202.

Zhenyu Hou, Yukuo Cen, Ziding Liu, Dongxue Wu, Baoyan Wang, Xuanhe Li, Lei Hong, and Jie Tang. 2023. Mtdiag: An effective multi-task framework for automatic diagnosis. In *Proceedings of AAAI*, pages 14241–14248.

Zefa Hu, Xiuyi Chen, Haoran Wu, Minglun Han, Ziyi Ni, Jing Shi, Shuang Xu, and Bo Xu. 2023. Matching-based term semantics pre-training for spoken patient query understanding. In *Proceedings of ICASSP*, pages 1–5.

Angsheng Li and Yicheng Pan. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339.

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 SIGIR*, page 544–554.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *arXiv preprint arXiv:2406.00922*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of AAAI*, pages 13362–13370.

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 EMNLP-IJCNLP*, pages 5033–5042.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. 2022a. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Proceedings of NeurIPS*, volume 35, pages 18864–18877.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022b. "my nose is running." "are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. In *Proceedings of IJCAI*, pages 4266–4272.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022c. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Proceedings of the Natural Language Processing and Chinese Computing*, pages 447–459.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Harald Merckelbach, Brechje Dandachi-FitzGerald, Daniel van Helvoort, Marko Jelicic, and Henry Otgaar. 2019. When patients overreport symptoms: More than just malingering. *Current Directions in Psychological Science*, 28(3):321–326.

OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ryosuke Oshima, Seitaro Shinagawa, Hideki Tsunashima, Qi Feng, and Shigeo Morishima. 2023. Pointing out human answer mistakes in a goal-oriented visual dialogue. In *Proceedings of ICCV Workshops*, pages 4663–4668.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Lindsay Prior, Meirion R Evans, and Hayley Prout. 2011. Talking about colds and flu: the lay diagnosis of two common illnesses among older british people. *Social Science & Medicine*, 73(6):922–928.

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

Sudha Rao. 2017. Are you asking the right questions? teaching machines to ask clarification questions. In *Proceedings of ACL, Student Research Workshop*, pages 30–35.

C. Shannon. 1953. The lattice theory of information. *Transactions of the IRE Professional Group on Information Theory*, 1(1):105–107.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Terminology-aware medical dialogue generation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Zhenduo Wang, Zhichao Xu, Qingyao Ai, and Vivek Srikumar. 2024. An in-depth investigation of user response simulation for conversational search. *arXiv preprint arXiv:2304.07944*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of ACL*, pages 201–207.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019a. Asking clarification questions in knowledge-based question answering. In *Proceedings of EMNLP-IJCNLP*, pages 1618–1629.

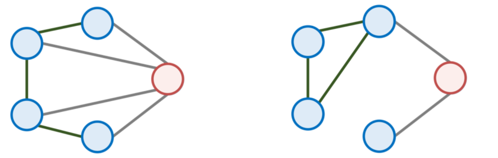
Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of ACL*, pages 6796–6814.

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

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

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 SIGKDD*, page 4763–4771.

Ziliang Zhao and Zhicheng Dou. 2024. Generating multi-turn clarification for web information seeking. In *Proceedings of WWW*, page 1539–1548.



(a) The worst scenario of self-contradiction (b) The best scenario of denial

Figure 5: Different scenarios after losing a node (using the example of 4 remaining nodes).

A Supplementary derivation of graph entropy

If there are n nodes remaining in the entity graph after a misreport, a self-contradiction involves at least $n-1$ edges connecting these nodes. As illustrated in Figure 5, in the worst-case scenario, the missing node was connected to every node in the original graph. Therefore, the minimum graph entropy is given by:

$$-\sum_{i=1}^{n} \frac{d_i - 1}{2(n-1)} \log_2 \frac{d_i - 1}{2(n-1)}. \qquad (10)$$

In the case of a denial, the best-case scenario occurs when the missing node in the original graph has only two edges. After the denial, the result is a connected graph with $n-1$ nodes and one isolated node, where the degree of the connected graph is $\mathrm{Vol}(G) - 4 \geq 2(n-1)$. Thus, the upper bound of entropy is:

$$\begin{aligned} &-\left(\textstyle\sum_{i=1}^{n-2} F(d_i) + F(d_{n-1} - 1)\right), \\ &F(d) = \tfrac{d}{\mathrm{Vol}(G)-4} \log_2 \tfrac{d}{\mathrm{Vol}(G)-4}. \end{aligned} \qquad (11)$$

Subtracting them yields a new equation that is evident when $n = 2$. For $n \geq 2$, the left-hand side is a monotonically increasing function with respect to n, while the right-hand side remains constant, indicating that if the equation holds for $n = 2$, it will also hold for any $n \geq 2$. Consequently, the lower bound for contradictions surpasses the upper bound for denials:

$$\begin{aligned} -\textstyle\sum_{i=1}^{n} \tfrac{d_i-1}{2(n-1)} \log_2 \tfrac{d_i-1}{2(n-1)} + \textstyle\sum_{i=1}^{n-2} F(d_i) \\ > -F(d_{n-1} - 1). \end{aligned} \qquad (12)$$

B Baselines

LLMs: (1) **DISC-MedLLM** (Bao et al., 2023), a trustworthy medical LLM adapted for multi-turn dialogues. (2) **HuatuoGPT2** (Chen et al., 2024), an advanced medical LLM trained on Chinese medical corpus.

Fine-tuned models: (1) **VRBot** (Li et al., 2021), a medical dialogue generation model based on patient entity tracking and doctor entity learning; and (2) **DFMED** (Xu et al., 2023), this framework performs the generation by fine-tuning a backbone model with dual-flow modeling. (3) **EMULATION** (Xu et al., 2024), this framework relies on diagnostic reasoning analyses and aligns with clinician preferences through thought process modeling.

C Complementary Implementation Details

For comparisons with fine-tuned baselines on the response generation task, we follow the settings of previous works, allowing any doctor's response to serve as the target sentence. We employ AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.02. Reproducibility is ensured by fixing random seeds. The proposed model is trained on an A100. Training is conducted with a learning rate of 1e-4 for 10 epochs, with no adjustment of other hyperparameters apart from setting the maximum generation length to 160 tokens.

D Interact Settings

Addressing concerns about the quality of dialogue simulators (Wang et al., 2024), we adopted the method proposed by Chen et al. (2023) to employ LLMs as simulated patients. The interactive experiment was initiated by selecting instances where misreports were detected within the dialogue. We then input the dialogue context into LLMs and requested that they respond as patients, guided by the generated doctor's inquiries. We employed the graph entropy method to assess whether misreports within the dialogue were mitigated. Notably, although the ChatGPT series models are not specialized in the medical domain, they surpass most models in medical-related evaluations[5]. We build on the prompt settings provided by Chen et al. (2023) to simulate patient-doctor interactions, as illustrated in Table 6. Simulated patients are utilized to assess whether doctor responses can help patients clarify their actual health conditions, thereby mitigating misreports. The prompts above were constructed with the participation of medical professionals and demonstrated their effectiveness as simulations of patients and doctors.

[5] paperswithcode.com/sota/question-answering-on-medqa-usmle

Doctor	You are a doctor conducting a conversational consultation with a patient. **1.** Take the information from the [Dialogue History] into account, which may include incomplete, inaccurate, or inconsistent details in the patient's statement. **2.** Reference the 'Potentially Inaccurate Entity' and 'Related Entity' in [Medical Knowledge] to provide accurate medical advice and help resolve uncoordinated issues. **3.** Respond in a way that is concise, approachable, and compassionate. Ask follow-up questions to gather more details and may also correct errors. **4.** Keep the conversation natural, focusing on one or two key points at a time to ensure the patient feels supported and informed. **5.** The response should be bite-sized and not give too much information at once, which is similar to what the doctor did in dialogue history.
Patient	You are a patient engaging in a conversational consultation with a doctor. **1.** Consider the [Dialogue History] and [Patient's Symptoms] for context to continue the conversation naturally. **2.** Respond briefly, using one to two short sentences. Describe your symptoms in a vague and colloquial manner, avoiding professional terminology. **3.** Express emotions like worry, anxiety, or hesitation as appropriate to reflect your life experiences. **4.** Keep your responses natural and reflective of how a patient might realistically communicate their concerns.

Table 6: The prompts for simulated patients and doctors.

	P	R	F1	$F1_A$	$F1_D$	$F1_E$	$F1_M$	$F1_S$
$\boldsymbol{Is}$	34.48	44.44	38.83	100.0	50.00	0.0	35.90	30.49
$\boldsymbol{Is}$ w/o Detection	30.32	41.05	34.88	100.0	25.00	50.00	31.97	29.64
$\boldsymbol{De}$	43.78	47.07	45.37	57.32	61.11	33.33	40.41	36.24
$\boldsymbol{De}$ w/o Detection	41.76	48.66	44.95	54.37	70.32	33.33	32.31	40.85
$\boldsymbol{Co}$	30.21	43.86	35.78	50.00	42.86	100.0	70.59	29.73
$\boldsymbol{Co}$ w/o Detection	16.99	37.27	23.35	45.45	68.18	0.0	50.00	16.00

Table 7: Ablation results for entity prediction on samples containing different types of misreports. A-S denote attribute, disease, examination, medicine, and symptom.

E Supplementary Ablation Study

As illustrated in Table 7, PaMis outperforms the overall average on samples containing misreports, and the detection method is effective for all types of misreports, resulting in an improved F1 score when perform entity prediction task. The performance only declines on disease entities, possibly due to the tendency to select other types of entities to fill co-occurrence relationships rather than making direct diagnoses.