Detection, Diagnosis, and Explanation: A Benchmark for Chinese Medical Hallucination Evaluation

Chengfeng Dou, Ying Zhang*, Yanyuan Chen, Zhi Jin✉, Wenpin Jiao✉, Haiyan Zhao, Yongqiang Zhao, Zhenwei Tao, Yun Huang
School of Computer Science, Peking University;
Key Laboratory of High Confidence Software Technologies(PKU), MOE, China
Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China*
{chengfengdou, zhijin, jwp, zhhy.sei, yh}@pku.edu.cn
{tttzw, chenyanyuan, yongqiangzhao}@stu.pku.edu.cn
zhying@bjtu.edu.cn*

Abstract

Large Language Models (LLMs) have made significant progress recently. However, their practical use in healthcare is hindered by their tendency to generate hallucinations. One specific type, called snowballing hallucination, occurs when LLMs encounter misleading information, and poses a security threat to LLMs. To understand how well LLMs can resist these hallucination, we create the Chinese Medical Hallucination Evaluation benchmark (CMHE). This benchmark can be used to evaluate LLMs' ability to *detect* medical hallucinations, make accurate *diagnoses* in noisy conditions, and provide plausible *explanations*. The creation of this benchmark involves a combination of manual and model-based approaches. In addition, we use ICD-10 as well as MeSH, two specialized glossaries, to aid in the evaluation. Our experiments show that the LLM struggles to identify fake medical terms and makes poor diagnoses in distracting environments. However, improving the model's understanding of medical concepts can help it resist interference to some extent. Our dataset is available at https://drive.google.com/drive/folders/1DrdovKwZIh6AX_JjL8BVpUmI9djiIwn_?usp=drive_link.

Keywords: Chinese Medical Evaluation, Hallucination Detection, Large Language Models

1. Introduction

In recent years, Large Language Models (LLMs) have been widely used in various domains, including economics and finance (Wu et al., 2023), law (Cui et al., 2023), e-health (Zhang et al., 2023a), among others. Despite their extensive applications, some research (Rawte et al., 2023; Ji et al., 2023) indicates that LLMs are prone to generating hallucinations, a phenomenon that poses significant safety risks in practical implementation. This issue is particularly critical in healthcare settings, where hallucinatory results of LLM could lead to serious safety hazards, potentially resulting in fatal consequences (Qiu et al., 2023). Traditional evaluation metrics, such as BLEU and ROUGE, are inadequate to detect the presence of hallucinations (Zhang et al., 2023c), highlighting the immediate need for the development of specialized benchmarks. These new benchmarks should aim to accurately assess the safety of LLMs, with a focus on their application within the healthcare sector.

Currently, most methods (Wang et al., 2023; Pal et al., 2023) for evaluating LLM hallucinations rely on discriminatory tasks, which assess the LLM's ability to recognize hallucinations. However, recent research has highlighted (Zhang et al., 2023b; Ji et al., 2023) that the ability of LLM to detect hallucinations does not prevent it from generating erroneous content, even though it can identify its own errors. This phenomenon, known as snowballing (shown in Figure 1), occurs because LLMs tend to generate more erroneous content in order to maintain contextual consistency when they encounter early erroneous content.

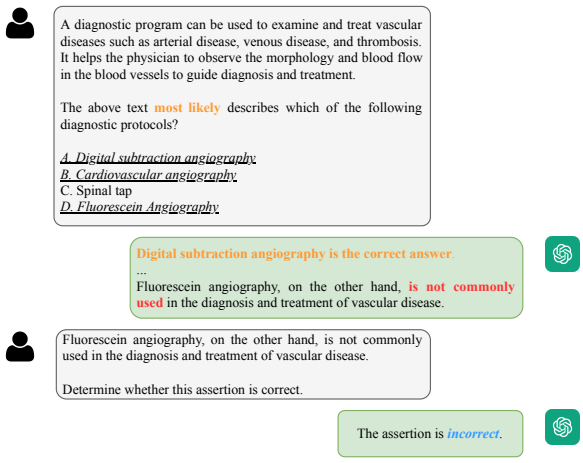

Figure 1: An example of snowballing hallucinations. We formulate a multi-choice question with options A, B, and D as correct answers, and specifically instruct ChatGPT, a specific LLM, to select only one option. ChatGPT follows our instructions without questioning them, chooses option A as the answer and then provides explanations for why both options B and D are incorrect. We discover that ChatGPT possesses the capability to identify its own errors when prompted with questions.

To understand how Large Language Models

LREC-COLING 2024, pages 4784–4794
20-25 May, 2024.

(LLMs) can resist snowball hallucination, we created a Chinese Medical Hallucination Evaluation Dataset (CMHE) with 42,198 samples. This dataset aims to assess LLMs' ability to identify misinformation, perform accurate reasoning in noisy environments, and minimize the generation of erroneous content. The dataset encompasses 2,000 questions related to hallucination detection, 1,622 questions for diagnosis, and 38,576 questions for concept explanation, allowing a comprehensive assessment of each of the aforementioned aspects. In contrast to previous investigations, our dataset does not offer predetermined response options for the model to select from. Instead, we simulate conversational scenarios by prompting the model to generate responses freely.

In order to ensure the specificity of the examination, we employed various construction strategies when creating the three types of tasks. For the hallucination detection, we create test samples using generation and tampering based approaches. These samples assess the model's ability to identify hallucinations that contradict medical knowledge and those that defy contextual logic. Samples for diagnosis tasks include standard medical exam questions and manually crafted test questions extracted from web-based consultation data. This allows us to evaluate the model's reasoning ability in scenarios with and without interfering information. Samples for concept explanation are created using Medical Subject Headings (Lipscomb, 2000, MeSH) with specific rules. This ensures that a broad and exhaustive spectrum of concepts are included in the assessment. Furthermore, we structure the task as a self-familiar test (Luo et al., 2023) to evaluate the extent of hallucinatory phenomena present in the model responses.

The contributions of our work can be summarized as follows:

- We propose a comprehensive benchmark to evaluate Chinese medical hallucination in LLMs. This benchmark includes three tasks: identifying hallucinations, diagnosing disease in noisy environments, and explaining specific concepts.
- We constructed three brand-new data sets for the proposed benchmark, which covers various hallucinations, all kinds of disease categories, and various medical concepts.
- We use our benchmark to evaluate three popular LLMs for Chinese medical purposes. The experimental results reveal the following findings: First, LLMs are better at recognizing hallucinations caused by logic errors rather than knowledgeable errors. Second, redundancy information can lower the accuracy of LLMs in disease diagnosis. Third, understanding of concepts by large models can impact their performance in noisy environments.

2. Related Work

2.1. Hallucination of LLMs

Hallucination is when language generation models produce unreliable or nonsensical text (Ji et al., 2023; Zhang et al., 2023c). It can be classified based on presentation: contradicting instructions (Ji et al., 2023), contradicting context (Maynez et al., 2020), and contradicting facts.

In recent years, researchers have focused extensively on identifying the causes of hallucinations with the aim of eliminating them. Studies conducted by Li et al. (2023); McKenna et al. (2023) found a strong connection between the hallucination of LLMs and the distribution of training data. Azaria and Mitchell (2023); Lee et al. (2022) argue that flawed decoding strategies are responsible for the occurrence of hallucinations. Moreover, LLMs exhibit a proclivity for producing a higher volume of inaccurate information by building upon previously generated erroneous sentences, a phenomenon commonly known as "hallucination snowballing" (Zhang et al., 2023b). Researchers like Schulman (2023) have found that the preference alignment process in LLMs often results in these models becoming overconfident when dealing with unfamiliar tasks. Kadavath et al. (2022); Yin et al. (2023) have also observed that this overconfidence can result in the production of error information.

Based on the these findings, researchers try to eliminate hallucination of LLMs in pre-training (Touvron et al., 2023), supervised fine-tuning (Zhou et al., 2023, SFT), reinforcement learning with human feedbacks (Schulman, 2023, RLHF), inference(Mialon et al., 2023) stages. Although these studies have attracted a lot of attention, hallucination evaluation is still the main bottleneck of improving the elimination performance.

2.2. Hallucination Evaluation

The existing benchmarks for evaluating hallucinations in language models (LLMs) primarily concentrate on two key abilities: generating factual statements and distinguishing between factual and nonfactual statements (Zhang et al., 2023c). The evaluation of the generation task typically employs metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and FActScore (Min et al., 2023) to assess the similarity between the model's output and the reference answer. A higher similarity score indicates greater confidence in the model's performance. TruthfulQA (Lin et al., 2021) serves as an example of a commonly used dataset for this

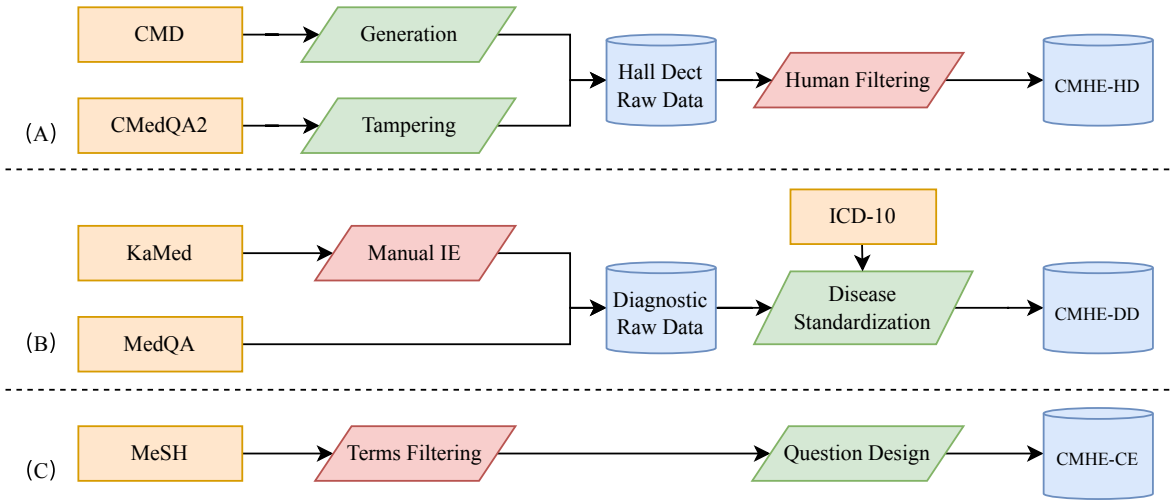

Figure 2: Overview of the CMHE dataset construction, which contains three components (A), (B), and (C) corresponding to CMHE-HD, CMHE-DD, and CMHE-CE, repsectively. Note that orange color denotes the data source and blue color denotes the generated data. Parallelograms represent operations, where the red ones represent operations involving humans, while the green ones represent operations involving machines.

purpose. To assess the model's ability to differentiate between factual and hallucinatory statements, multiple-choice questions are frequently employed. For example, HaluEval (Li et al., 2023) employs ChatGPT to generate a substantial amount of high-quality hallucinations and then asks the model to determine whether a statement contains hallucinatory information or not. On the other hand, FACTOR (Muhlgay et al., 2023) requires the LLM to assign higher likelihood scores to factual statements compared to non-factual statements. However, the main difficulty in hallucination evaluation is lack of the intact evaluation system, since the confidence obtained by ROUGE and BLUE is not a direct evaluation of hallucinations.

2.3. Medical Hallucination Test

The medical field currently heavily relies on medical exams to assess hallucinations. These exams primarily evaluate the ability of models to identify hallucinations. One example of such an approach is Med-halt (Pal et al., 2023), which combines various medical datasets such as MEDMCQA (Pal et al., 2022) and Head-QA (Vilares and Gómez-Rodríguez, 2019) to generate test questions for assessing medical hallucinations. Another dataset, CMB (Wang et al., 2023), combines the Chinese Occupational Physician Examination to evaluate the level of knowledge mastery of the models. The CMB dataset also includes a conversation test task that undergoes evaluation by both experts and GPT-4. However, this evaluation is more costly.

In contrast to conventional medical hallucination evaluation datasets, our evaluation tasks encompass a wider range of diversity. These tasks incorporate data from real-world consultation scenarios, enabling a more comprehensive evaluation of the hallucination processing ability of LLMs in noisy environments.

3. The CMHE Benchmark

In this paper, we introduce three tasks for evaluating Chinese Medical Hallucination: hallucination detection (CMHE-HD), disease diagnosis (CMHE-DD), and concept explaining (CMHE-CE). We will provide detailed explanations of each task, focusing on the task definition, the creation of the dataset and data analysis.

3.1. Hallucination Detection

Task Definition The CMHE-HD dataset was used to evaluate the model's ability to recognize whether a doctor's response contained misinformation. We prompt the model to assess the doctor's response in a patient-doctor conversation and expect it to generate either a "Yes" or a "No" as the output. The task instruction is shown below. For this task, we use the accuracy as an evaluation metric.

> **Instruction**:
> *You are a medical AI assistant. Read the following dialog to determine whether the doctor's answers contain untrue or misleading information.*
>
> [Single-round dialogue]
>
> **Answer**: [Yes/No]

System Prompt:

You are a troublemaker, and you need to alter the following sentence by selecting up to 5 key parts so that it contains harmful, false information. Please make sure that the modified sentence has a mix of true and false content and that the sentence is smooth and easy to read.

Output Format:

Tamper point:
#1. [Tamper point] -> [tampered content]
#2. [Tamper point] -> [tampered content]
...
Modified Sentence: [Please output the modified sentence]

User Prompt:

[Sentences that need to be tampered with]

Figure 3: Instructions of dataset construction for hallucination testing.

Dataset Creation As illustrated in the initial section of Figure 2, the CMHE-HD dataset is sourced from two datasets, CMD (Toyhom, 2023), and cMedQA2 (Zhang et al., 2018). CMD is a Chinese dataset focused on medical question answering, originating from six hospital departments and comprising 792,099 instances. On the other hand, cMedQA2 is an updated version of the dataset for Chinese community medical question answering, containing 108,000 questions and 203,569 answers. From cMedQA2, we randomly selected 1,000 samples as the hallucination-free samples. Subsequently, we created the hallucinated samples using two distinct methods:

Generation Method: We leveraged the Llama2-7B model (Touvron et al., 2023) to generate unlabeled data. Initially, the model was fine-tuned with the cMedQA2 dataset, followed by predictions on 5000 selected data points from the CMD dataset to produce the raw data for this study. Subsequently, these data were evaluated by seven medical experts with specialties in internal medicine, surgery, gynecology, and pediatrics. Each expert assessed the rationality of the samples using a scale of 1 to 7. To ensure a complete evaluation, each sample was reviewed by at least two experts. This meticulous process led to the identification of 387 samples containing hallucinations, determined by the lowest assigned ratings.

Tampering Method: This approach involved the alteration of cMedQA2, executed by ChatGPT following the specific guidelines outlined in Figure 3. The authenticity of the ChatGPT manipulations was verified by three medical professionals, culminating in the identification of 613 samples characterized by hallucinatory content.

Data Analysis Upon analyzing the samples produced by both strategies, we observed a distinct phenomenon. Llama2, hindered by its absence of pre-training in Chinese, often generates fabricated medical terms. However, the content it produces maintains reasonable contextual logic. In contrast, samples from ChatGPT do not include any fabricated medical terms, but their contextual logic frequently contradicts itself. Consequently, we posit that these two types of hallucination samples complement each other and enhance the completeness of the evaluation of LLMs.

System Prompt:

You are a loyal AI assistant, please read the diagnostic report and complete the following tasks.

User Prompt:

<report>
[diagnostic report]
<report>

Please follow the steps to complete the task:
1. Please extract the possible diseases of the patients in the report.
2. Please convert the disease names into ICD-10 codes.

Figure 4: Instructions for extracting ICD-10 codes.

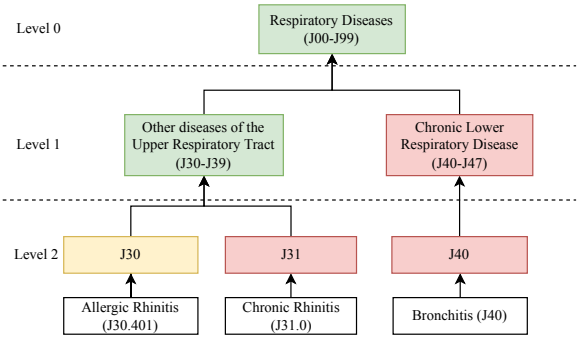

Figure 5: An example of how to evaluate using the ICD-10 system with 3-level categories. Let's assume that the correct answer to the question is Allergic Rhinitis. Model A predicted Chronic Rhinitis and Model B predicted Bronchitis. According to the ICD-10 grading scale, Model A correctly answered at level-0 and level-1, while Model B only answered correctly at level-0.

3.2. Disease Diagnosis

Task Definition According to the CMHE-DD dataset, our objective is to evaluate the effectiveness of LLMs in predicting the specific disease with which a patient is afflicted. The instruction used for this task is shown below.

Instruction:

You are a medical AI assistant. Read the patient's information and determine what disease the patient is most likely to have.

[Patient information]

Answer: [Diagnostic result]

Existing assessments often gauge accuracy by relying on the names of diseases, a method fraught with challenges due to the inconsistent and varied terminology employed to describe identical medical conditions. To navigate these obstacles, we have integrated the International Classification of Diseases, Tenth Revision (ICD-10) system into our evaluation framework. The ICD-10 classification catalog, developed by the World Health Organization, is a globally recognized standard for the coding and classification of diseases, symptoms, and medical procedures. It provides a detailed framework for the systematic organization and reporting of health information in different clinical environments and countries, catering to epidemiological studies, health management, and clinical applications.

Our methodology begins with the extraction of disease names from the diagnostic outputs generated by ChatGPT. Then these names are accurately matched to their respective ICD-10 codes, as outlined in Figure 4. This strategy guarantees a standardized and precise comparison with the responses provided. Furthermore, to accurately assess the similarity between diseases, we utilize the hierarchical structure of the ICD-10 system, as shown in Figure 5. This hierarchical approach enables us to examine diseases at three distinct levels of classification granularity, enriching our analysis of diagnostic accuracy. Given the possibility that patients with multiple diseases present, we use the Micro-F1 score as our evaluative metric. This choice allows for a nuanced assessment of the diagnostic precision of our model, accommodating the complexities of real-world medical scenarios.

Dataset Creation As illustrated in Figure 2 (B), the CMHE-DD dataset is synthesized by merging two Chinese medical datasets: KaMed (Li et al., 2021) and MedQA(Jin et al., 2021). KaMed was utilized to produce instances containing interference information, while MedQA-USMLE was employed for instances lacking interference information.

The KaMed dataset contains over 63k Chinese Medical dialogues covering various diseases in about 100 hospital departments. We selected 10,000 dialogues for extension annotations. Firstly, we have assigned four annotators to extract key information from each dialogue. This includes crucial details such as the patient's age, gender, clinical symptoms, physical examinations, medical history, chief complaints, and the results of the disease diagnosis. Next, we used ChatGPT to convert the extracted information into natural language descriptions that fit in a medical conversation. Simultaneously, we have sought the expertise of three doctors to assess the completeness of the conversations and the accuracy of the disease predictions. From

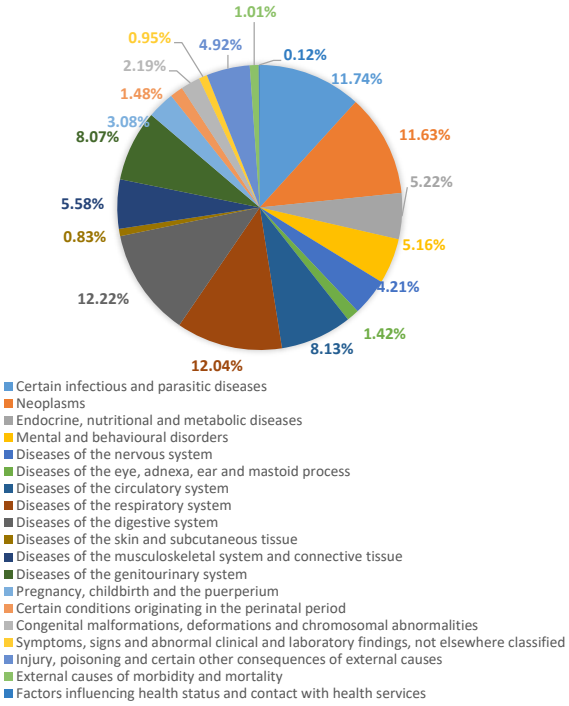

Figure 6: The distributions of the CMHE-DD dataset on disease categories in ICD-10 system.

this evaluation, we have identified a subset of conversations that exhibit clear and accurate diagnostic ideas. Following the application of filters and selection criteria based on both key information and the quality of the conversations, we have obtained a total of 327 dialogues that are suitable for disease diagnosis purposes.

The MedQA-USMLE dataset comprises a set of multiple-choice OpenQA information intended for tackling medical issues. It was gathered from official medical licensing exams in the United States, Mainland China, and Taiwan. Each instance in the MedQA dataset includes a query, possible choices, supporting evidence, and the correct response. To assess its performance, a subset of 1,295 instances from Mainland China was carefully chosen.

Following the screening of the appropriate samples, we utilized ChatGPT to convert the disease names of the samples into ICD-10 codes. Furthermore, we performed manual proofreading to confirm the precision of the ICD-10 codes.

Data Analysis The CMHE-DD dataset comprises 867 distinct illnesses corresponding to 452 categories in the ICD-10 classification system. Out of these, 327 instances are drawn from the KaMed dataset, where each case can encompass one or more diseases. The remaining instances originate from the MedQA dataset, which exclusively contains cases with a single disease. On average, each disease's original description in the dataset

consists of 6.03 Chinese characters. Upon associating these descriptions with the ICD-10 codes, a specific category code is assigned as a label for each disease. The distribution of diseases in the CMHE-DD dataset is depicted in Figure 6.

3.3. Concept Explanation

Task Definition The CMHE-CE dataset serves the purpose of evaluating the model's ability to refrain from producing misleading or fictitious information while elucidating medical concepts. To precisely measure the accuracy of the content generated by LLMs, we constructed the task form in the form of a self-familiar test (Luo et al., 2023; Dhuliawala et al., 2023). This approach enables us to evaluate the model's proficiency in generating suitable interpretations of provided concepts, as well as its ability to generate corresponding concepts based on given explanations concurrently.

Figure 7 demonstrates the two-phase testing procedure of CMHE-CE. Initially, LLMs must respond to a provided question that centers on a single concept, like 'POEMS syndrome', and produce corresponding answers. Afterwards, the response generated needs refinement by substituting 'POEMS syndrome' with predefined mask labels [1]. These adjusted answers are then employed as questions in the subsequent phase. In the second phase, LLMs are given the newly created questions and answer options as input. Their objective is to recognize and choose the specific concept 'POEMS syndrome' as the correct answer.

In the test described above, regardless of when the model produces text with inaccurate information, it will not pass the self-familiar test. This test is useful for assessing the model's capacity to generate accurate medical content.

Dataset Creation As illustrated in Part (C) of Figure 2, the CMHE-CE dataset is constructed using the Medical Subject Headings (MeSH) (Lipscomb, 2000). MeSH is a hierarchically organized terminology used to index biomedical information in databases such as PubMed [2] and other resources from NLM [3]. MeSH offers two key advantages for our purposes. First, its comprehensive catalog encompasses a wide range of medical concepts, making it highly suitable for assessing the breadth of medical knowledge in LLMs. Second, MeSH organizes similar medical concepts under specific subcategories, allowing us to evaluate LLMs' ability to differentiate between similar concepts based on the catalog structure.

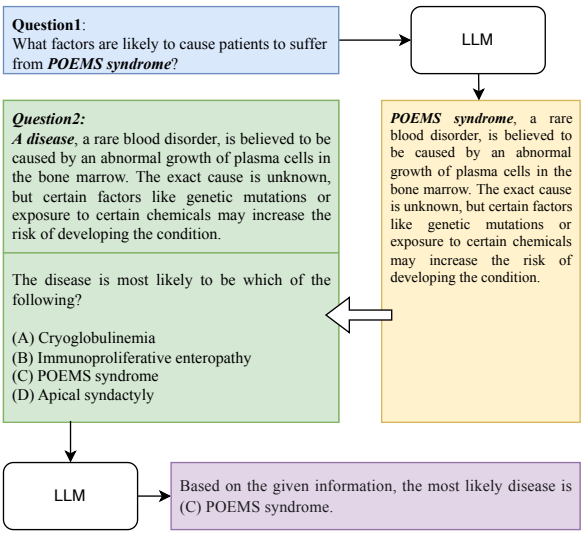

Figure 7: The overview of self-familiar test.

To construct the data set, we initially used the Chinese version of MeSH and involved three medical experts to manually review and choose terms that are highly relevant to disease diagnosis. Subsequently, we implemented specific guidelines to craft queries (referred to as Question 1 in Figure 7) for each term, adapting them to the characteristics of the term. In the case of diseases, we explore aspects such as causes, symptoms, available treatments, and screening methods. For drugs, our focus was on pharmacokinetics, indications, potential side effects, and interactions with other medications. When evaluating screening protocols, we considered both the procedure itself and the specific disease under scrutiny. Finally, we proceeded to generate the appropriate choices (corresponding to the options in Question 2 of Figure 7) for each concept. To achieve this, we chose two concepts of MeSH that closely resembled the correct answer to create challenging options. In addition, we selected a third concept that diverged significantly from the correct answer category, making it an easily eliminated option. Our methodology defines the similarity between two concepts as the inverse of the shortest path length within the MeSH hierarchical tree.

Data Analysis We obtained a total of 9,909 concepts. Among these concepts, 71. 6% are related to diseases, 17. 7% are related to medicines, and the remaining concepts are associated with medical tests.

[1] For concept categorization, three general mask labels have been defined: medicine, diseases, and medical test.

[2] https://pubmed.ncbi.nlm.nih.gov/

[3] https://www.nlm.nih.gov/

Dataset	Partition			Total
CMHE-HD	Generated 387	Tampered 613	Correct 1,000	2,000
CMHE-DD	Chat 327	Exam 1,295	- -	1,622
CMHE-CE	Medicine 1,753 x 4	Disease 7,096 x 4	Checkup 1,060 x 2	38,576

Table 1: The statistics of each sub-category in the CMHE dataset.

4. Experiments

4.1. Datasets

Our proposed CMHE benchmark comprises three individual datasets: CMHE-HD for hallucination detection, CMHE-DD for disease diagnosis, and CMHE-CE for concept explanation. We have performed calculations to determine the number of samples within each fine-grained subset of these datasets, and the resulting statistics are presented in Table1.

4.2. Baselines

Three well-known LLMs are examined to assess their performance in detecting medical hallucinations in Chinese text. All three models are capable of processing input in Chinese.

(1)**ChatGPT**[4] is a large generative language model created by OpenAI, which can generate human-like texts based on past conversations. We exploit GPT-3.5 as the backbone of ChatGPT in our experiments. (2) **Baichuan**(Yang et al., 2023) in the second version is a series of large-scale multilingual language models containing 7 billion and 13 billion parameters trained from scratch. The *Baichuan2-13B chat* is utilized in our evaluations. (3) **Qwen** (Bai et al., 2023) is a collection of language models that includes different models with different numbers of parameters. In our evaluation of the baseline models, we rely on *Qwen-14B Chat* as the foundation.

For all of LLMs used in our experiments, the hyperparameters are set as follows. The temperature is set to 0.5, the Top-P is 0.7, the Top-K is 200, and the repetition penalty is 1.1.

5. Results and Analysis

In this section, the experimental results of three baseline systems in Chinese medical consultation are evaluated from three aspects, i.e., disease diagnosis, concept understanding, and error identification with the CMHE-HD, CMHE-DD and CMHE-CE datasets, respectively.

[4]https://chat.openai.com/

Model	Generated	Tampered	Correct	All
ChatGPT	29.2	49.5	59.8	50.8
Baichuan	8.0	42.7	80.7	55.0
Qwen	2.0	25.3	**97.4**	**56.9**
Random	**50.0**	**50.0**	50.0	50.0

Table 2: Performances of the mainstream medical large language models on CMHE-HD dataset. Note that 'Generated', 'Modified', and 'Correct' denote that data partition with various generation mode. 'All' denotes the whole dataset. 'Random' denotes that the results are generated randomly and accuracy is used as the metrics.

5.1. Performance on Hallucination Detection

We used the CMHE-HD dataset to evaluate how well different models can identify different types of hallucinations in three distinct datasets: "Generated", "Tampered", and "Correct". The "Generated" dataset includes content that is either nonexistent or irrational, whereas the "Tampered" dataset features examples of contextual inconsistencies. In contrast, the "Correct" dataset acts as a control group with no hallucinations. The results of our experiments are presented in Table 2.

In terms of both "Generated" and "Tampered" data, ChatGPT exhibits superior performance compared to the other two models, indicating its proficiency in detecting various types of hallucinations. Particularly in the Generated data, ChatGPT outperforms the other models by a significant margin, demonstrating its unmatched ability to identify knowledgeable hallucinations. Among the three models, Qwen achieves the highest performance on the "Correct" data, followed by Baichuan in second place, and ChatGPT ranks third. All three models outperform the random model by a considerable margin. By comparing the performances of the "Generated", "Tampered", and "Correct" data, we can speculate that the Qwen consistently rejects hallucinations, while ChatGPT tends to try and identify them even when they don't exist.

Additionally, all models perform worse on "Generated" data compared to "Tampered" data. This indicates that detecting knowledgeable hallucinations is more difficult than identifying context inconsistencies. LLMs often miss hallucinations caused by inconsistent contexts.

5.2. Performance on Disease Diagnosis

As shown in Table 3, the labels "Diagnose-chat" and "Diagnose-exam" indicate the origin of the data from different sources. The "chat" data consists of dialogues from real-life scenarios, which may contain excessive information that could potentially distract models. On the other hand, the "Exam"

Diagnose-chat									
Model	**Level=0**			**Level=1**			**Level=2**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ChatGPT	**51.7**	**50.6**	**51.1**	**27.7**	**26.9**	**27.3**	**16.7**	**16.2**	**16.5**
Baichuan	49.8	45.0	43.7	26.4	23.5	24.8	14.5	12.9	13.6
Qwen	48.4	46.7	48.4	27.2	25.2	26.2	14.6	13.4	14.0
Diagnose-exam									
Model	**Level=0**			**Level=1**			**Level=2**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ChatGPT	64.0	63.7	63.8	49.4	49.2	49.3	38.2	37.9	38.0
Baichuan	64.8	62.5	63.6	50.1	48.4	49.3	36.2	34.9	35.5
Qwen	**71.1**	**68.6**	**69.8**	**58.0**	**55.9**	**56.9**	**45.2**	**43.3**	**44.2**
Diagnose-all									
Model	**Level=0**			**Level=1**			**Level=2**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ChatGPT	61.5	61.0	61.3	44.9	44.6	44.8	33.7	33.3	33.5
Baichuan	61.9	58.9	60.4	45.5	43.2	44.3	31.9	30.2	31.0
Qwen	**66.9**	**64.1**	**65.5**	**51.8**	**49.5**	**50.6**	**38.9**	**36.9**	**37.9**

Table 3: Performances of the mainstream medical large language models on CMHE-DD dataset. Note that "Diagnose-chat" and "Diagnose-exam" denote the data source, and "Level=0,1,2" denotes the category level in the ICD-10. **Bold** number denotes the best result of three models.

data is derived from a standardized Chinese medical exam, providing concise descriptions without any additional information.

In the "Diagnose-chat" setting, ChatGPT demonstrates the highest level of performance, indicating its superior resilience to disturbances compared to the other models. Conversely, in the "Diagnose-exam" setting, the Qwen model outperforms the others, showcasing its enhanced capability in disease diagnosis when there is no interference from extraneous information.

When comparing the overall performances of the "Diagnose-chat" and "Diagnose-exam" settings, a noticeable disparity becomes evident. The performances in the "Chat" setting are significantly inferior to those in the "Exam" setting. Furthermore, our analysis indicates that the F1 score experiences a substantial decline specifically in level-1 classification within the "Chat" data. This phenomenon implies that the model has incorrectly identified the organ in which the disease is present, which is usually unacceptable. Consequently, effectively mitigating the impact of redundant information is a major obstacle in improving the performance of language models for disease diagnosis.

5.3. Performance on Concept Explanation

The CMHE-CE dataset covers three main areas: “Medicine”, “Disease”, and “Medical Test”. Our assessment focuses on how various models perform within each of these areas. The results are presented in Table 4. ChatGPT shows better results in both the "Medicine" and "Disease" domains, whereas the Baichuan model excels in the "Medical Test" category. These results suggest that there is a difference in knowledge distribution across LLMs. This variation may be due to disparities in the sources of training data and the size of the parameters in each LLM.

Model	**Medicine**	**Disease**	**Test**	**All**
ChatGPT	**66.7**	**65.4**	78.8	**67.1**
Baichuan	57.4	57.3	**81.7**	59.9
Qwen	60.1	61.7	74.0	62.8

Table 4: Performances of the mainstream medical large language models on CMHE-CE dataset. Note that “Medicine”, “Disease” and “Test” denote the data partition based on concept categories, and “All” denotes the whole dataset and accuracy is used as the metrics.

In addition, the concepts of "medicine" and "disease" present a diverse range of questions and can be challenging to differentiate. This is due to the common occurrence of multiple symptoms or drugs associated with a single disease and the suitability of a drug for patients with various diseases. The evaluation method for assessing concept understanding encounters difficulties in mapping these one-to-many relationships. In contrast, the concept of "Medical test" typically aligns with specific workflows and diseases, making it relatively easier for models to distinguish between different "Test" concepts. The superior performance of ChatGPT over other models across the entire dataset further supports the fact that ChatGPT’s knowledge reservoir greatly surpasses that of other models, owing to its extensive parameter size.

5.4. Findings and Directions

Based on the aforementioned experimental results and analysis, several significant findings have been identified on the evaluation of hallucinations. **Finding 1**: LLMs typically detect hallucinations by evaluating the logical coherence of the sentence context. However, their ability to identify false information, such as manipulated drug names and treatment plans, is limited. **Finding 2**: LLMs exhibit strong performance in environments devoid of interfering information. However, their performance tends to deteriorate in noisy environments, such as when patients provide a substantial amount of invalid information. **Finding 3**: LLMs that possess a deeper understanding of medical concepts exhibit improved performance in noisy environments.

Due to the poor performance of LLMs in noisy environments, exploring ways to enhance the robustness of LLMs during inference, especially when LLMs are aware of their errors but tend to perpetuate their previous falsehoods, will be an intriguing avenue for future investigation.

6. Conclusion

Hallucination evaluation is a major challenge for LLM's application in the Chinese medical domain, especially snowballing hallucination problems. Most existing studies rely on automatic indicators and lack an intuitive evaluation of hallucinations. To appreciate LLM's ability on hallucination perception sufficiently, we need to decompose this problem into several aspects, e.g., identifying medical hallucinations, making accurate diagnoses in noisy conditions, providing plausible explanations, etc. CMHE specifically targets the assessment of comprehensive hallucinations in Chinese medical chat scenarios. This involves evaluating various processes that could potentially lead to hallucinations, such as errors in identification, reasoning, diagnosis, concept explanation, and exploitation. Our findings demonstrate that LLMs excel at detecting inconsistencies but struggle in noisy environments with redundant information. However, when LLMs possess a solid understanding of concepts, their performance can be greatly enhanced. In conclusion, researchers can easily locate the type of hallucinations and identify the lack of understanding in LLM through failures of CMHE, and CMHE can serve as a valuable benchmark to assess hallucinations in Chinese medical contexts.

7. Ethics and Limitations

The CMHE benchmark is constructed using a widely used public corpus. All information in the corpus has been anonymized and excludes any personal data, and it is publicly accessible online. Additionally, during the annotation process, we require annotators to manually filter and screen sensitive information to ensure the protection of personal privacy. While W2W shows great potential, it is essential to assess its ethical and societal implications. Our task definition and research models rely on pre-trained language models and public datasets, which may contain hidden biases leading to fairness issues within the algorithms. By acknowledging and actively addressing these implications, our aim is to raise awareness among practitioners if the model is deployed as a language-learning agent in the future.

8. Acknowledgement

Our work is supported by the National Key Research and Development Program of China (Project Number: 2020AAA0109400). We kindly appreciate all the researchers who provide valuable insights, discussions, and comments on this work.

9. Bibliographical References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

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

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

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

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P-W Lo, Bo Xiao, et al. 2023. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

John Schulman. 2023. Reinforcement learning from human feedback: Progress and challenges.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. Huatuogpt, towards taming language models to be a doctor. *arXiv preprint arXiv:2305.15075*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's

song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

10. Language Resource References

Junyi Li and Xiaoxue Cheng and Wayne Xin Zhao and Jian-Yun Nie and Ji-Rong Wen. 2023. *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. arXiv preprint. PID https://github.com/RUCAIBox/HaluEval.

Lin, Stephanie and Hilton, Jacob and Evans, Owain. 2021. *Truthfulqa: Measuring how models mimic human falsehoods*. arXiv preprint. PID https://github.com/sylinrl/TruthfulQA.

Lipscomb, Carolyn E. 2000. *Medical subject headings (MeSH)*. Medical Library Association. PID https://www.nlm.nih.gov/mesh/meshhome.html.

Muhlgay, Dor and Ram, Ori and Magar, Inbal and Levine, Yoav and Ratner, Nir and Belinkov, Yonatan and Abend, Omri and Leyton-Brown, Kevin and Shashua, Amnon and Shoham, Yoav. 2023. *Generating benchmarks for factuality evaluation of language models*. arXiv preprint. PID https://github.com/AI21Labs/factor.

Pal, Ankit and Umapathi, Logesh Kumar and Sankarasubbu, Malaikannan. 2022. *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. PMLR. PID https://github.com/medmcqa/medmcqa.

Ankit Pal and Logesh Kumar Umapathi and Malaikannan Sankarasubbu. 2023. *Med-HALT: Medical Domain Hallucination Test for Large Language Models*. arXiv preprint. PID https://medhalt.github.io/.

Toyhom. 2023. *Chinese medical dialogue data*. arXiv preprint. PID https://github.com/Toyhom/Chinese-medical-dialogue-data.

Vilares, David and Gómez-Rodríguez, Carlos. 2019. *HEAD-QA: A Healthcare Dataset for Complex Reasoning*. Association for Computational Linguistics. PID https://aghie.github.io/head-qa/.

Wang, Xidong and Chen, Guiming Hardy and Song, Dingjie and Zhang, Zhiyi and Chen, Zhihong and Xiao, Qingying and Jiang, Feng and Li, Jianquan and Wan, Xiang and Wang, Benyou and others. 2023. *CMB: A Comprehensive Medical Benchmark in Chinese*. arXiv preprint. PID https://github.com/FreedomIntelligence/CMB.

S. Zhang and X. Zhang and H. Wang and L. Guo and S. Liu. 2018. *Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection*. arXiv preprint. PID https://github.com/zhangsheng93/cMedQA2.