

# LLM Sensitivity Evaluation Framework for Clinical Diagnosis

Chenwei Yan<sup>1,2</sup>, Xiangling Fu<sup>1,2,†</sup>, Yuxuan Xiong<sup>1,2</sup>, Tianyi Wang<sup>1,2</sup>,  
Siu Cheung Hui<sup>3</sup>, Ji Wu<sup>4,5</sup>, Xien Liu<sup>4,†</sup>

<sup>1</sup>School of Computer Science, Beijing University of Posts and Telecommunications

<sup>2</sup>Key Laboratory of Trustworthy Distributed Computing and Service(BUPT), Ministry of Education

<sup>3</sup>Nanyang Technological University

<sup>4</sup>Department of Electronic Engineering, Tsinghua University <sup>5</sup>College of AI, Tsinghua University

{chenwei.yan, fuxiangling, buptwty, bupt\_xyx}@bupt.edu.cn

asschui@ntu.edu.sg, {xeliu, wuji\_ee}@mail.tsinghua.edu.cn

## Abstract

Large language models (LLMs) have demonstrated impressive performance across various domains. However, for clinical diagnosis, higher expectations are required for LLM's reliability and sensitivity: thinking like physicians and remaining sensitive to key medical information that affects diagnostic reasoning, as subtle variations can lead to different diagnosis results. Yet, existing works focus mainly on investigating the sensitivity of LLMs to irrelevant context and overlook the importance of key information. In this paper, we investigate the sensitivity of LLMs, i.e. GPT-3.5, GPT-4, Gemini, Claude3 and LLaMA2-7b, to key medical information by introducing different perturbation strategies. The evaluation results highlight the limitations of current LLMs in remaining sensitive to key medical information for diagnostic decision-making. The evolution of LLMs must focus on improving their reliability, enhancing their ability to be sensitive to key information, and effectively utilizing this information. These improvements will enhance human trust in LLMs and facilitate their practical application in real-world scenarios. Our code and dataset are available at <https://github.com/chenwei23333/DiagnosisQA>.

## 1 Introduction

Recently, Large Language Models (LLMs) have demonstrated their capabilities in various real-world tasks (Kocmi and Federmann, 2023; Liu et al., 2023a). In the medical domain, many kinds of evaluations have been proposed to assess LLMs' understanding of medical knowledge and their accuracy in medical licensing examinations. Although many LLMs have achieved impressive performance in these evaluations, even approaching

<sup>†</sup>Corresponding authors.

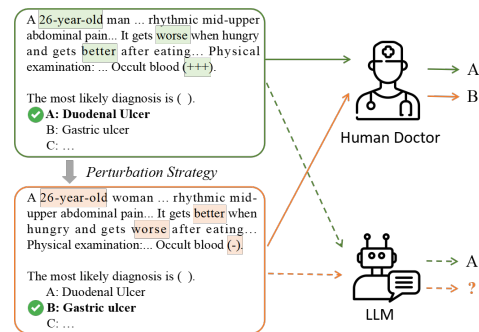


Figure 1: Human doctors are sensitive to key medical information. How do LLMs perform when key information is perturbed?

the comprehension level of human doctors in specific tasks (Singhal et al., 2023), they have not yet been widely adopted in real clinical scenarios (Chung et al., 2024). One main reason is due to the reliability and interpretability of LLM-generated results. As illustrated in Figure 1, in the clinical diagnosis process, human doctors usually pay much attention to certain key information, such as the presence or absence of certain symptoms, and positive or negative results of certain tests. It is because subtle changes in these key information can lead to different diagnostic outcomes. These key medical information include gender, age, clinical symptoms, medical checkup results and so on. They provide essential information for doctors to reach a more accurate diagnosis. However, whether LLMs can behave as human doctors to maintain the same level of sensitivity to such key medical information, and provide an accurate and practical assessment and diagnosis is one of the critical factors in ensuring their reliability.

Previous works have shown that LLMs are sensitive to certain information. For example, Pezeshkpour and Hruschka (2023) and Zheng et al. (2023) observed that LLMs are sensitive

to the order of options in multiple-choice questions (MCQs), while [Shi et al. \(2023\)](#) found that LLMs can be easily distracted by irrelevant context. However, existing works focus mainly on studying LLMs' sensitivity to changes on insignificant data aspects such as the order of options in MCQs. In real clinical practice, there are higher expectations for the sensitivity of LLMs, i.e., LLMs need to maintain a high level of sensitivity to key medical information as human doctors. When key medical information changes, LLMs should be able to detect it and ensure their responses remain sensible. This is crucial for enhancing their reliability and increasing human trust in LLMs. Currently, there is a lack of research on studying the sensitivity of LLMs to key medical information from a clinical perspective for evaluating the capabilities and limitations of LLMs.

In this paper, we propose a framework, named LLMSemEval, to evaluate the sensitivity of LLMs to key medical information, and assess their reliability for the clinical diagnosis. Then, we evaluate LLM's sensitivity by comparing their performance on the original and generated derived datasets. The results indicate that GPT-4 outperforms the other models, but there remains a gap in its sensitivity for real-world clinical application. Our contributions are summarized as follows:

- We propose a LLM sensitivity evaluation framework, named LLMSenEval, which provides a systematic approach for assessing the sensitivity of LLMs to key medical information in clinical diagnosis. To the best of our knowledge, this is the first work on proposing an evaluation framework to study the sensitivity of LLMs for clinical diagnosis.
- We propose a design schema for sensitivity evaluation, consisting of key information and perturbation strategy. We introduce four types of key medical information for clinical decision-making: age, gender, clinical symptoms, and checkup results. For each type of key medical information, we propose two perturbation strategies on change and removal.
- Moreover, we construct a diagnosis-related QA dataset, DiagnosisQA, based on the publicly available MedQA dataset. By applying the proposed perturbation strategies, we generate a series of sensitivity-focused datasets for evaluation. The

dataset has been publicly available on Github: <https://github.com/chenwei23333/DiagnosisQA>.

- The experimental results reveal that all the evaluated LLMs, including the best-performing GPT-4, have limitations in their sensitivity to key medical information. These findings point to new directions for developing more reliable medical LLMs.

## 2 Related Work

### 2.1 LLM Evaluation in the Medical Domain

The widespread adoption of LLMs ([Brown et al., 2020](#); [OpenAI, 2023](#); [Ouyang et al., 2022](#)) has recently attracted much interest in evaluating their performance for various applications. For medical domain, [Meng et al. \(2022\)](#) proposed a biomedical knowledge probing benchmark MedLAMA, and evaluated the medical knowledge understanding of multiple pre-trained large language models. [Singhal et al. \(2023\)](#) tested the performance of LLM on factuality, comprehension, reasoning, possible harm and bias. [Zhou et al. \(2024\)](#) investigated the degree and coverage of LLMs in encoding and mastering medical knowledge at four facets.

In addition to evaluating their capabilities of understanding medical knowledge, some assessments are conducted to determine whether LLMs can pass medical licensing examinations and tackle real-world clinical problems. [Liévin et al. \(2022\)](#) conducted an evaluation using the US Medical Licensing Examination and found that GPT-3.5 reaches the passing score. Moreover, [Cai et al. \(2023\)](#) proposed a Chinese medical examination benchmark, MedBench, for assessing the reasoning abilities of LLMs. [Liu et al. \(2023b\)](#) proposed CMExam, based on the Chinese National Medical Licensing Examination, and tested the performance of multiple general and medical domain-specific LLMs. However, although LLMs are able to achieve promising performance on many medical benchmarks, they still face many challenges for real-world applications, including issues on reliability and interpretability ([Zhang et al., 2023](#); [Zhou et al., 2024](#); [Cai et al., 2023](#)), and there is a lack of research focusing on sensitivity in medical applications.

### 2.2 Sensitivity Evaluation of LLMs

Many studies and practices have observed the limitations in the sensitivity of LLMs. In general, LLMs are sensitive to prompts, instructions and

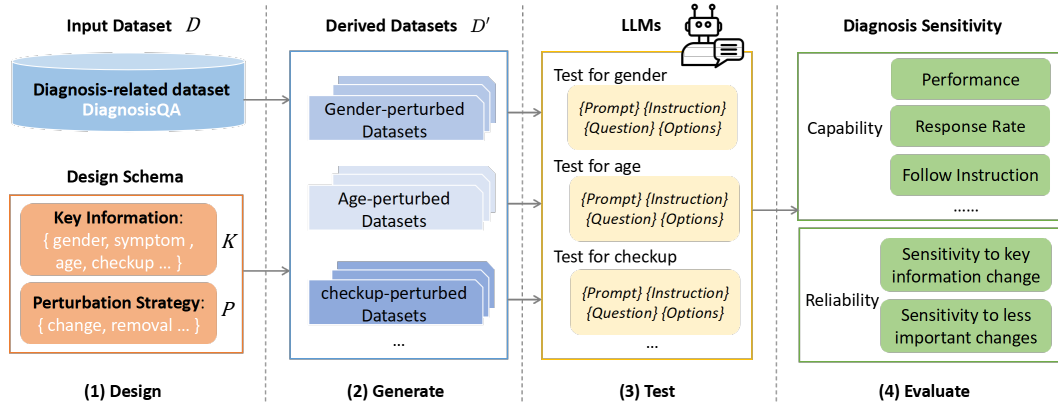


Figure 2: The proposed framework LLMSenEval for LLM sensitivity evaluation.

assigned roles to the extent that the inclusion of a sentence conveying reward or penalty in the prompt can affect their performance (Bsharat et al., 2024; Xu et al., 2023; Li et al., 2023).

Recent studies have shown that LLMs are sensitive to small changes in input context. Pezeshkpour and Hruschka (2023) observed that LLMs are sensitive to the order of options in multiple-choice questions, and Zheng et al. (2023) pointed out that LLMs may exhibit a preference for specific options. Moreover, Shi et al. (2023) found that LLMs can be easily distracted by irrelevant context.

Currently, the existing sensitivity studies mainly focus on how LLMs react to changes on insignificant data aspects. However, in clinical scenarios, apart from being undistracted by irrelevant information, a reliable LLM needs to be highly sensitive to important information. To our best knowledge, there is currently no related study on the evaluation of the sensitivity of LLMs to key medical information for clinical decision-making.

### 3 Proposed Framework

In this section, we propose an evaluation framework for LLM sensitivity, called LLMSenEval. As illustrated in Figure 2, the LLMSenEval consists of four steps: Design, Generate, Test and Evaluate.

#### 3.1 Design

In the Design, we propose a design schema, consisting of key information and perturbation strategy.

**Key Information.** It is crucial to identify the required key information for clinical diagnosis. In the proposed framework, gender, age, symptoms and checkup results are identified as key medical information, denoted as  $K =$

$\{k_{gender}, k_{age}, k_{symptom}, k_{check}\}$ . Gender and age are fundamental demographic characteristics in medical research, closely related to the occurrence and prognosis of numerous diseases. Clinical symptoms directly reflect a patient’s physical condition, providing vital clues to narrow down the range of possible medical conditions of the patient. Combined with the medical checkup results, a comprehensive assessment and diagnosis can then be made. Overall, these key information provide important clues for doctors to make a correct diagnosis.

**Perturbation Strategy.** Perturbing the key information is the basis for investigating the sensitivity of LLMs. We propose two perturbation strategies, namely key information change and key information removal, which are denoted as  $P$ . Based on the DiagnosisQA dataset, the details and examples of these strategies for perturbing each type of key medical information are shown in Table 1.

#### 3.2 Generate

Given a clinical diagnosis dataset  $D$  containing  $n$  documents, denoted as  $D = \{d_1, d_2, \dots, d_n\}$ , we extract the key information values for each document  $d_i$  ( $1 \leq i \leq n$ ) to generate a key-value dictionary  $Info$  as follows:

$$Info = \{k : V_k\}, \quad (1)$$

where  $k \in K$  is one of the four types of key medical information, and  $V_k$  is the corresponding value of  $k$  extracted from  $d_i$ . In our work, the values of key medical information are extracted using regular expression matching and keyword recognition (Chapman et al., 2001).

Then, we apply the perturbation strategies  $P$  to the key medical information  $K$ . The document generated by perturbing the key

Key Medical Information	Perturbation Strategy	Method	Example	Derived Dataset
Gender	Gender Change	Swap the gender	female -> male	$D_{GC}$
	Gender Removal	Use gender-neutral words	man -> patient	$D_{GR}$
Age	Age Change	Increase/decrease by 20%	10->12	$D_{AC}$
	Age Removal	Delete patient age	2-year-old boy -> boy	$D_{AR}$
Symptom	Symptom Change	Swap presence/absence	with -> without	$D_{SC}$
	Symptom Removal	Delete one clinical symptom	fever $\wedge$ cough -> cough	$D_{SR}$
Checkup	Results Change	Alter examination results	(+) -> (-)	$D_{CC}$
	Results Removal	Delete one result	CT reveals... -> None	$D_{CR}$

Table 1: Perturbation strategies on key medical information in the experiments.

information  $k \in K$  using perturbation strategy  $p \in P$  is denoted as  $d_i^{kp} \in D_{kp}$ . As a result, we obtain eight derived datasets:  $D' = \{D_{GC}, D_{GR}, D_{AC}, D_{AR}, D_{SC}, D_{SR}, D_{CC}, D_{CR}\}$  after applying the two perturbation strategies to the four types of key information.

### 3.3 Test

In this step, we evaluate the initial clinical diagnosis dataset  $D$  and the derived datasets  $D'$ . The input of LLMs is structured into prompt, instruction, question and options.

**Prompt.** The input prompt includes the role (who you are), the task (what the input is) and the requirement (how to do). For clinical diagnosis, the prompt is written as "You are a medical expert. The following is a medical exam question. Please give the correct option and the corresponding explanation."

**Instruction.** The instruction is given as "Return the result in JSON format with the following keys: Answer, Explanation.", which specifies the return format of LLMs.

**Question and Options.** The question and options are obtained from the QA question contents in  $D$  and its derived datasets  $D'$ .

### 3.4 Evaluate

The final step is to evaluate the responses using multiple metrics.

**Evaluation Principle.** The capabilities of LLMs can be evaluated from different perspectives. For example, comprehension and reasoning can assess their overall performance in understanding instructions and providing accurate predictions. Besides, sensitivity can be evaluated based on both changes in key information and less significant changes. A reliable LLM should be responsive to changes in

key information while maintaining stability under unimportant variations.

**Response Evaluation.** The ideal response should be a JSON-formatted string with the following information: question, options, answer and explanation, each containing a non-empty value. However, some responses do not follow the given instructions, and they just return the answer in plain text format. Therefore, there is a need to preprocess the responses through regular expression matching to extract the answer in the required format.

As illustrated in Figure 3, a valid response is a non-empty response, with the answer option chosen from the candidate options. In contrast, an invalid response occurs when LLMs refuse to provide a response, resulting in the "Answer" and "Explanation" fields being set to null. Additionally, if the responses do not include any options or diagnoses that match the candidate options, the values of "Answer" and "Explanation" are also set to null. When computing the metrics, such as accuracy, both valid and invalid responses are included, and empty responses are considered as wrong answers.

## 4 Experiments

In this section, we discuss the experiments on evaluating the sensitivity of LLMs for clinical diagnosis based on the proposed LLMSenEval framework. We first adopt the four types of key medical information discussed in the framework and propose two perturbation strategies. We then construct the DiagnosisQA dataset and generate the derived datasets for evaluation. After that, we select the LLMs for testing and define the metrics for evaluating the models' performance.

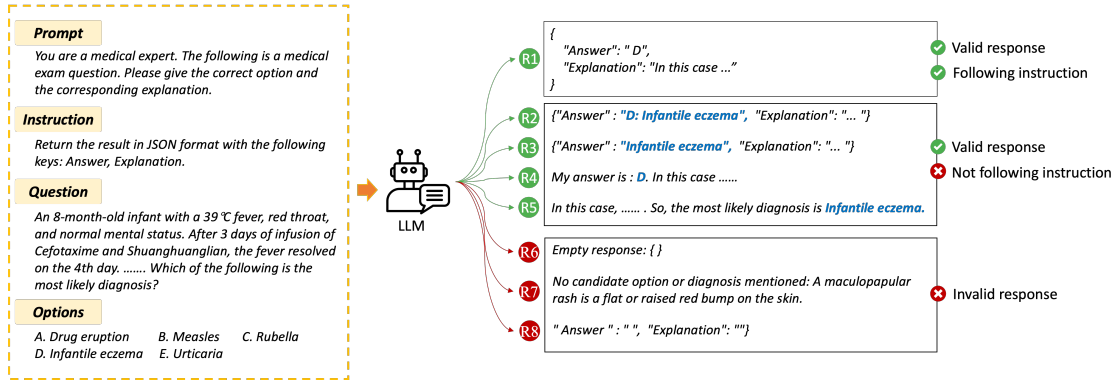


Figure 3: Valid and invalid responses from LLMs.

Dataset	# Total	# 5-Option	# 4-Option
DiagnosisQA	4,603	3,696	907
$D_{GC}, D_{GR}$	3,965	3,236	729
$D_{AC}, D_{AR}$	4,008	3,342	666
$D_{SC}, D_{SR}$	3,463	2,731	732
$D_{CC}, D_{CR}$	3,439	2,903	536

Table 2: Statistics of the DiagnosisQA dataset and eight derived datasets.

## 4.1 Dataset Generation

In the experiments, we construct a diagnosis-related QA dataset based on the MedQA dataset (Jin et al., 2021), by filtering out QA questions related to fundamental concepts and retaining those on case-based scenarios. Each case-based QA question provides details about patient symptoms, and checkup results, and concludes with "What is the most likely diagnosis?".

As a result, we obtain a dataset, named *DiagnosisQA* dataset, comprising 4,603 QA questions. Each question is accompanied by four or five candidate options. After applying perturbation strategies on DiagnosisQA, we obtain eight derived datasets. The QA questions that could not be perturbed due to missing key information are not included in the derived datasets. The statistics of the derived datasets are shown in Table 2.

Furthermore, we annotate the derived datasets with the help of four professional physicians and experts, confirming that the answers to each question in these datasets have been reviewed and corrected by human doctors, thereby ensuring the correctness and testability of the QA questions.

## 4.2 Selection of LLMs for Testing

We select five state-of-the-art LLMs for evaluation. From the GPT series, we choose the extensively evaluated versions: GPT-3.5-Turbo-0613 and the latest GPT-4-Turbo-2024-04-09. From the LLaMA series, we select LLaMA2-7b-chat (Touvron et al., 2023). Additionally, we also include Google’s Gemini-Pro and Anthropic’s Claude-3-haiku-20240307 for testing in the experiments.

For LLaMA2-7b-chat, we perform local deployment and inference. For the other models, we use APIs to invoke the service. All the model parameters are set with temperature of 0 and token limit of 1024. All the questions from the DiagnosisQA dataset and derived datasets are used for testing in the experiments.

## 4.3 Evaluation Metrics

In the experiments, we evaluate the LLM’s performance in their capability and sensitivity. To do this, we use four standard evaluation metrics, namely accuracy, macro precision, macro recall and macro F1-score. Among them, accuracy is used as the primary metric for sensitivity evaluation.

Moreover, we also introduce two additional metrics to better evaluate the LLM’s capability on its instruction understanding and instruction following. The first metric is the response rate (RR), which measures the number of valid responses of the model. The second metric is the followed-instruction rate (FIR), which evaluates how well the model follows the given instructions. These two metrics are defined as:

$$RR = \frac{\#validR}{N}, \quad (2)$$

$$FIR = \frac{\#followedInstructionR}{\#validR}, \quad (3)$$

Metrics	GPT-3.5	GPT-4	Gemini	LLaMA2-7b	Claude3
Accuracy	61.38	<b>78.95</b>	64.39	24.74	<u>65.37</u>
Precision	58.32	<b>68.78</b>	<u>62.00</u>	28.00	58.80
Recall	54.69	<b>65.43</b>	<u>53.74</u>	18.89	53.67
F1-score	55.14	<b>67.03</b>	<u>57.31</u>	16.44	55.92
RR	<b>99.95</b>	96.00	96.83	<u>97.84</u>	94.92
FIR	<u>98.92</u>	<b>99.98</b>	68.43	0	94.03

Table 3: Performances results (%) of LLMs on the DiagnosisQA dataset without any perturbations. The results are obtained from an average of two trials. The best scores are in **bold**, and the second-best scores are underlined.

where  $N$  is the total number of responses,  $\#validR$  is the number of valid responses, and  $\#followedInstructionR$  is the number of valid responses that follows the instructions.

## 5 Performance Analysis on DiagnosisQA

Table 3 shows the overall performance results of LLMs on DiagnosisQA dataset without any perturbations. GPT-4 exhibits the highest accuracy at 78.95%, significantly outperforming the second-ranked Claude3 by 13.58%, highlighting its superior performance on medical diagnosis tasks. Furthermore, GPT-4 also demonstrates robust results in precision, recall and F1-score with 68.78%, 65.43% and 67.03%, respectively, all of which also surpass the other models, indicating a well-balanced capability in both precision and recall. In contrast, LLaMA2-7b, likely due to its smaller parameter size, achieves much lower scores, with 24.74% in accuracy. Other models, such as Gemini and GPT-3.5, show moderate performance, with Gemini achieving an accuracy of 64.39% and GPT-3.5 reaching 61.38%.

In addition to these metrics, Table 3 also shows the performance results in RR and FIR. All models show high response rate, exceeding 94%. However, in terms of adherence to the given instructions, GPT-4 demonstrates an FIR of 99.98%, the highest among the five models. Conversely, LLaMA2-7b scores 0% in FIR, indicating its responses require more post-processing, as it fails to generate the instructed formatted outputs.

Overall, GPT-4 demonstrates the best overall performance based on DiagnosisQA, with its significantly higher accuracy and strict adherence to specified formats, showcasing its strong potential to serve as a medical AI assistant.

## 6 Sensitivity Analysis on Derived Datasets

To evaluate the sensitivity of LLMs, we conduct further experiments on eight derived datasets. Each dataset, based on manual annotation, contains two categories of questions. The first category is called *Same Answer Subset (SAS)*, where the correct answers to the QA questions remain unchanged after perturbations, as the diagnosis results are not affected by changes in specific key information. The second category is called *Different Answer Subset (DAS)*, where the correct answers are changed due to the influence of perturbations. In these questions, the correct answers are updated to other correct options. In addition, as some perturbed questions do not contain the correct answer option, we add an additional option "None of the above". Furthermore, some perturbed questions introduce logical inconsistency for medical diagnosis, for instance, after a gender swap, the patient's gender will no longer align with the organs or diseases mentioned in the question. As such, we add another option "The question contains inconsistency" to indicate such situations.

Overall, each QA question contains only one correct answer. Note that it is important to distinguish the difference between these two categories of questions. For example, a patient's gender may not affect the diagnosis of a common illness like cold, while a patient's symptoms or checkup results can significantly influence the diagnosis results.

Following this, we conduct experiments to investigate sensitivity from two perspectives. First, we evaluate whether LLMs have low sensitivity to questions in the *Same Answer Subset (SAS)* by measuring the change in accuracy before and after perturbations. A smaller difference indicates lower sensitivity, as these changes do not affect the final diagnosis results. Ideally, the accuracy should remain stable in this subset, as the correct diagno-

sis is not affected by the perturbations. Second, we evaluate whether LLMs have high sensitivity to questions in the *Different Answer Subset (DAS)* by measuring their accuracy. Higher accuracy indicates greater sensitivity to the changes, as the perturbations directly impact the correct answer.

### 6.1 Sensitivity Analysis on Gender

Table 4 shows the analysis results on gender. After applying gender change and removal perturbations, most question answers remain unchanged, as reflected in the *Same Answer Subset*. For this subset, the accuracy of LLMs varies, with some LLMs showing improvements while others degrading the performance. GPT-3.5 exhibits low sensitivity to gender changes and removals, as the  $\Delta$  accuracy lies within its standard error of 0.49%. In contrast, GPT-4 shows the most significant changes among the models. Its  $\Delta$  accuracy increases by 1.45% when gender is changed and by 1.98% when the gender is removed.

Subset		SAS		DAS	
Perturbation (Total)		Change (3,587)	Removal (3,961)	Change (378)	Removal (4)
Metric	SE	$\Delta$ accuracy		accuracy	
GPT-3.5	$\pm 0.49$	<b>-0.06</b>	<b>+0.22</b>	0	0
GPT-4	$\pm 0.22$	+1.45	+1.98	<u>32.01</u>	0
Gemini	$\pm 0.35$	-1.31	-1.11	0.79	0
Claude3	$\pm 0.22$	<b>-0.06</b>	+0.33	1.32	0
LLaMA2	$\pm 0.41$	<b>+0.36</b>	+1.06	0	0

Table 4: The sensitivity results on gender.  $\Delta$  indicates the difference in accuracy (%) before and after perturbations. The results on SAS that are within the SE range are highlighted in **bold**. The best results on DAS are underlined.

For the *Different Answer Subset*, where perturbations result in changes to the correct answers, GPT-4 displays exceptional sensitivity. GPT-4 is able to identify such changes, answering 121 questions from 378 questions correctly with an accuracy of 32.01%, far surpassing other models such as Gemini and Claude3, which can only answer 3 and 5 questions correctly, respectively.

### 6.2 Sensitivity Analysis on Age

Table 5 shows the analysis results on age. For the *Same Answer Subset*, all models consistently exhibit low sensitivity to both age changes and re-

Subset		SAS		DAS	
Perturbation (Total)		Change (4,004)	Removal (3,986)	Change (4)	Removal (22)
Metric	SE	$\Delta$ accuracy		accuracy	
GPT-3.5	$\pm 0.49$	<b>+0.47</b>	<b>-0.23</b>	25.00	0
GPT-4	$\pm 0.22$	+0.38	-1.20	<u>50.00</u>	0
Gemini	$\pm 0.35$	<b>-0.35</b>	-0.85	0	0
Claude3	$\pm 0.22$	+0.25	-0.40	25.00	<u>4.55</u>
LLaMA2	$\pm 0.41$	+0.47	-0.45	0	0

Table 5: The sensitivity results on age.

movals. Among them, the  $\Delta$  accuracy of GPT-3.5 remains within the standard error range under both change and removal perturbations. Although the  $\Delta$  accuracy of other models exceeds their standard error ranges, most deviations are below 0.5%, except in the case of age removal, where Gemini shows a decrease of 0.85% in  $\Delta$  accuracy and GPT-4 experiences a decline of 1.2%.

For the *Different Answer Subset*, due to the limited total number of related samples, GPT-4 can correctly answer 2 questions from 4 questions (i.e., 50%) under age change perturbation, while GPT-3.5 and Claude3 can only answer 1 question correctly with an accuracy of 25%. For age removal perturbation, Claude3 is the only model that can answer a question correctly from 22 questions (i.e., 4.55%), while all the other models are unable to answer any questions correctly. Overall, GPT-4 shows the highest sensitivity to age changes among the LLM models, while Claude3 achieves the highest sensitivity to age removal.

### 6.3 Sensitivity Analysis on Symptom

Table 6 shows the analysis results on symptom. For the *Same Answer Subset*, Gemini shows a drastic decrease in  $\Delta$  accuracy, dropping by 4.69% for symptom changes and 2.34% for removals. This shows its high sensitivity to symptom-related changes that should not influence the final diagnosis. In contrast, GPT-3.5 demonstrates the most stable performance, with a small decrease of 0.23% in  $\Delta$  accuracy for symptom changes, and an increase of 0.24% after symptom removal, both falling within its standard error range. The analysis results have suggested that GPT-3.5 is less sensitive to symptom changes that are unrelated to diagnosis results compared to other models.

For the *Different Answer Subset*, GPT-4 outper-

Subset		SAS		DAS	
Perturbation (Total)		Change (542)	Removal (2,395)	Change (2,921)	Removal (1,068)
Metric	SE	$\Delta$ accuracy		accuracy	
GPT-3.5	$\pm 0.49$	<b>-0.23</b>	<b>+0.24</b>	0.10	1.97
GPT-4	$\pm 0.22$	+1.88	-0.70	<u>3.08</u>	3.65
Gemini	$\pm 0.35$	-4.69	-2.34	1.37	<u>3.75</u>
Claude3	$\pm 0.22$	-1.41	-1.88	0.86	0.94
LLaMA2	$\pm 0.41$	-5.87	-4.93	0.03	0.47

Table 6: The sensitivity results on symptom.

forms other models, correctly answering 90 questions from 2,921 questions (i.e., 3.08%) for symptom change. In terms of symptom removal, Gemini can correctly answer 40 questions from 1068 questions (i.e., 3.75%), slightly surpassing GPT-4’s accuracy of 3.65%. The results indicate that GPT-4 is most sensitive to symptom changes, while Gemini is most sensitive to symptom removal. Other models, such as LLaMA2 and Claude3, show difficulty in achieving good performance after perturbations.

#### 6.4 Sensitivity Analysis on Checkup

Table 7 shows the analysis results on checkup. For the *Same Answer Subset*, GPT-4 exhibits a slight increase of 0.37% in  $\Delta$  accuracy for checkup changes and a good performance for checkup removals. This suggests that GPT-4 achieves low sensitivity to irrelevant checkup perturbations. In contrast, GPT-3.5 and Claude3 experience more noticeable accuracy changes, particularly GPT-3.5, which shows a decrease of 4.26% for checkup changes and 5.90% for removals in  $\Delta$  accuracy, indicating high sensitivity to information that should not impact the diagnosis results.

Subset		SAS		DAS	
Perturbation (Total)		Change (1,719)	Removal (3,028)	Change (1,720)	Removal (411)
Metric	SE	$\Delta$ accuracy		accuracy	
GPT-3.5	$\pm 0.49$	-4.26	-5.90	0.52	4.14
GPT-4	$\pm 0.22$	<b>0</b>	+0.37	<u>4.13</u>	<u>5.25</u>
Gemini	$\pm 0.35$	-2.50	-2.68	1.74	2.19
Claude3	$\pm 0.22$	-1.83	-0.31	0.87	1.70
LLaMA2	$\pm 0.41$	<b>-0.06</b>	-1.76	0.23	0.73

Table 7: The sensitivity results on checkup.

For the *Different Answer Subset*, GPT-4 once again stands out, correctly answering 71 and 22 questions from 1720 checkup change questions and 411 checkup removal questions, with an accuracy of 3.08% and 3.65%, respectively. The results demonstrate GPT-4’s ability to accurately detect significant checkup changes, suggesting it maintains the high sensitivity in scenarios where the diagnosis is affected by checkup changes.

## 7 Discussion

Figure 4 shows the overall sensitivity performance of five LLMs. The bar chart illustrates the average  $\Delta$  accuracy under two perturbations for each model on the *Same Answer Subset* across the four kinds of key medical information. Meanwhile, the line chart depicts the total number of correct answers provided by the LLMs in the *Different Answer Subset*.

From the perspective of whether LLMs have low sensitivity to questions in the *Same Answer Subset*, GPT-3.5 is less sensitive to gender, age and symptom information, while GPT-4 is less sensitive to checkup.

From the perspective of whether LLMs have high sensitivity to questions in the *Different Answer Subset*, GPT-4 stands out with the highest number of correct answers (345), while Gemini follows with 122 correct answers.

Overall, The results demonstrate GPT-4’s high sensitivity to changes in key medical information that can alter diagnosis results, while maintaining moderate sensitivity to changes that do not affect diagnosis results. Gemini ranks as the second-best model in terms of sensitivity for the *Different Answer Subset*; however, it also shows high sensitivity to less important changes, particularly symptom-related changes, which negatively impacts the robustness of the LLM. In contrast, Claude3 shows low sensitivity to the less important changes, but it also shows low sensitivity to the perturbations that affect diagnosis results, suggesting it may pay less attention to details.

However, even though GPT-4 has the highest sensitivity to important changes in key medical information among all models, demonstrating its potential as a medical diagnosis AI assistant, there remains a significant gap in its reliability when applied to real-world scenarios, as the accuracy only achieves 5.28% (totally answering 345 questions from 6,528 questions correctly). Not to mention,



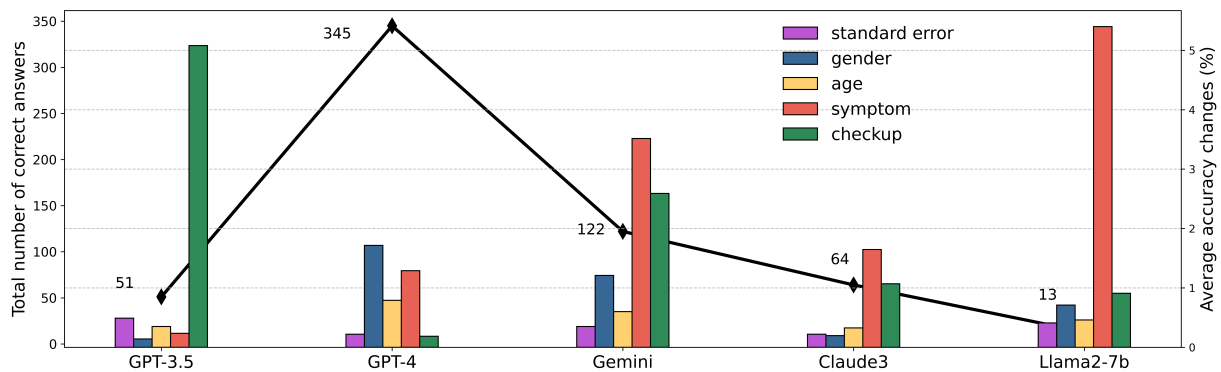


Figure 4: The overall sensitivity performance of five LLMs. The bar chart shows the average difference in accuracy on the Same Answer Subset, and the line chart shows the total correct answers provided by the LLMs on the Different Answer Subset.

GPT-4 is the most sensitive to gender perturbations among five LLMs. In conclusion, there is still a long way to go in reducing the sensitivity of LLMs to unimportant perturbations while enhancing their sensitivity to crucial ones.

## 8 Conclusion

In this paper, we investigated the sensitivity of large language models (LLMs) to key medical information in clinical diagnostic decision-making. Our work is the first to explore the limitations of LLM’s sensitivity from a clinical perspective. First, we proposed a LLM sensitivity evaluation framework, within which we introduced four types of key medical information and designed two kinds of perturbation strategies. Based on this framework, we constructed a diagnosis-related dataset, DiagnosisQA, along with eight derived datasets. We then evaluated five state-of-the-art LLMs on these datasets. The evaluation results reveal the limitations in large language model’s ability to effectively capture key medical information. Most LLMs exhibit poor sensitivity to the key medical information. Although large language models can achieve good performance on medical benchmarks, they still have substantial limitations in their sensitivity to key information in clinical diagnosis.

## 9 Limitation

The dataset employed in the experiment was refined from MedQA. While these case-based QA questions closely resemble actual Electronic Medical Records (EMRs) used in clinical diagnosis, there are notable differences: (1) the information within the QA questions has been expert-validated and is highly relevant to problem-solving, whereas clinical

EMRs in practice are generally longer and more comprehensive; (2) QA questions cover a limited number of diagnoses, leaving many common diagnoses uncovered, and the sensitivity of uncovered diagnoses is not investigated. Consequently, this presents a significant challenge for large language models. When applying LLMs to clinical practice, there remains a gap compared to our investigated results.

The decision-making process in clinical diagnosis is highly sensitive to numerous crucial factors. In this context, we have only discussed four key medical information elements: gender, age, symptoms, and checkup results. However, there are other aspects not covered, such as family medical history. All of this information plays a pivotal role in guiding doctors to make accurate diagnosis.

## Acknowledgments

This research was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0506501), National Key R&D Program of China (2021YFC2500803) and Program of China Scholarship Council (Project ID:202306470061).

## References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario

- Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2024. [Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4](#). *Preprint*, arXiv:2312.16171.
- Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. [Med-bench: A large-scale chinese benchmark for evaluating medical large language models](#). *Preprint*, arXiv:2312.12806.
- Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. [A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries](#). *Journal of Biomedical Informatics*, 34(5):301–310.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams](#). *Applied Sciences*, 11(14):6421.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). *Preprint*, arXiv:2302.14520.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. [Large language models understand and can be enhanced by emotional stimuli](#). *Preprint*, arXiv:2307.11760.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023a. [Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation](#). *Preprint*, arXiv:2305.01210.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023b. [Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset](#). *Preprint*, arXiv:2306.03030.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2022. [Can large language models reason about medical questions?](#) *arXiv*.
- Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. [Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- 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](#). *Preprint*, arXiv:2203.02155.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. [Large language models sensitivity to the order of options in multiple-choice questions](#). *arXiv preprint arXiv:2308.11483*.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *International Conference on Machine Learning*, pages 31210–31227. PMLR.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutí Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyang Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. 2023. Enhancing small medical learners with privacy-preserving contextual prompting. *arXiv preprint arXiv:2305.12723*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

Yuxuan Zhou, Xien Liu, Chen Ning, and Ji Wu. 2024. [Multifaceteval: Multifaceted evaluation to probe llms in mastering medical knowledge](#). *Preprint*, arXiv:2406.02919.

## A The Impact of Prompts

To investigate the effects of prompts, we conduct a comparative experiment on the DiagnosisQA with empty prompt in the model input, but keeping the fields on Instruction, Question and Options. The results are shown in Figure 5.

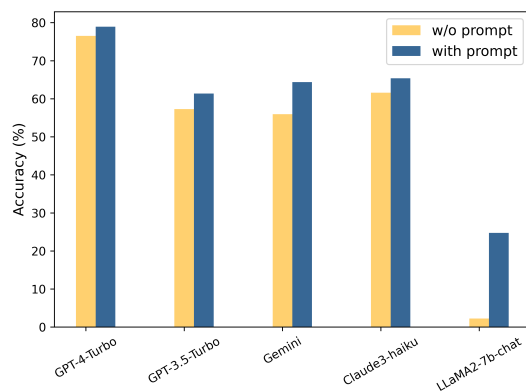


Figure 5: The impact of prompts on the performance of LLMs on the DiagnosisQA dataset.

Overall, the use of prompts leads to a general improvement in the performance of all LLMs, indicating that prompts play a vital role in enhancing the understanding capability of LLMs. By incorporating elements such as the role, task and requirements. Specifically, for those LLMs with good performance, prompts bring a certain improvement. However, for the model with smaller parameters, i.e. LLaMA-7b, its performance drops sharply without prompts. Without prompts, LLaMA-7b can hardly understand the model input.

## B Case Study

To demonstrate the impact of key medical information on diagnosis, we give an example from DiagnosisQA, illustrated in Table 8. The gender change perturbation is applied to the question. The original question involves a female patient, and the gender change perturbation causes a conflict between the patient's gender and situation. Consequently, the correct answer becomes "The question contains inconsistencies". In the evaluation, we find that both GPT-4 and Gemini select option F and note the conflict between the male patient and the described symptoms and candidate diseases in their explanations. They also recommended patient to take further examinations. However, GPT-3.5 and Claude3 did not recognize this issue.

	Original	With Gender Change Perturbation
Question	<p>A 14-year-old <u>girl</u> is brought to the pediatrician by her mother. The <u>girl's</u> mother states that <u>she</u> began having her period 6 months ago. The patient states that after <u>she</u> first period <u>she</u> has had a period every 10 to 40 days. <u>Her</u> menses have ranged from very light flow to intense and severe symptoms. Otherwise, the patient is doing well in school, is on the track team, and has a new <u>boyfriend</u>. <u>Her</u> temperature is 98.1°F (36.7°C), blood pressure is 97/58 mmHg, pulse is 90/min, respirations are 14/min, and oxygen saturation is 99% on room air. Physical exam demonstrates an obese <u>girl</u> but is otherwise within normal limits. Which of the following is the most likely diagnosis?</p>	<p>A 14-year-old <u>boy</u> is brought to the pediatrician by <u>his</u> mother. The <u>boy's</u> mother states that <u>he</u> began having his period 6 months ago. The patient states that after <u>his</u> first period <u>he</u> has had a period every 10 to 40 days. <u>His</u> menses have ranged from very light flow to intense and severe symptoms. Otherwise, the patient is doing well in school, is on the track team, and has a new <u>girlfriend</u>. <u>His</u> temperature is 98.1°F (36.7°C), blood pressure is 97/58 mmHg, pulse is 90/min, respirations are 14/min, and oxygen saturation is 99% on room air. Physical exam demonstrates an obese <u>boy</u> but is otherwise within normal limits. Which of the following is the most likely diagnosis?</p>
Options	<p>A. Normal development [Correct Answer Before Perturbation]            B. Polycystic ovarian syndrome            C. Pregnancy            D. Premenstrual dysphoric disorder            E. Premenstrual tension            F. The question contains inconsistencies [Correct Answer After Perturbation]</p>	
GPT-4		<p>F. The scenario describes a 14-year-old boy experiencing menstrual periods, which is <b>physiologically impossible as males do not have the reproductive organs necessary for menstruation</b>. The presence of 'periods' in a male patient suggests a different underlying issue that may be related to hormonal imbalances, endocrine disorders, or other medical conditions that need to be investigated further. <b>None of the provided options (A-E) are applicable to a male patient</b> experiencing these symptoms. ✓</p>
GPT-3.5		<p>B. The most likely diagnosis in this case is polycystic ovarian syndrome (PCOS). PCOS is a hormonal disorder that affects women of reproductive age. It is characterized by irregular menstrual periods, which can range from very light flow to intense and severe symptoms. Other common symptoms of PCOS include obesity. The patient's age, irregular periods, and obesity are all suggestive of PCOS. The other options are less likely based on the given information. ✗</p>
Gemini		<p>F. <b>The question contains inconsistencies</b>. The patient is a 14-year-old boy who has been having menstrual periods for 6 months. This is not normal development, <b>as boys do not typically have periods</b>. Polycystic ovarian syndrome, pregnancy, premenstrual dysphoric disorder, and premenstrual tension are all conditions that can cause irregular periods in women, but <b>they are not typically seen in boys</b>. The patient should be evaluated further to determine the cause of irregular periods. ✓</p>
Claude3		<p>B. Polycystic ovarian syndrome (PCOS) is a common endocrine disorder that affects women of reproductive age. It is characterized by irregular periods, excess androgen levels, and the presence of cysts on the ovaries. PCOS can cause a variety of symptoms, including acne, hirsutism, weight gain, and infertility. In this case, the patient is a 14-year-old boy who has been having irregular periods for the past 6 months. He is also obese, which is a common symptom of PCOS. ✗</p>

Table 8: A case study of the perturbed questions with the answer of "The question contains inconsistencies".