MultifacetEval: Multifaceted Evaluation to Probe LLMs in Mastering Medical Knowledge

Yuxuan Zhou, **Xien Liu***, **Chen Ning** and **Ji Wu**
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
{zhou-yx21, nc22}@mails.tsinghua.edu.cn, {xeliu, wuji_ee}@mail.tsinghua.edu.cn

Abstract

Large language models (LLMs) have excelled across domains, also delivering notable performance on the medical evaluation benchmarks, such as MedQA. However, there still exists a significant gap between the reported performance and the practical effectiveness in real-world medical scenarios. In this paper, we aim to explore the causes of this gap by employing a multifaceted examination schema to systematically probe the actual mastery of medical knowledge by current LLMs. Specifically, we develop a novel evaluation framework **MultifacetEval** to examine the degree and coverage of LLMs in encoding and mastering medical knowledge at multiple facets (comparison, rectification, discrimination, and verification) concurrently. Based on the MultifacetEval framework, we construct two multifaceted evaluation datasets: MultiDiseK (by producing questions from a clinical disease knowledge base) and MultiMedQA (by rephrasing each question from a medical benchmark MedQA into multifaceted questions). The experimental results on these multifaceted datasets demonstrate that the extent of current LLMs in mastering medical knowledge is far below their performance on existing medical benchmarks, suggesting that they lack depth, precision, and comprehensiveness in mastering medical knowledge. Consequently, current LLMs are not yet ready for application in real-world medical tasks. The codes and datasets are available at https://github.com/THUMLP/MultifacetEval.

1 Introduction

The rapid advancement of large language model (LLM) technology has achieved great success in various domains [Romera-Paredes *et al.*, 2023; Madani *et al.*, 2023; Boiko *et al.*, 2023]. Current LLMs encode extensive knowledge through pretraining on massive unlabeled data. Some are further finetuned on supervised datasets to be adapted to specific downstream tasks. Recently, famous general LLMs like GPT-4 [OpenAI, 2023] and Gemini-pro [Team *et al.*, 2023], as well as medical-domain-specific LLMs such as Med-PaLM [Singhal *et al.*, 2023a], are reported to have encoded vast medical knowledge and achieved significant performance on several medical benchmarks, surpassing previous state-of-the-art models by a considerable margin [Kung *et al.*, 2023; Nori *et al.*, 2023a; Nori *et al.*, 2023b]. Nevertheless, despite their impressive performance on existing benchmarks, these LLMs still face challenges in addressing real-world medical problems [Thirunavukarasu *et al.*, 2023; Clusmann *et al.*, 2023; Wornow *et al.*, 2023]. This leads to a significant gap between evaluation results and practical performance in the medical domain. Therefore, in this paper, we aim to study the underlying causes of this gap by systematically investigating the **depth** of medical knowledge mastery in current LLMs.

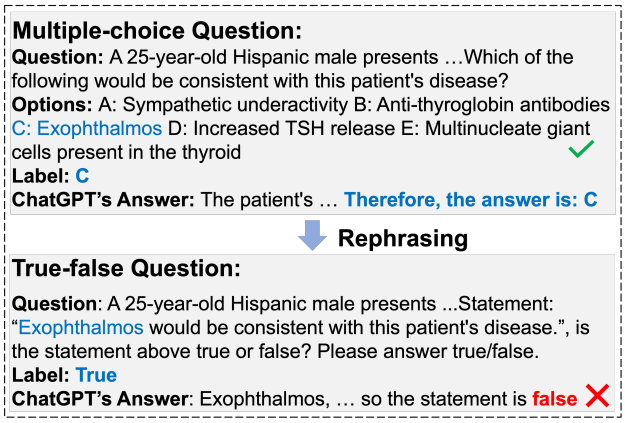


Figure 1: GPT-3.5-turbo responding to medical exam problems assessing the same knowledge point but in different formats.

Several medical benchmarks have been proposed to measure LLMs' capacities in the medical domain. Most current medical benchmarks assess LLMs by medical question-answering tasks [Jin *et al.*, 2021; Pal *et al.*, 2022; Hendrycks *et al.*, 2020; Jin *et al.*, 2019; Ben Abacha *et al.*, 2017; Ben Abacha *et al.*, 2019; Singhal *et al.*, 2023a]. Other benchmarks also evaluate LLMs in the forms of medical dialogue [Zeng *et al.*, 2020] or other traditional NLP tasks (e.g., relation extraction, NER) based on medical corpora [Zhang *et al.*,

*Corresponding author

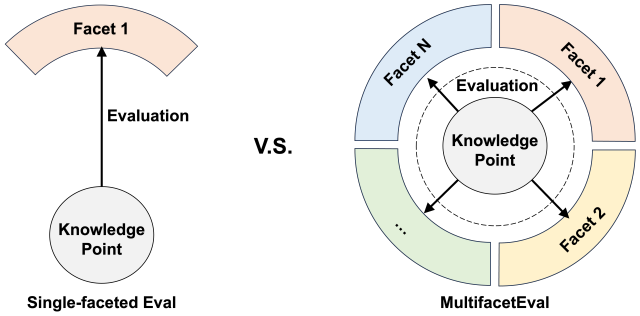


Figure 2: Principle of the proposed multifaceted evaluation.

2022]. Nevertheless, most existing medical benchmarks rely on a specific question type (e.g., multiple-choice questions) to evaluate LLMs. Therefore, they may overestimate the performance of current LLMs, as certain LLMs may have been finetuned for specific question types. Consequently, their performance on specific question types would significantly surpass those on other questions. Even if some benchmarks evaluate LLMs' medical capabilities from various facets, each facet is evaluated based on distinct sets of knowledge points. Therefore, the outcomes of these benchmarks still cannot reflect LLMs' mastery on the same knowledge points across diverse facets. Meanwhile, we found it necessary to conduct **multifaceted** evaluation on the same knowledge point. Figure 1 illustrates GPT-3.5-turbo's response to two medical exam problems assessing the same knowledge point but in different question types. The multiple-choice question is extracted from the United States Medical Licensing Examination (USMLE). In contrast, the true-false question is adapted from the original question by substituting the phrase "Which of the following" with the correct option, evaluating LLMs' ability to verify statements based on corresponding medical knowledge. Although GPT-3.5-turbo successfully chooses the symptom of the patient's disease, it judges the statement that is consistent with MCQ's answer as false, conflicting with its previous prediction. This highlights the importance of conducting multifaceted evaluations on identical medical knowledge points for a systematically analysis of LLMs' knowledge mastery.

In contrast to existing evaluation benchmarks, *current education systems generally utilize various assessment methods, including assignments, quizzes, projects, and exams, to evaluate students' comprehensive mastery of the same knowledge point from multiple facets*. Inspired by this, we propose a novel multifaceted evaluation approach **MultifacetEval** to evaluate the actual medical knowledge mastery of current LLMs from multiple facets. Figure 2 illustrates the principle of this approach. Specifically, we generate a series of questions for each knowledge point of interest with various question types. These questions emphasize evaluating this knowledge point from different facets, including comparison, discrimination, verification, and rectification capabilities. Therefore, the proposed approach would provide a more comprehensive evaluation of LLMs' medical knowledge mastery compared with conventional medical benchmarks that rely on a single evaluation facet. The proposed approach also possesses strong versatility, as it can generate multifaceted questions by directly crafting them based on knowledge points in medical knowledge bases or rephrasing questions on existing medical evaluation benchmarks.

To validate the effectiveness of the proposed multifaceted evaluation method, we apply the proposed method to construct two new evaluation datasets based on a medical knowledge database and a medical benchmark MedQA [Jin *et al.*, 2021], respectively. A total of 13 well-known general and medical LLMs are evaluated on these datasets. The experimental results indicate that current LLMs lack a comprehensive, precise, and in-depth mastery of medical knowledge despite their considerable performance on existing medical benchmarks. Moreover, the results demonstrate that current LLMs possess excellent comparison capability, while they have not well mastered other capabilities such as discrimination, verification, and rectification in the medical domain. Our contributions can be summarized as follows:

- We propose a novel multifaceted evaluation schema (MultifacetEval) to evaluate LLMs' medical knowledge mastery on the same knowledge point from various facets instead of a single facet in existing benchmarks. The proposed method can more accurately evaluate the LLMs' mastery of medical knowledge.
- Based on the proposed method, we generate two novel multifaceted datasets, **MultiDiseK** and **MultiMedQA**, based on a medical knowledge base and a well-known medical benchmark MedQA, respectively. The performance of these two datasets can more comprehensively reflect LLMs' mastery of medical knowledge.
- The experimental results reveal that the genuine extent of medical knowledge mastery in current LLMs is significantly lower than that evaluated by existing medical benchmarks. Additionally, we observe substantial variations in LLMs' performance across different facets.

2 Related Work

Large Language Models on Medical Tasks Recently, some famous LLMs are reported to encode medical knowledge and achieve considerable performance on existing medical benchmarks. General LLMs such as Flan-PaLM and GPT-4 are reported to achieve state-of-the-art performance on multiple datasets [Singhal *et al.*, 2023a; Nori *et al.*, 2023b]. For example, they achieve accuracies of 67.6 and 90.2 on a medical exam benchmark MedQA [Jin *et al.*, 2021], largely surpassing the prior SOTA models. Several LLMs specially pretrained or finetuned on the medical corpora, such as Med-PaLM, Med-PaLM2 [Singhal *et al.*, 2023b], ClinicalCamel [Toma *et al.*, 2023], and Med42 [Christophe *et al.*, 2023], are also proposed to address problems in the medical domain and achieve high performance on various medical benchmarks. However, these models cannot tackle problems in real medical scenarios. Our study aims to investigate the gap between the high evaluation performance and the limited practical effectiveness of existing LLMs in the medical domain.

Medical Evaluation Benchmarks Current medical evaluation benchmarks can be classified into three classes: (1)

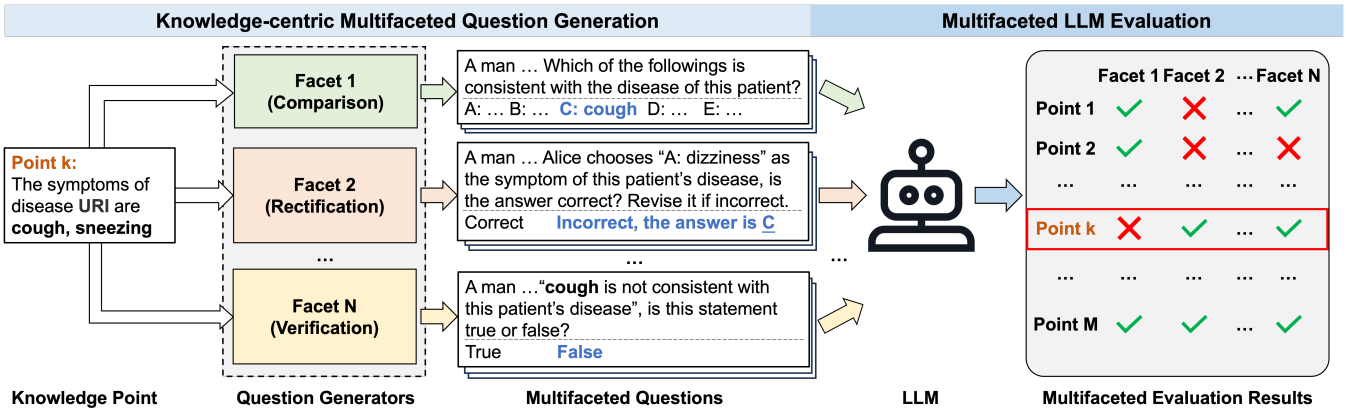


Figure 3: Framework of the proposed multifaceted evaluation approach that evaluates LLMs' medical knowledge mastery from various facets.

Question-answering datasets with problems collected from different sources, including medical exams [Jin *et al.*, 2021; Pal *et al.*, 2022; Hendrycks *et al.*, 2020], scientific literature [Jin *et al.*, 2019], and consumer health questions [Ben Abacha *et al.*, 2017; Ben Abacha *et al.*, 2019; Singhal *et al.*, 2023a]; (2) medical dialogue datasets [Zeng *et al.*, 2020; Yang *et al.*, 2020]; (3) datasets [Peng *et al.*, 2019; Zhang *et al.*, 2022] involving conventional NLP tasks (NER, relation extraction, NLI) on medical corpora. Some of these datasets assess LLMs from a single facet. Others evaluate LLMs with multiple tasks, while the tasks are constructed on different groups of knowledge points. In this paper, we design a new evaluation method to evaluate LLMs' mastery of the same knowledge point from multiple facets.

3 Knowledge-Centric Multifaceted Evaluation

3.1 Multifaceted Evaluation Schema

The proposed multifaceted evaluation approach is motivated by the current education systems, where various assessment methods, including assignments, projects, and exams, are employed to comprehensively evaluate whether students have truly mastered a particular knowledge point. Given a knowledge point k and a series of N evaluation facets $\mathbf{f} = [f^1, f^2, \cdots, f^N]^{\mathbf{T}}$, the performance of an LLM (M) evaluated through the proposed multifaceted evaluation is:

$$\mathbf{f}_{\mathrm{k}}(\mathrm{M}) = [f^1_{\mathrm{k}}(\mathrm{M}), f^2_{\mathrm{k}}(\mathrm{M}), \cdots, f^N_{\mathrm{k}}(\mathrm{M})]^{\mathbf{T}} \quad (1)$$

Where f^i_{k} denotes the specific questions designed to emphasize the evaluation of the i^{th} facet related to the knowledge point k, and $f^i_{\mathrm{k}}(\mathrm{M}) \in \{0, 1\}$ is the evaluation outcome: $f^i_{\mathrm{k}}(\mathrm{M}) = 1$ if all questions in f^i_{k} are answered correctly, and 0 otherwise. Compared with single-faceted evaluation, the proposed multifaceted evaluation method conveys more comprehensive information about the mastery of a specific knowledge point. The proposed multifaceted evaluation schema demonstrates strong **transferability**: it can be applied in other domains by adjusting evaluation facets and question generation strategies.

3.2 Facets of Medical Knowledge Mastery

We employ the proposed evaluation schema to systematically probe current LLMs in mastering medical knowledge. Considering the characteristics of medical scenarios, we set $N = 4$ and design a total of four evaluation facets of capabilities that are essential for solving real medical problems:

Comparison (f^1): The ability to compare different medical entities/events and choose the most suitable one that meets some criteria. It is crucial for medical applications such as diagnosis and drug recommendation.

Rectification (f^2): The capability to identify errors in the medical process (treatment, diagnosis) and offer corresponding corrections. Rectification plays an important role in medical scenarios such as computer aided diagnosis.

Discrimination (f^3): The capacity to recognize and differentiate between medical concepts accurately. Discrimination of medical concepts is the bedrock of medical applications such as clinical decision support and personalized medicine.

Verification (f^4): The ability to determine the veracity of a statement based on the acquired knowledge. Such capability is highly demanded in the quality assessment of electronic health records and laboratory results.

3.3 Multifaceted Medical Evaluation Framework

Built on the facets discussed above, we design a multifaceted medical evaluation framework to comprehensively evaluate mastery of medical knowledge by LLMs from these evaluation facets. Figure 3 illustrates an overview of the proposed multifaceted evaluation framework. Given a set of medical knowledge points $\mathcal{K}$, the framework evaluates LLMs' mastery of medical knowledge through two steps:

Multifaceted Question Generation In the first step, we generate multiple questions from diverse evaluation facets for each knowledge point in the set: $\mathrm{k} \rightarrow \{f^i_{\mathrm{k}} | 1 \leq i \leq N\}$, where $\mathrm{k} \in \mathcal{K}$. Specifically, we design four question types, including multiple-choice questions, revision questions, multiple-answer questions, and true-false questions, to emphasize the evaluation of the comparison, rectification, discrimination, and verification facets, respectively:

(1) Multiple-Choice Questions: We maintain the multiple-choice questions (MCQ) applied in existing benchmarks

Question:
A 25-year-old Hispanic male presents …Which of the following would be consistent with this patient's disease?
A: Sympathetic underactivity B: Anti-thyroglobin antibodies C: Exophthalmos D: Increased TSH release E: Multinucleate giant cells present in the thyroid
Alice's answer: A. Please determine whether her answer is correct, and if it's incorrect, provide the correct answer.
Answer: Correct / Incorrect, the correct answer is ____

Figure 4: Example of the proposed revision question.

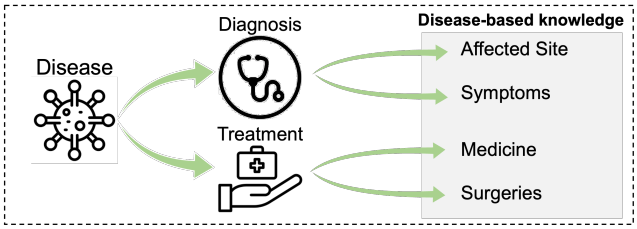


Figure 5: Aspects of disease-related knowledge in DiseK.

to emphasize the evaluation of the comparison facet. A multiple-choice question comprises a question and multiple options (typically 4). To answer multiple-choice questions accurately, participants must compare the given options and select the most suitable choice that fits the question.

(2) Revision Questions: We design a new question type named "revision" question (RQ) to focus on evaluating the rectification capabilities of LLMs. Figure 4 illustrates an example of this question type. A revision question comprises a multiple-choice question and a **provided option** (not necessarily correct) to this question. Participants are asked to recheck the correctness of the given option based on the question, and revise the answer appropriately if needed.

(3) Multiple-Answer Questions: We consider leveraging multiple-answer questions (MAQ) to highlight the evaluation of the discrimination capability. In contrast to MCQs, a multiple-answer question consists of several options, with one or more aligning with the given question. Effectively answering MAQs requires a comprehensive and precise mastery of discriminative knowledge for all options, as MAQ answers cannot be determined through option-wise comparison.

(4)True-false Questions: We utilize true-false questions (TFQ) to emphasize the assessment of the verification facet. A true-false question generally presents a statement that can be verified based on the corresponding medical knowledge and information provided in the statement. True-false questions do not include options that may provide clues, requiring ones to mastery medical knowledge accurately.

Multifaceted LLM Evaluation In the next step, we evaluate the LLM M with the generated multifaceted questions to obtain comprehensive evaluation results on each knowledge points: $f_{\mathrm{k}}^i \rightarrow f_{\mathrm{k}}^i(\mathrm{M})$ where $k \in \mathcal{K}$ and $1 \leq i \leq N$. Finally, the proposed evaluation framework produces comprehensive evaluation outcomes for M across all the knowledge points: $\{\mathbf{f}_{\mathrm{k}}(\mathrm{M}) | \mathrm{k} \in \mathcal{K}\}$. To reflect the LLM's comprehensive mastery of individual knowledge points, we define a knowledge point k is **mastered** by M under facets $\{f^1, f^2 ..., f^N\}$, if the function

$$r_{\mathrm{k}}(\mathrm{M}) = \prod_{i=1}^{N} f_{\mathrm{k}}^i(\mathrm{M}) \tag{2}$$

equals 1. Here, $r_{\mathrm{k}}(\mathrm{M}) = 1$ only when $f_{\mathrm{k}}^i(\mathrm{M}) = 1$ holds for $1 \leq i \leq N$, indicating accurate answers to all questions from these facets. The overall performance can therefore be represented as the *proportion of mastered knowledge points*:

$$p(\mathrm{M}) = \frac{1}{|\mathcal{K}|} \sum_{\mathrm{k} \in \mathcal{K}} r_{\mathrm{k}}(\mathrm{M}) \tag{3}$$

4 Experiments

4.1 Experiment Setup

MultiDiseK Dataset Generation Medical knowledge bases explicitly contain knowledge points that can be directly utilized in the proposed method. In this paper, we introduce a disease-centric knowledge base (**DiseK**) and construct a multifaceted evaluation dataset (**MultiDiseK**) from it. DiseK is annotated by 20 medical experts for about 3 months. It consists of 1,000 common diseases, accompanied by 4 fundamental aspects of medical knowledge (illustrated in Figure 5). These aspects are closely associated to the clinical decision-making process, involving diagnosis and treatment. Therefore, LLMs must acquire these aspects of medical knowledge to be applicable in clinical decision support systems (CDSS) [Wu *et al.*, 2018; Liang *et al.*, 2019].

The MultiDiseK dataset is constructed based on medical knowledge points in DiseK by using carefully crafted question templates. For the comparison facet, an MCQ (**4-option**) is generated for each aspect of disease knowledge, where options are formed by selecting attributes that either belong or do not belong to the specified disease. Revision questions are generated by rephrasing each MCQ into two questions, providing either the correct choice or a randomly selected incorrect choice. Multiple-answer questions are generated similarly to MCQs, but with selecting 1-3 attributes as correct options. For true-false questions, we randomly choose an attribute with 50% probability associated with the disease and 50% not associated. Participants are then asked to determine whether the given attribute is associated with the disease. For all questions crafted above, we also generate a corresponding negated version by incorporating negation words into the question and modifying the answers correspondingly. This is done to further assess the depth of knowledge mastery by LLMs. Finally, the constructed dataset encompasses a total of 3,167 disease-related knowledge points (some diseases may not have corresponding medications or surgeries), including 6,334 MCQs, 12,668 RQs, 6,334 MAQs, and 6,334 TFQs. More details of this dataset (e.g., question templates, dataset statistics) are presented in Appendix A and B.

MultiMedQA Dataset Generation To make our proposed multifaceted evaluation approach comparable with existing benchmarks, we further construct another dataset **MultiMedQA** based on a medical benchmark **MedQA** [Jin *et al.*, 2021], since several LLMs have achieved notable performance on this benchmark. MedQA is a medical exam dataset that contains **5-option** multiple-choice questions from the professional medical board exams of different sources. The

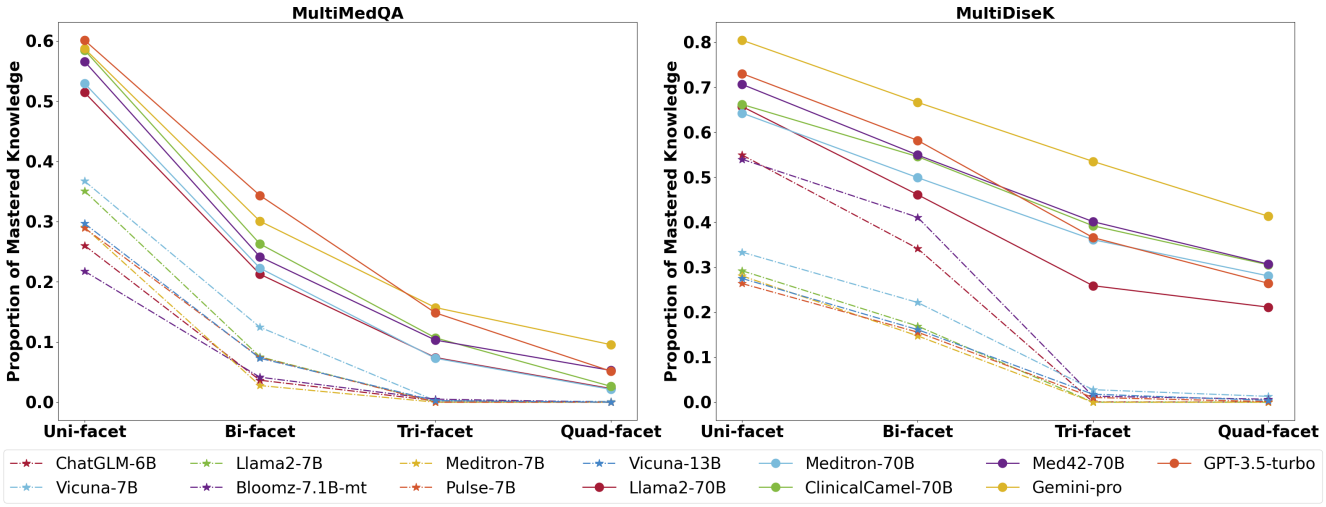


Figure 6: Proportion of mastered knowledge points ($p(\mathrm{M})$) evaluated by single-faceted and multi-faceted methods on two datasets. Dash dotted lines refer to LLMs with sizes under 70B, while solid lines denote LLMs larger than 70B. Evaluated facets are added following the sequence: comparison, verification, rectification, and discrimination.

question in MedQA typically consists of a patient's medical consultation record followed by a question related to the patient's situation (e.g., diagnosis, the next step in management, findings of diagnostic tests). Employing the multifaceted evaluation schema, we rephrase each MedQA question into various types, conducting a multifaceted evaluation of the medical knowledge points embedded in MedQA. To do so, we first selected 800 questions suitable for the multifaceted adaptation from the US exam part (1,273 questions) by regular expressions. After that, we rephrase them into multifaceted questions with heuristic rules. For revision-type questions, we generate them using a method similar to that applied in MultiDiseK construction. For multiple-answer questions, given the challenge of explicitly identifying the knowledge points in the original question, we adopt a solution by retrieving synonyms for the correct option and randomly replacing 0-3 incorrect options with these synonyms to generate a new question. Each question is further paired with a negated version by introducing a negation word in the question. True-false questions were generated by substituting the interrogative word/phrase (e.g., "which of the following") with either the correct or incorrect option selected from the remaining four options, and negated versions were also created using the same method. The resulting MultiMedQA dataset includes 800 MCQs, 1,600 RQs, 1,600 MAQs, and 3,200 TFQs. More details of MultiMedQA are provided in Appendix C.

Evaluation Setting We evaluate LLMs by five-shot learning on the proposed datasets. We report the performance of LLMs under two settings: (1) answer-only [Brown *et al.*, 2020]: prompting LLMs with only question-answer pairs; (2) Chain-of-Thought with Self-consistency (CoT+SC) [Wang *et al.*, 2022]: prompting LLMs multiple times with question-answer pairs and the chain-of-thoughts, aggregating the results by majority vote to obtain the final answer. For the latter setting, we generate CoTs following the method proposed in [Nori *et al.*, 2023b] and ask LLMs each question 5 times in our implementation. We only apply the answer-only setting for experiments in MultiDiseK since questions in MultiDiseK do not require sophisticated reasoning in medical cases. We use carefully designed regular expressions to extract answers and observe that they can retrieve answers successfully in most cases. More details are provided in Appendix D.

Metrics We employ **proportion of mastered knowledge points** ($p(\mathrm{M})$ in Sec.3.3) to measure the overall performance. For each facet, we employ accuracy ($\frac{\#\text{correctly answered questions}}{\#\text{all of the questions}}$) as the fine-grained metric. For MAQs, correct predictions require an exact match with the ground truth answers. Regarding RQs, correctness is determined when both the veracity of the chosen option and the original question's answer are accurately predicted. We observe that some LLMs "fortunately" achieve high accuracies on RQs by always staying consistent with the provided option since half of the RQs provide the correct options. Therefore, we revise the calculation of revision questions' performance to reduce this bias: $\mathrm{acc} = \frac{1}{N_o}\,\mathrm{acc}_T + \frac{N_o-1}{N_o}\,\mathrm{acc}_F$, where N_o is the number of options, acc_T and acc_F are the accuracies of RQs that provide correct and incorrect options, respectively. The accuracy calculated above is proven to reduce the impact of this bias, and we provide the corresponding proof in Appendix E.

Baseline Models We evaluate a total of 13 LLMs with varying sizes in this paper: (1) general LLMs: ChatGLM (6B) [Du *et al.*, 2022], Llama2 (7B,70B) [Touvron *et al.*, 2023], Vicuna (7B,13B) [Zheng *et al.*, 2023], Bloomz-mt (7.1B) [Muennighoff *et al.*, 2023], GPT-3.5-turbo [Ouyang *et al.*, 2022] and Gemini-pro [Team *et al.*, 2023]; (2) medical LLMs: Pulse (7B) [Xiaofan Zhang, 2023], Meditron (7B,70B) [Chen *et al.*, 2023], ClinicalCamel (70B) [Toma *et al.*, 2023], and Med42 (70B) [Christophe *et al.*, 2023]. We have not evaluated GPT-4 [OpenAI, 2023] and MedPaLM [Singhal *et al.*, 2023a], since GPT-4 is too expensive and MedPaLM is not publicly available yet.

Model	Comp.	Rect.	Disc.	Veri.	Average
Random	20.0	20.0	3.2	50.0	23.3
ChatGLM-6B	27.7	20.3	5.7	50.6	26.1
Vicuna-7B	21.0	17.7	2.1	49.4	22.6
Llama2-7B	20.8	23.0	0.1	49.6	23.4
Bloomz-7.1B-mt	25.4	11.9	5.5	50.1	23.2
Meditron-7B	20.6	18.8	0.0	48.9	22.1
Pulse-7B	19.9	14.9	0.7	49.2	21.2
Vicuna-13B	20.1	17.4	0.6	51.7	22.4
Llama2-70B	41.8	30.7	10.8	54.7	34.5
Meditron-70B	47.2	28.3	5.1	50.8	32.8
ClinicalCamel-70B	23.9	24.9	6.4	50.8	26.5
Med42-70B	**59.0**	44.8	**26.2**	57.5	**46.9**
Gemini-pro	41.0	37.2	12.5	**59.2**	37.5
GPT-3.5-turbo	45.5	**48.6**	12.5	58.1	41.2

(a) Results in the setting of Answer-only.

Model	Comp.	Rect.	Disc.	Veri.	Average
Random	20.0	20.0	3.2	50.0	23.3
ChatGLM-6B	26.0	17.2	6.8	49.1	24.8
Vicuna-7B	36.7	10.8	6.2	53.0	26.7
Llama2-7B	35.1	18.1	5.6	51.2	27.5
Bloomz-7.1B-mt	21.8	12.5	5.5	50.9	22.7
Meditron-7B	29.1	13.3	4.8	50.2	24.3
Pulse-7B	28.9	19.6	5.3	50.5	26.1
Vicuna-13B	29.7	15.8	5.7	52.1	25.8
Llama2-70B	50.6	35.4	10.1	58.1	38.5
Meditron-70B	53.1	33.9	10.8	56.4	38.5
ClinicalCamel-70B	58.2	37.5	12.0	61.3	42.2
Med42-70B	56.9	34.3	22.6	59.1	43.2
Gemini-pro	59.4	40.2	**34.5**	**64.2**	**49.6**
GPT-3.5-turbo	**60.4**	**50.1**	22.4	62.2	48.8

(b) Results in the setting of Chain-of-Thought Self Consistency.

Table 1: Five-shot accuracies on the **MultiMedQA** dataset across comparison (Comp.), rectification (Rect.), discrimination (Disc.), and verification (Veri.) capabilities. "Average" column denotes the macro average of accuracies across all facets.

Model	Comp.	Rect.	Disc.	Veri.	Average
Random	25.0	25.0	6.7	50.0	26.7
ChatGLM-6B	35.2	27.7	18.5	52.8	33.5
Vicuna-7B	29.5	24.9	14.0	55.1	30.9
Llama2-7B	27.8	25.5	15.7	55.2	31.0
Bloomz-7.1B-mt	34.0	21.0	17.6	53.3	31.5
Meditron-7B	27.4	25.2	12.5	50.1	28.8
Pulse-7B	26.1	22.2	2.8	52.9	26.0
Vicuna-13B	25.6	25.6	9.2	52.5	28.2
Llama2-70B	65.6	47.5	33.7	58.1	51.2
Meditron-70B	66.3	50.3	38.7	63.1	54.6
ClinicalCamel-70B	66.8	63.1	37.0	68.8	58.9
Med42-70B	72.5	57.3	37.1	64.4	57.8
Gemini-pro	**81.7**	**72.6**	**55.0**	**77.0**	**71.6**
GPT-3.5-turbo	74.1	59.1	44.8	63.7	60.4

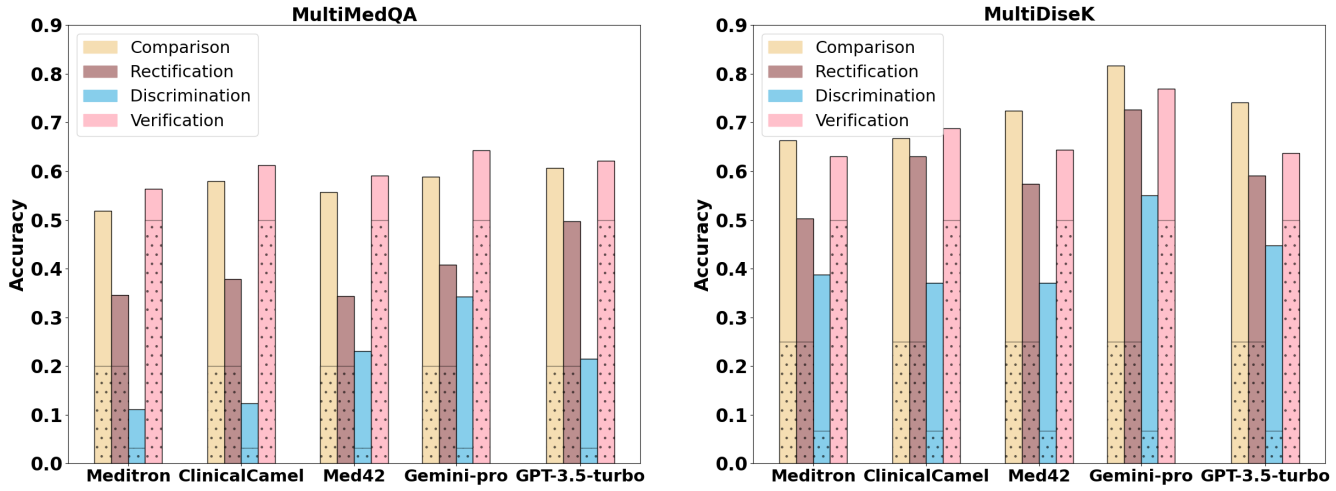
Table 2: Five-shot accuracies on the **MultiDiseK** dataset.

4.2 Results

Single-faceted vs. Multi-faceted We first compare the proposed multifaceted evaluation with the conventional single-faceted evaluation. Figure 6 illustrates the proportion of mastered knowledge points ($p(\mathrm{M})$) by LLMs on the proposed MultiMedQA and MultiDiseK datasets, evaluated using both single-faceted (comparison-type) and multifaceted methods[1]. These LLMs are reported to achieve high performance on existing medical benchmarks, including MedQA. We report the performance on the MultiMedQA achieved by the CoT+SC setting since these LLMs generally achieve higher performance in this setting. The experimental results indicate that all LLMs above 70B have effectively mastered a considerable number of knowledge points when evaluated solely from the comparison facet (i.e., **the original MedQA questions**), consistent with their reported performance on existing benchmarks. However, we observe a **sharp decline** in the proportion of mastered knowledge points across various LLMs as the number of evaluated facets increases. For example, GPT-3.5-turbo's performance evaluated by 4 facets is around 50% lower on MultiMedQA and 40% lower on MultiDiseK compared with the single-faceted results. Moreover, we observe that though several smaller LLMs (dash dotted lines) also perform notably under single-faceted evaluation, their performance nearly approaches zero when evaluated by ≥ 3 facets. In contrast, larger LLMs master more knowledge under multifaceted evaluation. We also study different sequences of adding evaluation facets in Appendix F and observe that the conclusions remain consistent. The results imply that current LLMs lack a comprehensive mastery of medical knowledge.

Comparison Across LLMs Table 1 and 2 compare LLMs performance across various datasets and settings. LLMs generally perform better on the MultiDiseK dataset since the questions do not involve analysis of specific medical cases. **Gemini-pro** achieves the highest performance on both datasets with 49.6 and 71.6 in average accuracy, respectively. GPT-3.5-turbo performs similarly to Gemini-pro on MultiMedQA (48.8) but significantly lags behind Gemini-pro on MultiDiseK (60.4). The discrepancy may be attributed to the broader coverage of disease knowledge by Gemini-pro compared with GPT-3.5-turbo, while its ability to apply medical knowledge in specific medical cases is similar to GPT-3.5-turbo. For open-source LLMs in 70B size, we find that several medical LLMs (Med42, ClinicalCamel) significantly surpass their base model LLama2-70B and achieve comparable performance compared to GPT-3.5-turbo on both datasets (46.9 for Med42 on MultiMedQA and 58.9 for ClinicalCamel on MultiDiseK). LLMs that are not larger than 13B perform only slightly better than random guessing. However, they achieve significantly higher performance in the comparison facet and perform similarly or even worse than random guessing on facets such as verification and rectification. One possible explanation is that these two facets represent higher-level

[1]Since current LLMs struggle to correctly answer both an affirmative question and its negation simultaneously, we remove negated questions in this analysis to ensure the visibility.



(a) Results on the MultiMedQA dataset.

(b) Results on the MultiDiseK dataset.

Figure 7: Multifaceted performance of LLMs across the evaluated facets on MultiMedQA and MultiDiseK. Hatched bars: random guessing performance. Solid bars above the hatched part: LLMs gain over random guessing. Meditron, ClinicalCamel, and Med42 are all 70B versions.

capabilities that can manifest only in LLMs with larger sizes. Notably, the comparison-type questions in MultiMedQA are directly sourced from the MedQA dataset. In our study, the performance of GPT-3.5-turbo on this facet (**60.4**) aligns closely with the reported performance in [Nori *et al.*, 2023b] (**60.2**), which could indicate the reliability of our findings. Comparing the Answer-only setting with the CoT+SC setting, we find that larger models significantly benefit more from CoT+SC (except Med42). The effect of CoT+SC varies across facets: for Gemini-pro, CoT+SC largely improves its performance in the comparison (+18.4) and discrimination (+22.0) facets, while it has a limited effect on verification (+5.0) and rectification facets (+3.0).

Comparison Across Multiple Facets We further compare the top-5 LLMs' performance across different capabilities facets in Figure 7. The performance on the MultiMedQA is reported under the CoT+SC setting as well. Note that the hatched bars represent the random guessing performance of the corresponding question type. The experimental results demonstrate that the evaluated LLMs typically exhibit the most significant improvement over random guessing in the **comparison facet**, followed by the rectification and discrimination facets, and lastly, the verification facet. The high performance on the comparison facet may be caused by the fact that current LLMs have seen more comparison-type questions (MCQs) in their training data to perform well on existing benchmarks. Rectification-type questions are more challenging than comparison-type questions because they require LLMs to determine the correctness of the provided answer and to revise it accurately. Discrimination-type questions also perform worse than comparison-type questions, probably because of their demand for LLMs to discern nuances between concepts instead of merely selecting the most suitable choice. Verification-type questions exhibit the lowest gain, likely due to the need for direct verification based on medical knowledge without additional information from options.

5 Conclusion and Discussion

In this paper, we propose a multifaceted evaluation approach, MultifacetEval, designed to probe the actual mastery of medical knowledge by current LLMs. Following this methodology, we construct two multifaceted evaluation datasets, MultiDiseK and MultiMedQA. The experimental results demonstrate that current LLMs' medical knowledge mastery is significantly lower than their performance on medical benchmarks suggests, indicating that the proposed MultifacetEval framework offers a more comprehensive assessment of LLMs' medical knowledge mastery. Furthermore, LLMs demonstrate significant variations in performance across different evaluation facets. These results suggest that **Current LLMs generally lack a deep, precise, and comprehensive mastery of medical knowledge**, which is the probable cause of the disparity between high performance on medical benchmarks and insufficient performance on real medical scenarios. Moreover, although some smaller LLMs are reported to achieve performance comparable to larger LLMs on several benchmarks, they achieve much lower performance on multifaceted datasets, indicating that their mastery of medical knowledge is not as comprehensive as that of larger LLMs.

The above conclusion also provides insights into the development of medical foundation models: (1) Medical foundation models need to be sufficiently large to master medical knowledge comprehensively, deeply, and precisely; (2) Their training should cover a diverse range of medical tasks rather than being restricted to specific ones, making them truly applicable in real-world scenarios.

Finally, it is worth noting that our study is only a first step in exploring the actual mastery of medical knowledge by LLMs. In the future, we plan to evaluate LLMs across additional facets relevant to real medical applications and expand the scale of knowledge points for evaluation, continuously enhancing the comprehensiveness, professionalism, and robustness of the proposed method.

Acknowledgments

The work is supported by National Key R&D Program of China (2021ZD0113402). We thank the anonymous reviewers for helpful comments and feedback.

References

[Ben Abacha *et al.*, 2017] Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*, 2017.

[Ben Abacha *et al.*, 2019] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019*, 2019.

[Boiko *et al.*, 2023] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Chen *et al.*, 2023] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[Christophe *et al.*, 2023] Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and Shadab Khan. Med42 - a clinical large language model. 2023.

[Clusmann *et al.*, 2023] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141, 2023.

[Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[Jin *et al.*, 2019] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Jin *et al.*, 2021] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[Kung *et al.*, 2023] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

[Liang *et al.*, 2019] Huiying Liang, Brian Y Tsui, Hao Ni, Carolina CS Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, Jiancong Chen, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine*, 25(3):433–438, 2019.

[Madani *et al.*, 2023] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, 2023.

[Muennighoff *et al.*, 2023] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Nori *et al.*, 2023a] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[Nori *et al.*, 2023b] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin,

Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Pal *et al.*, 2022] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[Peng *et al.*, 2019] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.

[Romera-Paredes *et al.*, 2023] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, pages 1–3, 2023.

[Singhal *et al.*, 2023a] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pages 1–9, 2023.

[Singhal *et al.*, 2023b] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Thirunavukarasu *et al.*, 2023] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[Toma *et al.*, 2023] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.

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

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[Wornow *et al.*, 2023] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

[Wu *et al.*, 2018] Ji Wu, Xien Liu, Xiao Zhang, Zhiyang He, and Ping Lv. Master clinical medical knowledge at certificated-doctor-level with deep learning model. *Nature communications*, 9(1):4352, 2018.

[Xiaofan Zhang, 2023] Shaoting Zhang Xiaofan Zhang, Kui Xue. Pulse: Pretrained and unified language service engine. 2023.

[Yang *et al.*, 2020] Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*, 2020.

[Zeng *et al.*, 2020] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, 2020.

[Zhang *et al.*, 2022] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, 2022.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.