Original Paper

Performance Gaps and Optimization Strategies in Chinese Medical Large Language Models Based on MedBench: Evaluation Study

Luyi Jiang[1*], MD; Jiayuan Chen[2*], MD; Lu Lu[2], MD; Xinwei Peng[2], MD; Lihao Liu[2], PhD; Junjun He[2], PhD; Jie Xu[2], DHM

[1]Shanghai Health Development Research Center (Shanghai Medical Information Center), Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Shanghai, China

[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China

*these authors contributed equally

Corresponding Author:

Jie Xu, DHM
Shanghai Artificial Intelligence Laboratory
No. 129 Longwen Road, Xuhui District
Shanghai 200235
China
Phone: 86 15221828927
Email: xujie@pjlab.org.cn

Abstract

Background: The evaluation and improvement of medical large language models (LLMs) are critical for their real-world deployment, particularly in ensuring accuracy, safety, and ethical alignment. However, existing frameworks are inadequate for dissecting domain-specific error patterns or addressing cross-modal challenges.

Objective: To address the limitations of current evaluation methods, this study aims to develop and apply a granular error taxonomy to systematically identify critical weaknesses in leading medical LLMs. The ultimate goal is to establish an actionable road map for enhancing their clinical robustness, safety, and overall trustworthiness.

Methods: This study introduces a granular error taxonomy developed through a systematic analysis of 10 top-performing models on MedBench (specifically AntAngelMed, Citrus-2.0, INF-Med, WHU_Med, zhuomuniao-Med, TeleChat2, hunyuan-med, UNI-GPT, fusiontech-Med, and GPT-4). Incorrect responses were categorized into 8 distinct types: omissions, hallucination, format mismatch, causal reasoning deficiency, contextual inconsistency, unanswered, output error, and deficiency in medical language generation. Based on these findings, we propose a tiered optimization strategy spanning 4 levels, from prompt engineering and knowledge-augmented retrieval to hybrid neuro-symbolic architectures and causal reasoning frameworks.

Results: Based on a comprehensive error analysis of 42,766 question responses generated by the top 10 models, the evaluation using the 8 defined metrics revealed significant vulnerabilities in the leading models. Despite achieving a high accuracy of 0.86 in medical knowledge recall, analysis across the 8 error categories identified omissions as the most prevalent issue, exhibiting a staggering 96.3% omission rate in critical reasoning tasks. Furthermore, safety and ethics evaluations showed alarming inconsistency under option-shuffled conditions, with a low robustness score of 0.79. Our analysis uncovers systemic weaknesses in the models' ability to enforce knowledge boundaries and perform multistep reasoning.

Conclusions: This work establishes an actionable road map for developing more clinically robust LLMs. By providing error-driven insights, it redefines evaluation paradigms, ultimately advancing the safety and trustworthiness of artificial intelligence in high-stakes medical environments and promoting its responsible application.

JMIR AI 2026;5:e86864; doi: 10.2196/86864

Keywords: medical LLMs; MedBench; incorrect responses; performance optimization; medical large language models; artificial intelligence; benchmarking

Introduction

In recent years, large language models (LLMs), empowered by massive text corpora and deep learning techniques, have demonstrated breakthrough advancements in cross-domain knowledge transfer and human-machine dialogue interactions [1]. Within the health care domain, LLMs are increasingly deployed across 9 core application scenarios, including intelligent diagnosis, personalized treatment, and drug discovery, garnering significant attention from both academia and industry [2,3]. A particularly important area of focus is the development and evaluation of Chinese medical LLMs, which face unique challenges due to the specialized nature of medical knowledge and the high-stakes implications of clinical decision-making. Hence, ensuring the reliability and safety of these models has become critical, necessitating rigorous evaluation frameworks [4].

Current research on medical LLM evaluation exhibits 2 predominant trends. On one hand, general-domain benchmarks (eg, HELM [5] and MMLU [6]) assess foundational model capabilities through medical knowledge tests. On the other hand, specialized medical evaluation systems (eg, MedQA [7] and C-Eval-Medical [8]) emphasize clinical reasoning and ethical compliance. Notably, the MedBench framework [9], jointly developed by institutions including Shanghai AI Laboratory, has emerged as the most influential benchmark for Chinese medical LLMs. By establishing a standardized evaluation system spanning 5 dimensions—medical language comprehension, complex reasoning, and safety ethics—it has attracted participation from hundreds of research teams.

However, existing studies remain constrained by several critical limitations. First, most evaluations focus on macro-level metrics (eg, accuracy and F_1-score) while lacking fine-grained error pattern analysis [10]. Second, conventional error taxonomies (eg, knowledge gaps and logical errors) fail to capture domain-specific deficiencies in medical contexts [11]. Third, standardized evaluation methodologies for cross-modal medical tasks (eg, radiology report generation) remain underdeveloped [12]. Furthermore, many legacy benchmarks struggle to differentiate the capabilities of state-of-the-art models, diminishing their utility in guiding model optimization.

The primary aim of this study is to address these challenges by proposing an innovative analytical framework for fine-grained error analysis. To address these challenges, this study leverages the MedBench database to propose an innovative analytical framework incorporating 8 error categories: omissions, hallucination, format mismatch, causal reasoning deficiency, contextual inconsistency, unanswered, output error, and deficiency in medical language generation. Through systematic analysis of error patterns across top-performing models, we reveal previously unidentified systemic weaknesses in Chinese medical LLMs, particularly in clinical pathway adherence and dialogue stability. This granular error analysis not only provides actionable insights for model refinement but also establishes a novel evaluation paradigm for developing safe and reliable medical AI systems.

Methods

Evaluation Framework and Metrics

To systematically assess the performance of LLMs in the medical field, we adopted a multidimensional evaluation framework under MedBench, encompassing objective multiple-choice questions (single-select and multiselect) and subjective open-domain questions. Accuracy and robustness were used as core metrics for objective tasks, while key information recall and hierarchical error taxonomy were used for subjective tasks (Table 1).

Table 1. Definition of assessment indicators.

Indicator	Definition
Accuracy	Accuracy refers to the proportion of correctly answered question versions among all model predictions. Each multiple-choice question was presented five times, with the same content but with the correct answer systematically rotated across options A–E. An accuracy of 100% indicates that the model correctly answered all five versions of the question.
Robustness	The model's robustness refers to its ability to produce fully consistent answers when presented with the same choice question where options are randomly shuffled.
Macro-recall	Macro-recall calculates the recall based on each of the answer points designed in the quiz question separately, and then takes the average of these recalls for the score.

Evaluation of Core Competency Performance

For objective questions, model accuracy was calculated by directly matching LLM-generated options against ground-truth answers. The 10 LLMs (specifically AntAngelMed [13], Citrus-2.0 [14], INF-Med [15], WHU_Med, zhuomuniao-Med, TeleChat2 [16], hunyuan-med [17], UNI-GPT, fusiontech-Med, and GPT-4 [18]) evaluated in this study comprise both publicly available models and proprietary models submitted by developers. For publicly accessible models (eg, GPT-4 and hunyuan-med), we accessed them via their official APIs or open-source platforms (detailed in Table S1 in Multimedia Appendix 1). For proprietary or competition-specific medical LLMs (eg, INF-Med and WHU_Med), model developers interacted directly with our MedBench platform via our standardized evaluation API. To ensure evaluation consistency across both access methods, we applied uniform prompt templates (detailed in Table S2 in Multimedia Appendix 1). These templates explicitly instructed the models to output their final answers in a standardized format. We then used rule-based text parsing

with regular expressions to extract the specific choices (eg, A, B, C, and D) from the models' raw output generations. Any responses that failed to yield recognizable options or deviated entirely from the instructed format were handled as invalid outputs. To rigorously validate robustness, answer choices for each multiselect question were shuffled across 10 permutations. A response was deemed valid only if the model consistently identified the correct answer or answers across all permutations, ensuring resistance to positional bias.

Error Patterns in Subjective Responses

Subjective responses were evaluated via macro-recall based on coverage of predefined key information points. To ensure evaluation objectivity, reproducibility, and analytical rigor, we systematically translated qualitative error descriptions into quantitative mathematical formulations. Errors were classified into 8 categories through a 3-expert consensus protocol using these formalized criteria: omissions (failure to address critical content), hallucination (fabricated claims), format mismatch (deviation from structured guidelines), causal reasoning deficiency (flawed logical chains), contextual inconsistency (contradictory statements), unanswered (no valid output), output error (technical failures), and deficiency in medical language generation (nonclinical phrasing). The detailed qualitative definitions, mathematical formulations, and illustrative examples are presented in Table 2.

Table 2. Incorrect responses, labels, and definitions.

Label	Define	Mathematical formulation	Illustrative example
Omissions	Model did not answer all the points scored	Let $K=\{k_1, k_2,..., k_N\}$ be the set of key points required by the ground truth, and M be the set of key points present in the model's response. The OR[a] is defined as: $OR = 1 - \frac{\lvert K \cap M \rvert}{\lvert K \rvert}$	The prompt asks for 3 typical influenza symptoms (K={fever, cough, fatigue}). The model only outputs "fever and cough" (M={fever, cough}). Thus, OR = 1 − 2/3 = 33.3%.
Hallucination	Model responses are not realistic or do not exist in the question stem	Let E_{model} be the set of medical entities or factual statements generated by the model, and C be the union of the provided context (question stem) and established medical knowledge bases (eg, UMLS[b]). The HR[c] is defined as: $HR = \frac{\lvert E_{model} \setminus C \rvert}{\lvert E_{model} \rvert}$	The provided patient history only mentions "hypertension." The model responds, "Considering the patient's history of hypertension and diabetes..." ("diabetes" is an ungrounded entity, belonging to E_{model} but not C).
Format mismatch	Model does not output the content in the characteristic format as required or the output is in the wrong format	Let Φ be the set of formatting requirements (eg, JSON structure and word and paragraph limits), and $L(\Phi)$ be the valid output space that satisfies these constraints. We define an indicator function I_{FM}: $I_{FM}(\text{Output}) = \begin{cases} 0, & \text{if Output} \in L(\Phi) \\ 1, & \text{if Output} \notin L(\Phi) \end{cases}$	The prompt strictly requires "Output in JSON format, containing 'diagnosis' and 'treatment' keys." The model instead generates a plain text paragraph, failing to adhere to the structural constraint.
Causal reasoning deficiency	Overinference or underinference of the content of model responses	We model the reasoning process as a directed graph $G=(V, E)$, where nodes V represent clinical states and edges E represent causal links. Let E^* be the edges of the gold-standard reasoning chain and $\hat{E}$ be the edges generated by the model. The deficiency distance (D_{CR}) is defined as: $D_{CR} = \lvert \hat{E} \setminus E^* \rvert + \lvert E^* \setminus \hat{E} \rvert$	Gold-standard reasoning chain: Pathogen infection → Airway inflammation → Cough. Underinference: "Cough is due to infection" (skipping the inflammation mechanism). Overinference: "Infection → Malignant tumor" (introducing a non-existent causal link).
Contextual inconsistency	Inconsistent or irrelevant model outputs	Let the model's response be a sequence of propositions $S=\{s_1, s_2,..., s_n\}$. Inconsistency is triggered if there is an internal logical contradiction (ie, $\exists i,j$ such that $s_i \wedge s_j \Longrightarrow$ False), or if the semantic similarity score $Sim(S, C)$ between the response S and the context C falls below a predefined threshold τ (ie, $Sim(S, C) < \tau$).	In the first paragraph, the model states, "Aspirin is contraindicated for this patient," but in the concluding prescription, it writes, "Recommend taking Aspirin 100 mg." The 2 propositions are mutually exclusive.
Unanswered	Model did not answer the question	Let $\Omega_{refusal}$ be the set of common AI refusal templates, and H(Output) be the effective information entropy of the output. The indicator function I_{UN} is defined as: $I_{UN} = \begin{cases} 1, & \text{if Output} \in \Omega_{\text{refusal}} \text{ or } H(\text{Output}) \approx 0 \\ 0, & \text{otherwise} \end{cases}$	The model outputs, "Sorry, as an AI, I cannot provide a medical diagnosis," or it simply generates meaningless punctuation marks.
Output error	Model output is completely incorrect	Based on clinical facts, we define an accuracy scoring function $Score_{\text{acc}}(\text{Output}, \text{Gold}) \in [0, 1]$. Given a lower bound of error tolerance ϵ, the output error indicator I_{OE} is defined as: $I_{OE} = \begin{cases} 1, & \text{if Output} < \epsilon \\ 0, & \text{otherwise} \end{cases}$	The prompt asks, "What is the normal resting heart rate for adults?" The model answers, "10 to 20 beats per minute" (a statement that completely deviates from objective medical facts).
Deficiency in medical language generation	The model's inability to translate everyday conversational content into clinically accurate medical text	Let T_{model} be the set of symptom and disease vocabulary used in the model's response, and $V_{standard}$ be standard clinical terminologies (eg, SNOMED CT and MeSH). The nonprofessional ratio (R_{np}) is defined as: $R_{np} = \frac{\lvert E_{model} \setminus C \rvert}{\lvert E_{model} \rvert}$	In a task requiring a professional discharge summary, the model uses colloquial terms like "tummy ache" and "runny nose" instead of standard medical terminologies like "abdominal pain" and "rhinorrhea."

[a]OR: omission rate.
[b]UMLS: Unified Medical Language System.
[c]HR: hallucination rate.

Ethical Considerations

This study involved the analysis of publicly available data and computational models (LLMs). It did not involve human participants, human data, human tissue, or animal subjects. Therefore, in accordance with the policies of *JMIR AI*, ethical approval from an institutional review board or research ethics board was not required. All methods were carried out in accordance with relevant guidelines and regulations.

Results

Core Competency Performance

From June to December 2024, standardized evaluations of 44,711 question responses generated by the top 10 models (AntAngelMed, Citrus-2.0, INF-Med, WHU_Med, zhuomuniao-Med, TeleChat2, hunyuan-med, UNI-GPT, fusiontech-Med, and GPT-4) across MedBench's 5 competency dimensions revealed distinct capability patterns. As shown in Table 3, models achieved peak accuracy (0.86) in medical knowledge question answering, demonstrating robust comprehension of foundational concepts. Health care safety and ethics emerged as a secondary strength (accuracy 0.80), yet its robustness score (0.79) under shuffled permutations highlighted persistent inconsistencies in ethical decision-making. Aligned with our 8 newly defined error metrics, critical gaps persisted in medical language understanding (challenges in parsing specialized terminology, often leading to deficiency in medical language generation) and complex medical reasoning (limited multistep diagnostic inference, frequently resulting in causal reasoning deficiency and omissions).

Table 3. Accuracy and robustness performance of the top 10 models.

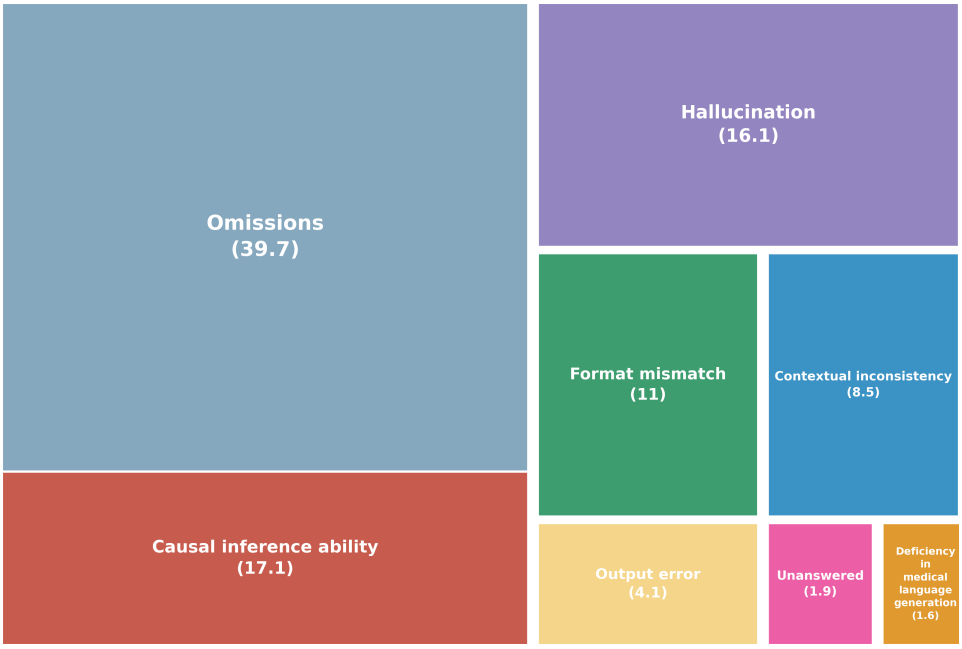
Dimension	Number of question responses	Accuracy	Robustness
Medical knowledge question answering	3703	0.86	0.94
Medical language generation	1530	0.76	—[a]
Medical language understanding	5784	0.71	—
Complex medical reasoning	891	0.72	—
Health care safety and ethics	32,802	0.80	0.79
Total	44,711	—	—

[a]Not applicable.

Error Patterns in Subjective Responses

A granular dissection of incorrect responses identified omissions of critical answer points as the dominant failure mode (4723/11,909 errors, 39.7%), reflecting systemic deficiencies in comprehensive reasoning despite superficially coherent outputs. Causal inference breakdowns (2037/11,909, 17.1%) and hallucinatory content generation (1913/11,909, 16.1%) constituted secondary yet consequential flaws, while format mismatch accounted for 1307 out of 11,909 errors (11%; Figure 1), underscoring the need for standardized output frameworks.

Figure 1. Proportions of incorrect responses in labels.



Domain-specific failure profiles further elucidated model limitations (Figure 2). In medical knowledge question answering, omissions dominated (1644/3703 errors, 44.4%), followed by hallucinations (868/3703, 23.4%), exposing weak knowledge boundary safeguards. For medical language understanding, 1984 out of 5784 errors (34.3%) stemmed from unaddressed contextual constraints, while 1303 out of 5784 errors (22.5%) arose from fractured causal reasoning chains, indicating deficits in clinical narrative modeling. Medical language generation exhibited severe hallucination (979/1530 errors, 64%), revealing semantic-clinical pragmatics disconnects. Strikingly, complex medical reasoning errors were overwhelmingly dominated by omissions (858/891 errors, 96.3%), fundamentally undermining reasoning reliability. As documented in Table 4, a granular error taxonomy delineates the classification schemas and clinical manifestations of these incorrect responses.

Figure 2. Proportions of incorrect responses in labels under different dimensions.

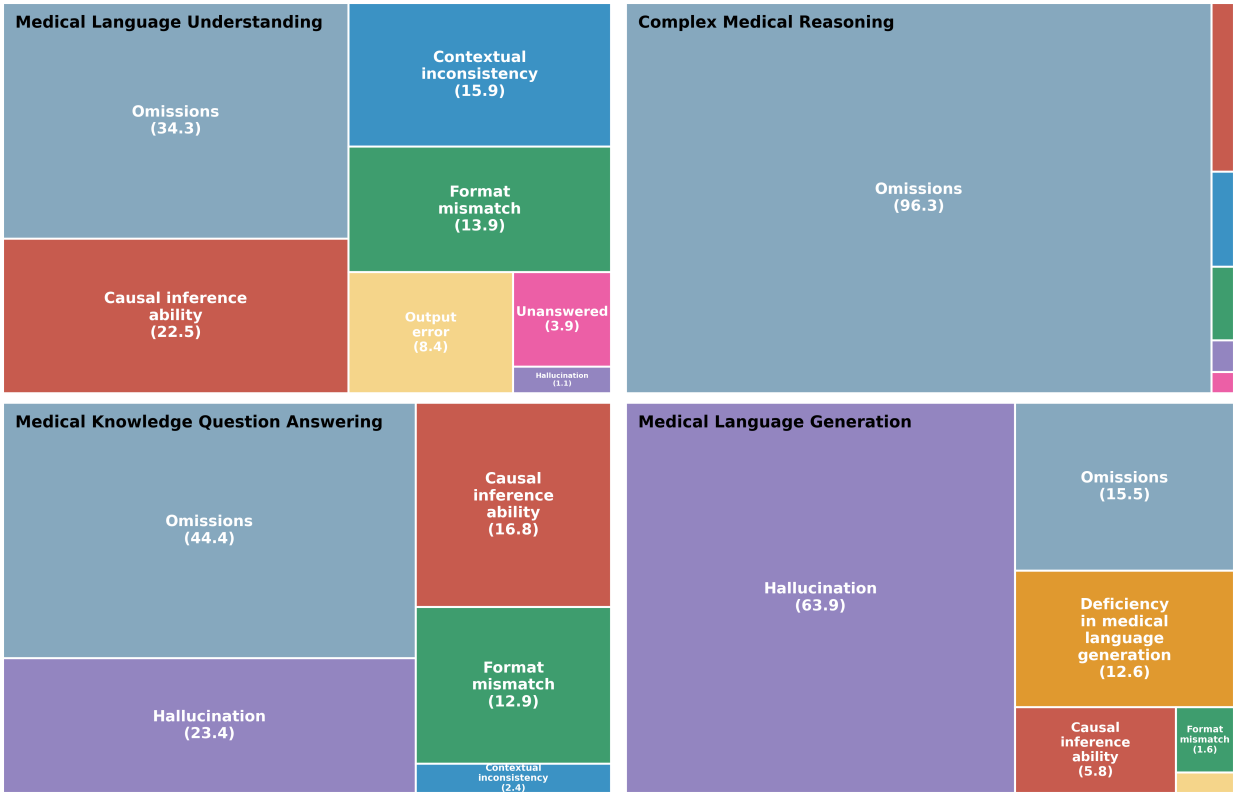


Table 4. Labeling of incorrect responses and corresponding examples.

	Question	Answer points	Model response
4.1 Omissions	肠结核有哪些临床表现？ What are the clinical manifestations of intestinal tuberculosis?	[“腹痛,” “大便习惯改变,” “腹部肿块,” “全身症状,” “低热,” “盗汗,” “贫血”] [“abdominal pain,” “change in bowel habits,” “abdominal mass,” “systemic symptoms,” “low-grade fever,” “night sweats,” “anemia”]	肠结核的临床表现主要包括：发热、盗汗、消瘦、腹痛、腹泻、便秘、腹部包块等。其中，腹痛多位于右下腹或脐周，呈隐痛或钝痛，餐后加重，排便后减轻。腹泻与便秘可交替出现，腹泻多为糊状便，无脓血。腹部包块多位于右下腹，质地较硬，活动度差，有压痛。部分患者可有肠梗阻表现，如腹胀、呕吐等。 The clinical manifestations of intestinal tuberculosis mainly include: fever, night sweats, emaciation, abdominal pain, diarrhea, constipation, abdominal mass, etc. Among them, abdominal pain is mostly located in the right lower abdomen or around the umbilicus, which is hidden pain or dull pain, aggravated after meals and alleviated after defecation. Diarrhea and constipation may occur alternately, and the diarrhea is mostly paste-like stool without pus and blood. Abdominal mass is mostly located in the right lower abdomen, hard texture, poor mobility, with pressure pain. Some patients may have intestinal obstruction manifestations, such as abdominal distension and vomiting.
4.2 Hallucination	粒细胞缺乏症的定义是什么? What is the definition of granulocyte deficiency?	[“中性粒细胞绝对计数,” “0.5 ×10^9/L”] [“Absolute neutrophil count,” “0.5 × 10^9/L”]	粒细胞缺乏症是一种严重的血液疾病，主要特征是血液中的中性粒细胞数量极度减少或完全缺失。中性粒细胞是白细胞的一种，主要负责抵抗感染。因此，粒细胞缺乏症患者极易发生严重的感染，病情危重，死亡率高。 粒细胞缺乏症的发生可能与多种因素有关，包括粒细胞生成减少、成熟障碍、消耗过多等。一些疾病，如再生障碍性贫血、急性白血病、感染、肿瘤骨髓浸润等，可能导致粒细胞减少症。维生素 B12 或叶酸缺乏也可能导致粒细胞成熟障碍。此外，系统性红斑狼疮、败血症、严重感染等可能导致粒细胞过度消耗。 粒细胞缺乏症的治疗主要以升白细胞药物为主，同时需要控制感染。肾上腺皮质激素、升白细胞药物、粒细胞输注等方法也可以用于缓解症状。患者需要注意休息，保持局部卫生，避免感染，并避免食用辛辣刺激性食物。 Granulocyte deficiency is a serious blood disorder characterized by an extreme decrease or complete absence of neutrophils in the blood. Neutrophils are a type of white blood cell that is primarily responsible for fighting off infections. Therefore, patients with granulocyte deficiency are highly susceptible to serious infections, critical conditions and high mortality rates. The development of granulocyte deficiency may be associated with a number of factors, including decreased production of granulocytes, impaired maturation, and excessive depletion. Some diseases, such as aplastic anemia, acute leukemia, infections, and bone marrow infiltration of tumors, may lead to granulocytopenia. Vitamin B12 or folate deficiency may also lead to impaired granulocyte maturation. In addition, systemic lupus erythematosus, sepsis, and severe infections may lead to excessive granulocyte depletion. Treatment of granulocyte deficiency is primarily based on leukocyte-boosting drugs, and infection control is required. Adrenocorticotropic hormone, leukocyte-boosting drugs, and granulocyte infusion can also be used to relieve symptoms. Patients need to pay attention to rest, maintain local hygiene, avoid infection, and avoid spicy and irritating foods.
4.3 Format mismatch	医学命名实体识别任务: 请从给定医学文本中，识别出命名实体。 实体识别的格式为: 药物: ”“ 医疗设备: ”“ 医院科室: ”“ 微生物类: ”“ 身体部位: “” 医疗程序: “” 医学检验项目: “” 症状和体征: “” 疾病: “ ” Medical Named Entity Recognition Task: Please identify named entities from a given medical text. The format for entity identification is: Drug: “” Medical device: “” Hospital department: “” Microorganisms: “” Body part: “” Medical Procedures: “” Medical Test Programs: “” Signs and symptoms: “” Diseases: “”	“药物”: [] “医疗设备”: [] “医院科室”: [] “微生物类”: [] “身体部位”: [] “医疗程序”: [] “医学检验项目”: [] “症状和体征”: [“FMR-1 基因的突变”] “疾病”: [“脆性 X 综合征”] “Drugs”: [] “Medical devices”: [] “Hospital units”: [] “Microbiology”: [] “Body parts”: [] “Medical procedures”: [] “Medical tests”: [] “Symptoms and signs”: [“Mutations in the FMR-1 gene”] “Diseases”: [“Fragile X syndrome”]	药物: 医疗设备: 医院科室: 微生物类: 身体部位: 医疗程序: 医学检验项目: 症状和体征: 疾病: 脆性 X 综合征 基因: FMR-1 基因 Drugs: Medical Equipment: Hospital Departments: Microbiology: Body Parts: Medical Procedures: Medical Test Program: Signs and symptoms: Disease: fragile X syndrome Gene: FMR-1 gene

	Question	Answer points	Model response
4.4 Causal reasoning deficiency	现病史 病史摘要 赵某，男，7 岁，3 天前无明显诱因出现腹部疼痛，伴寒战高热，最高体温 38.7°C，于当地行“退热止痛”治疗后未见明显好转。患儿于 1 天前出现皮肤黄染，同时伴巩膜黄染，精神差，营养状况较差。病儿自发病以来，精神睡眠差，饮食差，大便成形，偶呈陶土样色，尿呈浓茶样色。 主诉 腹痛高热 3 天，伴皮肤黄染 1 天。 体格检查 T38.5°C，P92 次/分，R22 次/分，Bp110/78mmHg。 发育正常，营养中等，自主体位，查体合作，全身皮肤及巩膜黄染，上肢皮肤可见散在皮下出血点，全身浅表未及肿大淋巴结。两肺呼吸音粗，未闻及干湿啰音。心率 92 次/分，律齐，未闻及病理性杂音。腹部略膨隆，肝脏肿大，表面光滑，质地坚硬，边缘圆钝，于肝区肋弓下可触及一肿物，形状规则，大小约 3cm×2cm，右上腹压痛，反跳痛（±），全腹叩诊鼓音，移动性浊音阴性，肠鸣可闻及。 辅助检查 实验室检查 血常规：WBC15.91×10^9/L，NEU89.7%，RBC3.74×10^12/L，Hb115g/L，PLT286×10^9/L。 肝功能：ALT70U/L，AST56U/L；总胆红素 176μmol/L，间接胆红素 25μmol/L，直接胆红素 151μmol/L；碱性磷酸酶 984U/L；谷氨酰转移酶 1065.5U/L。 影像学表现 腹部 CT 平扫 1.CT 检查显示胆总管呈巨大囊样扩张。 2.胆囊体积增大。 3.无肝内胆管远端分支扩张。 简述本例病人的诊断及诊断依据。 Present Medical History (1) Summary of history Zhao Mou, male, 7 years old, had abdominal pain with chills and high fever with a maximum temperature of 38.7°C 3 days ago without any obvious cause, and did not show significant improvement after local treatment of “antipyretic and analgesic”. The child developed skin yellowing 1 day ago, accompanied by scleral yellowing, poor mental health and poor nutritional status. Since the onset of the disease, the child had poor mental sleep, poor diet, formed stools, occasionally clay-like color, and urine was strong tea-like color. (2) Complaints Abdominal pain and high fever for 3 days, accompanied by skin yellowing for 1 day. Physical examination T38.5°C, P92 beats/min, R22 beats/min, Bp110/78mmHg. Development is normal, nutrition is moderate, autonomous position, cooperation in physical examination, generalized skin and sclera yellow staining, scattered subcutaneous hemorrhages can be seen on the skin of the upper limbs, generalized superficial not found enlarged lymph nodes. Respiratory sounds in both lungs were coarse, and no dry or wet rales were detected. Heart rate was 92 beats/minute, rhythmic, no pathological murmur was heard. The abdomen was slightly distended, the liver was enlarged, smooth surface, hard texture, rounded edges, a mass was palpable under the costal arch of the liver area, regular shape, size about 3cm×2cm, pressure pain in the right epigastric region, rebound pain (±), tympanic sound on percussion in the whole abdomen, mobile turbid sound was negative, and intestinal tinnitus could be heard. Auxiliary examination Laboratory examination Blood routine: WBC15.91×10^9/L, NEU89.7%, RBC3.74×10^12/L, Hb115g/L, PLT286×10^9/L. Liver function: ALT70U/L, AST56U/L; total bilirubin 176μmol/L, indirect bilirubin 25μmol/L, direct bilirubin 151μmol/L; alkaline phosphatase 984U/L; glutamyl transferase 1065.5U/L. Imaging manifestations Abdominal CT plain scan 1. CT examination showed a huge cystic dilatation of the common bile duct. 2. Increased gallbladder volume. 3. No dilatation of the distal branches of the intrahepatic bile duct. Briefly describe the diagnosis and diagnostic basis of this patient.	[“先天性胆管扩张症,” “腹部疼痛,” “高热,” “寒战,” “黄染,” “直接胆红素,” “囊样扩张”] [“congenital choledocholithiasis,” “abdominal pain,” “high fever,” “chills,” “yellow stain,” “direct bilirubin,” “cystic dilatation”]	本例患儿的诊断可能为胆道梗阻导致的阻塞性黄疸，具体原因需进一步明确。 诊断依据包括： 1. 临床症状：患儿出现腹痛、寒战、高热，随后出现皮肤和巩膜黄染，精神差，营养状况较差。这些症状提示可能存在感染和/或肝脏疾病。 2. 体格检查：皮肤和巩膜黄染，上肢皮下出血点，肝脏肿大，质地坚硬，右上腹压痛，这些体征进一步支持肝脏疾病和/或胆道问题。 3. 实验室检查：血常规显示白细胞升高，中性粒细胞比例增高，提示有感染的可能。肝功能异常，ALT 和 AST 轻度升高，总胆红素和直接胆红素显著升高，碱性磷酸酶和谷氨酰转移酶显著升高，这些都是胆道梗阻和肝细胞损伤的标志。 4. 影像学检查：腹部 CT 平扫显示胆总管呈巨大囊样扩张，胆囊体积增大，但无肝内胆管远端分支扩张。这表明胆道系统存在异常，可能是胆道梗阻的直接证据。 综合上述临床表现、体格检查、实验室检查和影像学检查，可以初步诊断为胆道梗阻导致的阻塞性黄疸。但是，需要进一步的检查，如胆道超声、MRCP（磁共振胰胆管造影）或 ERCP（内镜逆行胰胆管造影）等，以确定梗阻的具体位置和原因，可能是胆结石、胆道炎症、先天性胆道异常或其他原因所致。 The diagnosis of this child may be obstructive jaundice due to biliary obstruction, the exact cause of which requires further clarification. The diagnostic basis includes: 1. Clinical symptoms: the child presented with abdominal pain, chills, and high fever, followed by yellowing of the skin and sclera, poor mental health, and poor nutritional status. These symptoms are suggestive of possible infection and/or liver disease. 2. Physical examination: yellowing of the skin and sclera, subcutaneous hemorrhagic spots on the upper extremities, an enlarged, firm liver, and right upper abdominal tenderness are signs that further support the presence of liver disease and/or biliary tract problems. 3. Laboratory tests: Blood counts show elevated leukocytes and an increased percentage of neutrophils, suggesting infection. Liver function is abnormal with mild elevations of ALT and AST, significant elevations of total and direct bilirubin, and significant elevations of alkaline phosphatase and glutamyltransferase, all of which are markers of biliary obstruction and hepatocellular injury. 4. Imaging: CT scan of the abdomen shows a huge cystic dilatation of the common bile duct and an increase in gallbladder volume without dilatation of the distal branches of the intrahepatic bile ducts. This indicates an abnormality in the biliary system and may be direct evidence of biliary obstruction. Taking into account the above clinical manifestations, physical examination, laboratory tests and imaging studies, a preliminary diagnosis of obstructive jaundice due to biliary obstruction can be made. However, further investigations such as biliary ultrasound, MRCP (magnetic resonance cholangiopancreatography) or ERCP (endoscopic retrograde cholangiopancreatography) are needed to determine the exact location and cause of the obstruction, which may be due to gallstones, biliary inflammation, congenital biliary anomalies or other causes.
4.5 Contextual inconsistency	现病史 病史摘要 陈 XX，男，65 岁，近一周来反复排暗红色血便，伴头晕乏力。半天前再次排血便，量约 600ml，并出现神志淡漠，尿量减少，12 小时尿量为 200ml。无发热、恶心、呕吐，既往病史无特殊。 主诉 反复排血便 1 周，加重伴少尿半天。 体格检查 生命体征 T36.8°C，P105 次/分，R20 次/分，Bp90/60mmHg。体重 60kg。 自主体位，神志淡漠，呼之可对答。全身皮肤及巩膜较苍白，四肢偏凉。两肺呼吸音清晰，未闻及干湿啰音。心率 105 次/分，律齐，未闻及病理性杂音。腹部平软，肝脾不大，肠鸣音活跃。双肾区无叩痛，双下肢无水肿。 辅助检查 血液分析：白细胞计数 8.8×10^9/L，中性粒细胞比例 0.7，红细胞计数 2.8×10^12/L，血红蛋白 78g/L 尿液分析：尿比重 1.030，尿白细胞（-），尿潜血（-），尿蛋白（-） 大便潜血：（+），胃潜血：（-） 生化电解质：钾 4.5mmol/L，钠 138mmol/L，氯 103.5mmol/L，二氧化碳结合力 22mmol/L 肝肾功能：尿素氮 9.8mmol/L，肌酐 201μmol/L，谷丙转氨酶 10U/L，谷草转氨酶 20U/L 纤维结肠镜检查：提示降结肠及直肠可见弥漫出血点。 B 超检查：肝、胆、脾、胰及泌尿系 B 超无特殊。 简述本例病人的治疗原则。 Present Medical History Summary of history Chen XX, male, 65 years old, had repeated dark red blood stools with dizziness and fatigue in the past week. Half a day ago, he passed blood stool again, with a volume of about 600 ml, and appeared to be apathetic, with a decreased urine output of 200 ml in 12 h. There was no fever, nausea, or vomiting, and his past medical history was unremarkable. Complaints Recurrent bloody stools for 1 week, aggravated with oliguria for half a day. Physical examination Vital signs T36.8°C, P105 beats/min, R20 beats/min, Bp90/60mmHg. weight 60kg. He was in voluntary position, apathetic, and could answer when called. The skin and sclera were pale, and the limbs were cold. Breath sounds of both lungs were clear, and no dry or wet rales were heard. Heart rate was 105 beats/minute, rhythmic, no pathological murmur was heard. The abdomen was flat and soft, the liver and spleen were not large, and bowel sounds were active. There was no tenderness in both kidneys and no edema in both lower limbs. Auxiliary examination Blood analysis: white blood cell count 8.8×10^9/L, neutrophil ratio 0.7, red blood cell count 2.8×10^12/L, hemoglobin 78g/L. Urinalysis: urine specific gravity 1.030, urine leukocytes (-), urine occult blood (-), urine protein (-) Fecal occult blood: (+), gastric occult blood: (-) Biochemical electrolytes: potassium 4.5 mmol/L, sodium 138 mmol/L, chloride 103.5 mmol/L, carbon dioxide binding capacity 22 mmol/L Liver and kidney function: urea nitrogen 9.8 mmol/L, creatinine 201 μmol/L, alanine aminotransferase 10 U/L, azelaic aminotransferase 20 U/L Fiber colonoscopy: suggesting diffuse bleeding spots visible in the descending colon and rectum. Ultrasound examination: liver, gallbladder, spleen, pancreas and urinary tract ultrasound is not special. Briefly describe the principles of treatment for this patient.	[“止血,” “输血,” “液体平衡,” “电解质及酸碱平衡”] [“Hemostasis,” “Transfusion,” “Fluid balance,” “Electrolyte and acid-base balance”]	1.该病例出血原因考虑为结肠病变，反复出血考虑为慢性失血，半天前大量出血为慢性失血急性加重。病人体重偏轻，平时饮食习惯不佳，半天前大量出血后尿量减少，考虑血容量不足，急性肾损伤。根据病人的血红蛋白水平及结肠镜检查结果，应予以输血补液，纠正贫血，维持水电解质平衡，同时请肛肠外科会诊讨论是否手术治疗。 2.在补充血容量后，病人尿量增加，肌酐可缓慢下降至正常。若肌酐继续上升，考虑行肾脏替代治疗。 3.病人的贫血主要原因为消化道出血，为缺铁性贫血，待病人病情稳定后，可予以补充铁剂。 1. The cause of bleeding in this case is considered to be colonic lesions, repeated bleeding is considered to be chronic blood loss, and massive bleeding half a day ago is an acute exacerbation of chronic blood loss. The patient’s light weight, poor dietary habits in general, and decreased urine output after massive bleeding half a day ago are considered to be hypovolemia and acute kidney injury. According to the patient’s hemoglobin level and colonoscopy results, blood transfusion and fluid replacement should be given to correct anemia and maintain the water-electrolyte balance, and at the same time, we should ask the anorectal surgery department to consult and discuss whether surgical treatment. 2. After supplementing blood volume, the patient’s urine output increases, and creatinine may slowly decrease to normal. If creatinine continues to rise, consider renal replacement therapy. 3. The patient’s anemia is mainly caused by gastrointestinal bleeding, which is iron deficiency anemia. After the patient’s condition is stabilized, iron supplements can be given.

	Question	Answer points	Model response
4.6 Unanswered	现病史 病史摘要 病人 2017 年 8 月 15 日夜间驾驶摩托车被出租车横向撞伤，当时诉腹部胀痛、左膝疼痛，急症入六安市人民医院，给予保守治疗腹痛未见明显好转，且腹痛加重。2017 年 8 月 17 日转至安医大二附院，检查提示腹腔积液明显，急诊剖腹探查。术中见腹盆腔大量肠液样液体，约 2000ml，全小肠肠壁炎症水肿，距屈氏韧带约 45cm 处小肠系膜撕裂，致约 40cm 小肠发黑，并可见 2 处破裂口。行小肠部分切除肠吻合+小肠结肠浆肌层修补术。术后病人持续高热，最高 40 摄氏度，低血压、乳酸持续性升高，最高 7.66mmol/L，转至 ICU 给予呼吸机辅助呼吸、血管活性药物等治疗，病人未见明显好转。于 2017 年 8 月 21 日夜间 8 时转入我院，入院时病人神志淡漠，呼吸机辅助呼吸，心率增快，约 130 次/分；血压 130/80mmHg 左右；体温 38.3°C，给予拔除腔静脉导管；抽取血培养，亚安培南+伏立康唑抗感染；给予抑酸、抑酶，静脉营养支持；监测膀胱压；病人全身组织水肿明显，限制液体输入，给予蛋白后利尿（24 小时共出尿液约 7000ml）。 主诉 车祸外伤术后多脏器功能障碍 3 天。 体格检查 （1）一般情况 神志模糊，精神欠佳，慢性病容，表情痛苦，发育正常，营养中等，匀称，抬入病房，被动体位，查体合作，语言正常，声音低微，对答切题。 （2）专科情况 腹部膨隆，腹式呼吸存在，无腹壁静脉曲张，未见肠型及蠕动波，无压痛及反跳痛，未触及包块。腹部正中可见一手术伤口，少量脓性液体，腹部可见右侧两根腹腔引流管。肝脾肋下未触及，未触及胆囊，Murphy 征阳性。腹部鼓音区未及，无移动性浊音。肝上界位于右锁骨中线第 5 肋间，下界位于右季肋下缘，肝区无叩痛，脾浊音区正常，胆囊区无叩痛。肠鸣音弱，1 次/分，未闻及血管杂音及摩擦音。 辅助检查 （1）实验室检查 白细胞计数 9.5×10^9/L，红细胞计数 3.28×10^12/L，血红蛋白 98g/L，血小板计数 498×10^9/L，C 反应蛋白 74.9mg/L。 （2）X 线检查 床边胸片提示两肺炎症，左侧胸腔少量积液。 （3）CT 检查 提示腹腔广泛渗出。 病情变化 2017 年 8 月 22 日 19 时体温升高，最高 39°C，20 时血压下降，最低 77/34mmHg；心率升高，最高 140 次/分。给予加快输液速度，补充晶体液；同时给予白蛋白、红悬、血浆等补充胶体液；给予冷沉淀改善凝血功能；22 时病人血压恢复至 120/80mmHg 左右，心率恢复至 80 次/分。 该病人可下哪些诊断？ Present Medical History Summary of medical history The patient was driving a motorcycle at night on August 15, 2017 and was hit by a cab horizontally, at that time, he complained of abdominal distension and pain, left knee pain, and was admitted to the Lu'an People's Hospital in an emergency, and was given conservative treatment for abdominal pain which did not show any significant improvement and worsening of the abdominal pain. He was referred to the Second Affiliated Hospital of the Anhui University of Medical Science and Technology on August 17, 2017, and his examination suggested that there was obvious accumulation of fluid in the abdominal cavity, and he was urgently admitted to the hospital by caesarean section. Intraoperatively, a large amount of intestinal fluid-like fluid was seen in the abdominopelvic cavity, about 2,000 ml, inflammation and edema of the intestinal wall of the whole small intestine, and the mesenteric laceration of the small intestine about 45 cm from the flexural ligament, which resulted in the blackness of the small intestine of about 40 cm and 2 rupture orifices were seen. Partial resection of small bowel with intestinal anastomosis + small bowel colonic plasma muscle layer repair was performed. Postoperatively, the patient had persistent high fever up to 40 degrees Celsius, hypotension, persistent elevation of lactate up to 7.66 mmol/L. He was transferred to ICU and given ventilator-assisted respiration, vasoactive drugs, etc., and the patient did not show significant improvement. Transferred to our hospital at 8:00 p.m. on the night of August 21, 2017, at the time of admission, the patient was apathetic, ventilator-assisted respiration, heart rate increased, about 130 beats/min; blood pressure about 130/80 mmHg; temperature 38.3 °C, given to remove the vena cava catheter; blood culture was taken, subamphetamine + voriconazole anti-infective; acid suppression, enzyme suppression, intravenous nutritional support was given; bladder pressure was monitored; the patient's systemic tissue Edema was obvious, fluid intake was restricted, and post-protein diuresis was given (a total of about 7000 ml of urine was produced in 24 hours). Main complaint Multi-organ dysfunction for 3 days after car accident trauma. Physical examination (1) General condition Ambiguous, mentally ill, chronic look, painful expression, normal development, moderate nutrition, well-proportioned, carried into the ward, passive position, cooperative in body checking, normal speech, low voice, tangential answers. (2) Specialized conditions The abdomen was distended, abdominal respiration was present, no abdominal wall varices, no bowel pattern or peristaltic wave was seen, no pressure pain or rebound pain, and no mass was palpable. A surgical wound with a small amount of purulent fluid was seen in the middle of the abdomen, and 2 abdominal drains on the right side were seen in the abdomen. Liver and spleen were not palpated subcostally, gallbladder was not palpated and Murphy's sign was positive. Abdominal tympanic area was not reached and there was no mobile turbidity. The upper border of the liver was located at the 5th intercostal space in the right midclavicular line and the lower border was located at the lower border of the right quaternary ribs; there was no percussion pain in the hepatic region, the splenic turbid zone was normal, and there was no percussion pain in the gallbladder region. Bowel sounds were weak, 1 beats/minute, no vascular murmur or friction sounds were heard. Auxiliary examination (1) Laboratory tests White blood cell count 9.5×10^9/L, red blood cell count 3.28×10^12/L, hemoglobin 98g/L, platelet count 498×10^9/L, C-reactive protein 74.9mg/L. (2) X-ray examination Bedside chest radiograph suggested inflammation in both lungs and a small amount of fluid in the left pleural cavity. (3) CT examination Suggests extensive exudation in the abdominal cavity. Disease change At 19:00 on August 22, 2017, the body temperature increased, with a maximum of 39°C. At 20:00, the blood pressure decreased, with a minimum of 77/34 mmHg; the heart rate increased, with a maximum of 140 beats/min. Accelerated infusion rate was given to replenish crystalloid; albumin, red suspension, and plasma were also given to replenish colloid fluids; cold precipitation was given to improve coagulation; at 2200 hours, the patient's blood pressure recovered to about 120/80 mmHg, and the heart rate recovered to 80 beats/min. What diagnosis can be made in this patient?	["腹腔感染," "脓毒症," "多脏器功能障碍"] ["abdominal infection," "sepsis," "multiple organ dysfunction"]	_[a]

	Question	Answer points	Model response
4.7 Output error	任务: 将给定实体进行归一化，从候选的若干个 ICD-10 诊断标准词中选择出与原诊断描述匹配的词。 输入医学短语: 右侧睾丸胚胎癌 候选标准词: 甲状腺恶性肿瘤##癌##肺继发恶性肿瘤##恶性肿瘤放疗##转移性癌 子宫颈恶性肿瘤##鳞状细胞癌##高血压 麦克尔憩室##小肠憩室炎 抑郁发作##脑血管病 颊部血管畸形 睾丸恶性肿瘤##胚胎性癌 归一化术语: TASK: Normalize a given entity to select from a candidate number of ICD-10 diagnostic criteria words that match the original diagnostic description. Enter the medical phrase: Embryonal carcinoma of the right testis Candidate Criteria Words: Thyroid malignancy##Cancer##Secondary malignant tumor of lung##Radiotherapy for malignancy##Metastatic cancer malignant tumor of the cervix##squamous cell carcinoma##hypertension Michael's diverticulum##Small bowel diverticulitis Depressive episodes ##Cerebrovascular disease Buccal vascular malformations Testicular malignancy ##embryonic carcinoma Normalized Terminology:	睾丸恶性肿瘤##胚胎性癌 Testicular malignant tumors##Embryonal carcinoma	鳞状细胞癌##颊部血管畸形##糖尿病##脐部感染 Squamous cell carcinoma ##Buccal vascular malformation ##Diabetes ##Umbilical cord infection
4.8 Deficiency in medical language generation	任务：假设你是一位专业医生，请根据以下医患对话记录，为患者总结问诊报告。 要求输出格式为: 主诉: “” 现病史: “” 辅助检查: “” 既往史: “” 诊断: “” 建议: “” 问诊对话: 患者: 宝宝四个月，最近几天老吐豆腐渣那种。是怎么回事？前几天拉肚子，在吃药 医生: 您好 患者: 你好 医生: 每天几次大便 患者: 请问宝宝四个月 医生: 咳嗽，流鼻涕，发烧不 患者: 最近老吐豆腐渣那种 患者: 偶尔咳嗽 患者: 不发烧 患者: 偶尔流清鼻涕 医生: 每天几次大便 患者: 这几天都是四五次 医生: 宝宝现在是母乳喂养吗 医生：吃奶好不 患者: 恩母乳 患者: 好 医生: 精神状态怎么样 患者: 好 医生: 现在用什么药呢 患者: 金盆贝康，和蒙脱石 医生: 化验大便没 患者: 还有一些小颗白药 患者: 没那 医生: 腹泻几天了 患者: 有四五天了 医生: 这种情况建议去医院化验大便看看排除肠炎 TASK: Assuming you are a medical professional, summarize the consultation report for the patient based on the following transcript of the doctor-patient conversation. The required output format is: Chief complaint:“” Present medical history:“” Ancillary examination: “” Past history: “” Diagnosis: “” Recommendations: “” Interview dialog: PATIENT: The baby is 4 months old and has been spitting up the tofu crumb kind for the last few days. What's going on? He had diarrhea a few days ago and was on medication Doctor: Hello Patient: Hello Doctor: How many times a day Patient: Your baby is 4 months old. Doctor: cough, runny nose, fever not Patient: Recently, he has been vomiting Patient: occasional cough Patient: No fever Patient: occasional runny nose Doctor: How many times a day Patient: 4 or 5 times in the past few days Doctor: Is your baby breastfeeding? Doctor: How is breastfeeding? Patient: Yes, breast milk Patient: Yes Doctor: How is your mental state? Patient: Good Doctor: What kind of medicine are you using now? Patient: Golden Ibuprofen and Montelukast Doctor: Have you had any stool tests? Patient: And some small white pills Patient: No Doctor: How many days have you had diarrhea? Patient: 4 or 5 days Doctor: In this case, it is recommended to go to the hospital for a stool test to rule out enteritis.	{“主诉”: [“呕吐,” “咳嗽,” “流清涕”], “现病史”: [“呕吐,” “豆腐渣样,” “咳嗽,” “流清涕,” “金盆贝康,” “蒙脱石”], “诊断”: [“肠炎”], “建议”: [“医院,” “化验”]} {“complaints”: [“vomiting,” “cough,” “runny nose”], “medical history”: [“vomiting,” “tofu scum-like,” “cough,” “runny nose,” “Gold Ibuprofen,” “montelukast”], “diagnosis”: [“enterocolitis”], “advice”: [“hospital,” “laboratory”]}	主诉: “宝宝四个月，最近几天老吐豆腐渣那种，偶尔咳嗽，不发烧，偶尔流清鼻涕。” 现病史: “患儿四个月大，近期出现呕吐（豆腐渣样），偶尔咳嗽，无发烧，偶尔流清鼻涕，近四五天大便四五次，目前正在服用金盆贝康、蒙脱石和未说明名称的小颗白药。” 辅助检查: “未进行大便化验。” 既往史: “无特殊既往病史。” 诊断: “怀疑婴儿胃肠道疾病，需排除肠炎。” 建议: “建议去医院进行大便化验，以排除肠炎。继续母乳喂养，观察宝宝的精神状态和吃奶情况，注意宝宝的脱水症状。” Complaint: “The baby is 4 months old, the last few days old vomiting tofu dregs kind, occasional cough, no fever, occasional runny nose.” Present Medical History: “The child is 4 months old and has recently developed vomiting (tofu dregs kind), occasional cough, no fever, occasional runny nose, 4 or 5 bowel movements in the last 4 or 5 days, and is currently taking Gold Ibuprofen, Montelukast, and unspecified name of a small white pill.” SUPPORTING EXAMINATION: “No stool tests were performed.” PAST HISTORY: “No specific past medical history.” DIAGNOSIS: “Gastrointestinal disease is suspected in the infant and enterocolitis needs to be ruled out.” RECOMMENDATION: “It is recommended to go to the hospital for a stool test to rule out enteritis. Continue breastfeeding, observe the baby's mental status and milk intake, and watch for signs of dehydration.”

[a]Not available.

Discussion

Principal Findings

Leveraging the authoritative medical LLM evaluation system MedBench, we conducted an analytical assessment of current mainstream models in the health care field and compiled the incorrect responses from the top 10 models in the evaluation rankings.

The robustness score of 0.79 for health care safety and ethics under shuffled options permutations reveals concerning inconsistency in ethical decision-making scenarios. This could potentially be attributed to inadequate safety mechanisms or insufficiently comprehensive datasets concerning drug contraindications and medical ethics information [19]. For example, the model may generate incorrect drug recommendations, ignore the patient's allergy history or drug contraindications, and directly threaten patient safety. At the same time, the model may lack an understanding of medical ethics, such as patient privacy protection or fairness, resulting in output that does not meet specifications. Therefore, enhance model safety protocols by integrating rigorously curated medical cases and ethical scenarios, particularly those involving privacy protection, patient rights, and health care equity [20-24]. Embed international medical ethics guidelines and local legal frameworks as hard constraints for content generation. Strengthen pharmacological safety [24-26] through expanded datasets covering drug contraindications, interactions [27], and duplicate medication risks. Medical domain-specific models should prioritize these safety-critical considerations [23] during development and deployment.

In medical knowledge question answering, issues such as information omission (1644/3703, 44.4%) and hallucinations (868/3703, 23.4%) reveal gaps in the training corpus, which lacks comprehensive coverage of key medical scenarios. This leads to incomplete or fabricated outputs [28,29], such as omitting critical details or generating nonexistent content, posing risks in high-stakes medical contexts. To address this, models need domain-specific datasets, including medical textbooks and clinical guidelines, to improve reliability. In medical language understanding, omissions (1984/5784, 34.3%) and causal reasoning (1303/5784, 22.5%) deficiencies show models struggle to fully grasp questions, resulting in overinterpretation or insufficient inference [30,31]. For example, they may miss diagnostic criteria or fail to connect symptoms to diseases. Enhancing this requires better training data and architectural improvements, such as integrating causal reasoning modules or external knowledge graphs. In complex medical reasoning, omissions severely hinder performance. Tasks such as diagnosing patients require synthesizing multiple data points, but models often fail to integrate information effectively, leading to fragmented outputs [32]. For instance, they might overlook critical diagnoses or prioritize unlikely conditions. Improving this demands hierarchical reasoning frameworks or hybrid systems combining neural networks with symbolic logic alongside training on annotated case studies and real-world scenarios.

The improvement suggestions for medical LLMs can be structured into 4 levels based on technical complexity and expected impact, progressing from simpler, high-yield enhancements to more advanced, long-term innovations. At the first level, low-cost, high-return optimizations, the focus is on refining "surface-level" elements such as data quality, prompt engineering, and parameter fine-tuning [33,34]. For instance, cleaning noisy medical text, designing specialized prompt templates [34,35] for clinical scenarios, and using techniques such as Low-Rank Adaptation (LoRA) for efficient fine-tuning can significantly enhance baseline performance with minimal model modifications. Moving to the second level, domain-specific adaptation, the goal is to bolster the model's medical expertise through methods such as knowledge-augmented retrieval [36], multitask joint training [37], and ethical constraint integration. This includes linking to authoritative medical databases [38-40] for real-time evidence retrieval and training on diverse tasks such as diagnostic reasoning and medical record generation [41] to improve overall competency. At the third level, architectural advancements, the focus shifts to upgrading the model's structure or introducing hybrid systems to tackle complex medical challenges. Examples include combining symbolic logic with neural networks [42] to enhance diagnostic interpretability or designing modular reasoning frameworks [43-45] to minimize error propagation, thereby addressing tasks requiring advanced cognition or extended reasoning chains. Build real-time validation [46,47] pipelines combining rule-based filters and API-driven fact-checking. Finally, at the fourth level, cutting-edge exploration, the emphasis is on long-term technological innovation, such as multimodal pretraining that integrates text [48], imaging, and genomic data, transitioning from correlation-based learning to causal reasoning models [43,44], and leveraging digital twin technology to simulate virtual patient cohorts for training. While these efforts are technically demanding and may yield limited immediate benefits, they lay the groundwork for future breakthroughs in medical AI. This tiered improvement framework, progressing from simple to complex and from high-impact to foundational changes, not only addresses current model limitations but also charts a clear path toward developing reliable and trustworthy medical AI systems.

Limitations

While MedBench provides a standardized evaluation framework, relying solely on this benchmark may not fully capture the unstructured noise and ambiguity inherent in real-world clinical environments, potentially limiting the generalizability of our findings to diverse health care settings. Furthermore, this study represents a static snapshot of the top 10 performing models, excluding emerging architectures and specialized open-source tools that might exhibit different behavior. The manual development of the error taxonomy also introduces a degree of subjectivity, particularly when distinguishing between overlapping error types, such as causal reasoning deficiencies and hallucinations, in complex medical scenarios.

Additionally, the current evaluation is primarily restricted to textual modalities, leaving the models' capabilities in processing imaging or genomic data, critical components of modern diagnostics discussed in our future outlook, largely unexplored. Finally, the proposed 4-level optimization strategy remains a theoretical framework derived from our diagnostic insights. While this work identifies systemic weaknesses and prescribes an actionable road map, the empirical implementation and quantitative validation of these specific interventions fall outside the current scope, necessitating future research to verify their efficacy in enhancing clinical robustness and safety.

Conclusion

The current mainstream Chinese medical models face challenges across multiple dimensions, including medical knowledge question answering, language generation, and complex reasoning, while demonstrating room for improvement in safety mechanisms and ethical constraints. Based on a hierarchical improvement framework, we propose a progressive optimization pathway spanning from foundational data optimization and domain-specific knowledge enhancement to architectural innovation. These findings not only provide actionable directions for enhancing the clinical applicability of medical LLMs but also validate the benchmarking value of MedBench in advancing medical AI technologies. Future research should focus on exploring multimodal medical data integration and constructing causal reasoning mechanisms, thereby facilitating the leapfrog development of medical LLMs from knowledge association to clinical decision support systems.

Acknowledgments

Please note that no AI tools were used at any stage of the manuscript development.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets used in this study are available through the MedBench open platform. Access to the data can be obtained by contacting the MedBench team or the corresponding author.

Authors' Contributions

LJ, JC, and JX designed the study. JC, L Lu, and XP were responsible for writing the manuscript. JC collected and sorted out the data. JC performed the statistical analysis. L Liu and JH reviewed and edited the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of the evaluated medical large language models and prompt templates used.
[DOCX File (Microsoft Word File), 167 KB-Multimedia Appendix 1]

References

1. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
3. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: progress, application, and challenge. arXiv. Preprint posted online on Dec 11, 2023. [doi: 10.48550/arXiv.2311.05112]
4. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health. May 2020;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
5. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. arXiv. Preprint posted online on Nov 16, 2022. [doi: 10.48550/arXiv.2211.09110]
6. Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv. Preprint posted online on Sep 7, 2020. [doi: 10.48550/arXiv.2009.03300]
7. Huang Y, Bai Y, Zhu Z, et al. C-eval: a multi-level multi-discipline Chinese evaluation suite for foundation models. Preprint posted online on Nov 6, 2023. [doi: 10.48550/arXiv.2305.08322]
8. Liu M, Hu W, Ding J, et al. MedBench: a comprehensive, standardized, and reliable benchmarking system for evaluating Chinese medical large language models. Big Data Min Anal. 2024;7(4):1116-1128. [doi: 10.26599/BDMA.2024.9020044]

9. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
10. Li'evin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? arXiv. Preprint posted online on Dec 20, 2022. [doi: 10.48550/arXiv.2207.08143]
11. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. Apr 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
12. Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress. Science. May 24, 2024;384(6698):842-845. [doi: 10.1126/science.adn0117] [Medline: 38768279]
13. Group (2025) AntAngelMed: a large-scale medical MOE model. Hugging Face. URL: https://huggingface.co/MedAIBase/AntAngelMed [Accessed 2026-07-08]
14. Wang G, Gao M, Yang S, et al. Citrus: leveraging expert cognitive pathways in a medical language model for advanced medical decision support. arXiv. Preprint posted online on Feb 26, 2025. [doi: 10.48550/arXiv.2502.18274]
15. Liu Z, Hou Z, Zhu G, Sang Z, Xie C, Yang H. InfiMed: low-resource medical mllms with advancing understanding and reasoning. arXiv. Preprint posted online on Oct 8, 2025. [doi: 10.48550/arXiv.2505.23867]
16. Wang Z, Liu X, Yao Y, et al. Technical report of TeleChat2, TeleChat2.5 and T1. arXiv. Preprint posted online on Jul 29, 2025. [doi: 10.48550/arXiv.2507.18013]
17. Sun X, Chen Y, Huang Y, et al. Hunyuan-Large: an open-source MOE model with 52 billion activated parameters by Tencent. arXiv. Preprint posted online on Nov 6, 2024. [doi: 10.48550/arXiv.2411.02265]
18. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Dec 19, 2023. [doi: 10.48550/arXiv.2303.08774]
19. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. J Med Internet Res. Aug 1, 2024;26:e60083. [doi: 10.2196/60083] [Medline: 38971715]
20. Wang H, Fu W, Tang Y, et al. A survey on responsible LLMs: inherent risk. arXiv. Preprint posted online on Jan 16, 2025. [doi: 10.48550/arXiv.2501.09431]
21. Zhou J, Li H, Liao X, et al. A unified method to revoke the private data of patients in intelligent healthcare with audit to forget. Nat Commun. 2023;14(1):6255. [doi: 10.1038/s41467-023-41703-x]
22. Haque MA, Siddique S, Rahman MM, et al. SOK: exploring hallucinations and security risks in AI-assisted software development with insights for LLM deployment. Presented at: 2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA); Sep 1-4, 2025; Varna, Bulgaria. [doi: 10.1109/IDSTA66210.2025.11202778]
23. Chandra M, Sriraman S, Verma G, et al. Lived experience not found: LLMs struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. arXiv. Preprint posted online on Oct 24, 2024. [doi: 10.48550/arXiv.2410.19155]
24. De Vito G, Ferrucci F, Angelakis A. HELIOT: LLM-based CDSS for adverse drug reaction management. Preprint posted online on Sep 24, 2024. [doi: 10.48550/arXiv.2409.16395]
25. Sahoo P, Singh AK, Saha S, Chadha A, Mondal S. Enhancing adverse drug event detection with multimodal dataset: corpus creation and model development. Presented at: Findings of the Association for Computational Linguistics: ACL 2024; Aug 11-16, 2024; Bangkok, Thailand. URL: https://aclanthology.org/2024.findings-acl.667/ [Accessed 2026-06-28]
26. De Vito G, Ferrucci F, Angelakis A. LLMs for drug-drug interaction prediction: a comprehensive comparison. arXiv. Preprint posted online on Feb 9, 2025. [doi: 10.48550/arXiv.2502.06890]
27. McCoy RT, Pavlick E, Linzen T. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy. [doi: 10.18653/v1/P19-1334]
28. Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. arXiv. Preprint posted online on Apr 22, 2019. [doi: 10.48550/arXiv.1904.09751]
29. Wang L, Shen Y. Evaluating causal reasoning capabilities of large language models: a systematic analysis across three scenarios. Electronics (Basel). 2024;13(23):4584. [doi: 10.3390/electronics13234584]
30. Joshi N, Saparov A, Wang Y, He H. LLMs are prone to fallacies in causal inference. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024; Miami, Florida. [doi: 10.18653/v1/2024.emnlp-main.590]
31. Kerner T. Domain-specific pretraining of language models: a comparative study in the medical field. arXiv. Preprint posted online on Jul 19, 2024. [doi: 10.48550/arXiv.2407.14076]

32. Ni S, Bi K, Guo J, Yu L, Bi B, Cheng X. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 27 to Aug 1, 2025; Vienna, Austria. [doi: 10.18653/v1/2025.acl-long.1184]
33. Liu Y, Xu J, Zhang LL, et al. Beyond prompt content: enhancing LLM performance via content-format integrated prompt optimization. arXiv. Preprint posted online on May 21, 2025. [doi: 10.48550/arXiv.2502.04295]
34. Lin Y, Hasan M, Kosalge R, et al. Visual template inference for data extraction from documents. arXiv. Preprint posted online on Jan 11, 2025. [doi: 10.48550/arXiv.2501.06659]
35. Kovács Á, Recski G. LettuceDetect: a hallucination detection framework for RAG applications. arXiv. Preprint posted online on Feb 24, 2025. [doi: 10.48550/arXiv.2502.17125]
36. Toma A, Xie R, Palayew S, Lawler P, Wang B. WangLab at MEDIQA-CORR 2024: optimized LLM-based programs for medical error detection and correction. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. URL: https://aclanthology.org/2024.clinicalnlp-1.59/ [Accessed 2026-06-28]
37. Qin L, Zhang Y, Liang H, Jatowt A, Yang Z. Listening to patients: detecting and mitigating patient misreport in medical dialogue system. Presented at: Findings of the Association for Computational Linguistics: ACL 2025; Jul 27 to Aug 1, 2025; Vienna, Austria. 2024.URL: https://aclanthology.org/2025.findings-acl.135/ [Accessed 2026-06-28]
38. Liu W, Tang J, Cheng Y, Li W, Zheng Y, LiangX. MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. Presented at: Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022; Sep 24-25, 2022; Guilin, China. [doi: 10.1007/978-3-031-17120-8_35]
39. Li R, Wang X, Yu H, et al. Exploring LLM multi-agents for ICD coding. arXiv. Aug 14, 2024. [doi: 10.48550/arXiv.2406.15363]
40. Shen C, Chen Z, Luo D, Xu D, Chen H, Ni J. Exploring multi-modal data with tool-augmented LLM agents for precise causal discovery. Presented at: Findings of the Association for Computational Linguistics: ACL 2025; Jul 27 to Aug 1, 2024; Vienna, Austria. URL: https://aclanthology.org/2025.findings-acl.36/ [Accessed 2026-07-08]
41. Wang R, Sun K. DSPy-based neural-symbolic pipeline to enhance spatial reasoning in LLMs. Neural Netw. Jan 2026;193:108022. [doi: 10.1016/j.neunet.2025.108022] [Medline: 40939460]
42. Cui Y, He P, Zeng J, et al. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. Presented at: Findings of the Association for Computational Linguistics: ACL 2025; Jul 27 to Aug 1, 2025; Vienna, Austria. URL: https://aclanthology.org/2025.findings-acl.956/ [Accessed 2026-06-28]
43. Wu Z, Zeng Q, Zhang Z, Tan Z, Shen C, Jiang M. Enhancing mathematical reasoning in LLMs by stepwise correction. arXiv. Preprint posted online on Oct 16, 2024. [doi: 10.48550/arXiv.2410.12934]
44. Zhong L, Du Z, Zhang X, Hu H, Tang J. ComplexFuncBench: exploring multi-step and constrained function calling under long-context scenario. arXiv. Preprint posted online on Jan 17, 2025. [doi: 10.48550/arXiv.2501.10132]
45. Yang S, Sun R, Wan X. A new benchmark and reverse validation method for passage-level hallucination detection. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023; Singapore. URL: https://aclanthology.org/2023.findings-emnlp.256/ [Accessed 2026-06-28]
46. Varshney N, Yao W, Zhang H, Chen J, Yu D. A stitch in time saves nine: detecting and mitigating hallucinations of LLMs by validating low-confidence generation. arXiv. Preprint posted online on Aug 12, 2023. [doi: 10.48550/arXiv.2307.03987]
47. Liu W, Ma X, Zhu Y, et al. Sliding windows are not the end: exploring full ranking with long-context large language models. arXiv. Preprint posted online on Dec 19, 2024. [doi: 10.48550/arXiv.2412.14574]
48. Xia Y, Dittler D, Jazdi N, Chen H, Weyrich M. LLM experiments with simulation: large language model multi-agent system for simulation model parametrization in digital twins. Presented at: 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA); Sep 10-13, 2024:1-4; Padova, Italy. [doi: 10.1109/ETFA61755.2024.10710900]

Abbreviations

LLM: large language model
LoRA: Low-Rank Adaptation

Edited by Yanshan Wang; peer-reviewed by Minghan Li, Shuchu Han; submitted 31.Oct.2025; final revised version received 11.Jun.2026; accepted 12.Jun.2026; published 15.Jul.2026

Please cite as:
Jiang L, Chen J, Lu L, Peng X, Liu L, He J, Xu J

Performance Gaps and Optimization Strategies in Chinese Medical Large Language Models Based on MedBench: Evaluation Study
JMIR AI 2026;5:e86864
URL: https://ai.jmir.org/2026/1/e86864
doi: 10.2196/86864