CMedCalc-Bench: A Fine-Grained Benchmark for Chinese Medical Calculations in LLMs

Yunyan Zhang[†] **Zhihong Zhu**[†] **Xian Wu**[*]
Tencent Jarvis Lab
{yunyanzhang, profzhu, kevinxwu}@tencent.com

Abstract

Large Language Models (LLMs) have demonstrated significant potential in medical diagnostics and clinical decision-making. While benchmarks such as MedQA and PubMedQA have advanced the evaluation of qualitative reasoning, existing medical NLP benchmarks still face two limitations: the absence of a Chinese benchmark for medical calculation tasks, and the lack of fine-grained evaluation of intermediate reasoning. In this paper, we introduce CMedCalc-Bench, a new benchmark designed for Chinese medical calculation. CMedCalc-Bench covers 69 calculators across 12 clinical departments, featuring over 1,000 real-world patient cases. Building on this, we design a fine-grained evaluation framework that disentangles clinical entity extraction from numerical computation, enabling systematic diagnosis of model deficiencies. Experiments across four model families, including medical-specialized and reasoning-focused, provide an assessment of their strengths and limitations on Chinese medical calculation. Furthermore, explorations on faithful reasoning and the demonstration effect offer early insights into advancing safe and reliable clinical computation.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency across diverse biomedical tasks (Wu et al., 2024), including medical knowledge retrieval, diagnostic reasoning, and clinical decision-making. Existing biomedical NLP benchmarks such as PubMedQA (Jin et al., 2019) and MedQA (Jin et al., 2021) predominantly focus on qualitative reasoning and textual comprehension. However, these benchmarks largely overlook quantitative computational tasks, thus limiting the applicability of LLMs in clinical scenarios where precise numerical calculations are fundamental.

Medical calculators are widely used by healthcare providers to support clinical decisions through quantitative assessments, directly influencing patient care quality and clinical outcomes (Green et al., 2019). Despite their widespread use, the ability of LLMs to reliably perform medical calculations remains underexplored. Initial benchmarks such as MedCalcBench (Khandekar et al., 2024), AgentMD (Jin et al., 2024), and OpenMedCalc (Goodell et al., 2023) have begun to address this gap but face two critical challenges: (1) the absence of a Chinese benchmark for medical calculation tasks, which leaves a major linguistic and clinical coverage gap; and (2) the lack of fine-grained evaluation of intermediate reasoning processes, as most benchmarks only assess final outputs without diagnosing where models fail (Zhu et al., 2025a,b).

To address these challenges, we introduce CMedCalc-Bench, the first Chinese clinical calculation benchmark designed for rigorous evaluation of medical LLMs. CMedCalc-Bench covers 69 clinically significant calculation tasks across 12 medical specialties, featuring 1,143 real-world cases. Crucially, it incorporates a four-stage evaluation framework that separately examines knowledge acquisition, parameter extraction, unit conversion, and calculation or comprehension, enabling systematic diagnosis of model deficiencies.

Experiments are carried out across four representative model families: open-source foundation, medical-specialized, advanced proprietary, and reasoning-focused. The results reveal substantial performance gaps across categories and task types. Reasoning-focused models achieve relative gains, yet all models display cascading errors.

Beyond computational accuracy, CMedCalc-Bench also considers safety. The Faithful Reasoning analysis assesses whether models can abstain when confronted with missing or contradictory inputs. The Demonstration Effect study further examines how exemplar choice influences refusal be-

[†]Equal contribution.
[*]Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25662–25671
November 4-9, 2025

	Lang.	Medical	Qual. Reasoning	Quant. Reasoning	Open-ended	FG-Eval
MedQA (Jin et al., 2021)	en	✔	✔	✗	✗	✗
MedMCQA (Pal et al., 2022)	en	✔	✔	✗	✗	✗
PubMedQA (Jin et al., 2019)	en	✔	✔	✗	✗	✗
MMLU (Hendrycks et al., 2020)	en	✔	✔	✗	✗	✗
MedJourney(Wu et al., 2024)	zh	✔	✔	✗	✔	✗
OlymMATH(Sun et al., 2025)	en&zh	✗	✗	✔	✔	✗
GSM8k (Cobbe et al., 2021)	en	✗	✗	✔	✔	✗
MATH (Hendrycks et al., 2021)	en	✗	✗	✔	✔	✗
MedCalc-Bench (Khandekar et al., 2024)	en	✔	✔	✔	✔	✗
OpenMedCalc (Goodell et al., 2023)	en	✔	✔	✔	✔	✗
AgentMD (Jin et al., 2024)	en	✔	✔	✔	✔	✗
CalcQA (Zhu et al., 2025a)	en	✔	✔	✔	✔	✗
CMedCalc-Bench (Ours)	**zh**	✔	✔	✔	✔	✔

Table 1: Comparison of the proposed CMedCalc-Bench with existing related benchmarks. "Lang." denotes the language focused on; "Qual." and "Quant." indicate qualitative and quantitative reasoning, respectively; "Open-ended" tasks require free-form answers; "FG-Eval" denotes whether fine-grained evaluation is supported.

havior, showing that unanswerable demonstrations substantially improve safe abstention.

In summary, our contributions are as follows: (1) establishing the first Chinese benchmark[1] tailored explicitly for clinical calculation tasks; (2) introducing a fine-grained evaluation strategy to pinpoint different computational weaknesses; and (3) providing extensive empirical analysis to clarify current LLM limitations and inform future research in Chinese medical computational capabilities.

2 Related Work

Most existing benchmarks for medical LLMs focus on multiple-choice questions. In English, MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU (Medical) Series (Hendrycks et al., 2021) are widely used. In Chinese, MedJourney (Wu et al., 2024) extends this setup by evaluating patient journeys with both multiple-choice and open-ended formats.

Beyond clinical evaluation, researchers have built datasets to measure mathematical calculation. For instance, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) cover problems from elementary to advanced levels, while OlympiadMATH (Sun et al., 2025) raises the difficulty to Olympiad-style tasks that require complex multi-step reasoning. Recently, clinical evaluation has moved toward combining LLMs with external calculators. AgentMD (Jin et al., 2024) and OpenMedCalc (Goodell et al., 2023) use medical calculators to support quantitative reasoning. CalcQA (Zhu et al., 2025a) builds on this with 100 calculator pairs derived from patient cases. Khandekar et al. (2024) further contribute annotated reasoning chains for over 1,000 clinician-validated cases.

Despite these advances, existing benchmarks still fall short in covering Chinese medical calculations. To this end, we introduce CMedCalc-Bench, a new fine-grained benchmark for Chinese medical calculations. Table 1 summarizes the differences between related benchmarks and ours.

3 CMedCalc-Bench

3.1 Task Categorization

In CMedCalc-Bench, we curated 69 calculators from the Medical Calculators of the Clinical Assistant of the People's Medical Publishing House,[2] which are widely adopted in medical practice across different departments. These calculators can be grouped into three categories: equation-based calculators (37), rule-based accumulators (20), and semantic-based quantifiers (12). Figure 1 presents example instances from each category.

➠ Equation-based calculators process numerical data (*e.g.*, age and weight) and categorical inputs (*e.g.*, gender) through predefined mathematical formulas to generate precise quantitative outputs. These outputs are typically continuous decimals representing medical parameters. ➠ Rule-based accumulators evaluate categorical inputs (*e.g.*, clinical criteria) and numerical data (*e.g.*, age thresh-

[1] https://github.com/Zhihong-Zhu/CMedCalc-Bench

[2] https://ccdas.pmphai.com/appformula/toPcIndex

Figure 1: Example instances of the proposed CMedCalc-Bench dataset.

olds) to generate discrete scores through additive rules. Each condition or criterion contributes a predefined point value, with the final sum categorizing risk or severity. Unlike equation-based calculators, rule-based accumulators prioritize clinical judgment codified into incremental scoring rather than mathematical formulas. Semantic-based quantifiers analyze qualitative clinical data, such as *imaging reports, pathology descriptions*, or *symptom narratives*, to generate quantitative classifications. Unlike equation-based calculators or rule-based accumulators, they interpret unstructured information to assign grades or risk tiers. For example, the NYHA Functional Classification categorizes heart failure severity based on symptom descriptions.

3.2 Data Collection

In this subsection, we explain how we built the dataset for the 69 calculation tasks in CMedCalc-Bench. We describe the process for collecting patient notes below, which followed three main steps.

Knowledge Preparation and Notes Retrieval. We first listed the attributes required by each of the 69 calculators and standardized their units. For equation-based calculators, we implement the original formulas; for rule-based accumulators and semantic-based quantifiers, we compiled the scoring and grading criteria from official guidelines.

To obtain patient notes, we collect anonymized records[3] from two widely used Chinese medical platforms. In total, we retrieve 37,149 patient notes. To maximize recall, each attribute is expanded into a synonym set of about three terms on average, and regular expressions are applied to capture diverse expressions in the clinical narratives. After filtering, 46 calculators remain with at least one matched note containing the required attributes.

Attribute Extraction and Answer Generation. For equation-based calculators, attribute values are extracted from patient notes and directly substituted into predefined formulas. For rule-based accumulators, scores are derived according to official scoring guidelines, with `GPT-4o` assisting in generating step-by-step reasoning that is subsequently checked against documented cases. The same workflow is applied to semantic-based quantifiers, where guideline-aligned entities are identified and mapped to grading criteria, and model outputs are further verified through manual review.

Data Verification and Expansion. We engaged three physicians to perform data verification. Each extracted case was first checked by one doctor for the correctness of attributes and answers. Another doctor ensured that the final answer did not appear verbatim in the note and removed sensitive identifiers such as names and hospitals. A meta-annotator conducted the final review and selected up to 20 high-quality notes for each calculator.

After verification, 46 calculators retained at least 5 eligible notes. Some calculators, particularly rule-based accumulators, had very few matches because the required subjective criteria were rarely

[3] https://www.iiyi.com/; https://www.dxy.cn/bbs/newweb/pc/case

Figure 2: Department diversity in the proposed CMedCalc-Bench. Each color corresponds to one high-level subject: General Assessment & Support, Critical & Systemic Care, and Organ Systems & Specialties. For visual clarity, only the most frequent classes are shown.

documented. To address this, we synthesized 331 additional cases by adapting translated examples from MedCalc-Bench (Khandekar et al., 2024).

Quality Control. Two primary annotators first independently labeled the entire dataset, yielding a Cohen's Kappa of $\kappa = 0.85$, indicating almost perfect agreement (Landis and Koch, 1977). To finalize the labels, a senior annotator then performed a full review, which involved adjudicating all 186 disagreements and additionally verifying all instances on which the primary annotators had agreed.

3.3 Data Analysis

Key Statistics. Figure 3 summarizes statistics of the proposed CMedCalc-Bench across different calculator subtypes, reporting the number of indicators, instances, average note length, and attribute complexity. The dataset covers 69 medical calculators, each containing 5–20 instances, resulting in a total of 1,143 instances. Each instance consists of: (1) the calculator name, (2) a patient note, (3) the ground-truth answer computed by the corresponding calculator, and (4) the calculation process, including extracted clinical entities (*e.g.*, lab values, vital signs) and step-by-step explanatory reasoning.

Department Diversity. The calculators in the proposed CMedCalc-Bench span 12 departments, as shown in Figure 3. These departments are further grouped into three broader categories: General Assessment & Support, Critical & Systemic Care, and Organ Systems & Specialties. Figure 2 illustrates the hierarchical structure of categories, departments, and calculators, highlighting the broad diversity encompassed by CMedCalc-Bench.

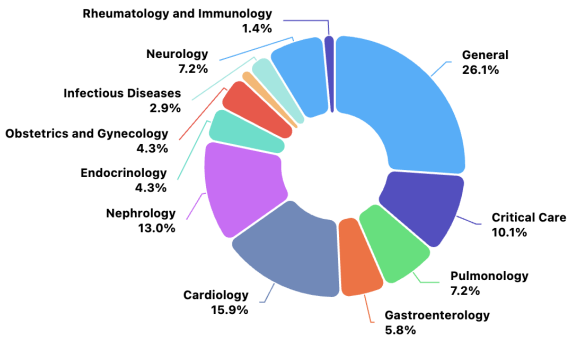
4 Evaluation

4.1 Settings

We have chosen four types of LLMs for evaluation: (1) Open-source foundation LLMs, including `Llama 3.1-8b` (Grattafiori et al., 2024) and `Qwen 2.5` (Yang et al., 2025), with parameter sizes ranging from 7b to 32b; (2) Medical specialized LLMs, including `HuatuoGPT-o1-7B` (Chen et al., 2024) and `Baichuan-M1-14B` (Wang et al., 2025); (3) Advanced proprietary LLMs, in-

Model	Equation-based			Rule-based			Semantic-based		
	D	ZC	OC	D	ZC	OC	D	ZC	OC
Open-Source Foundation Models									
Llama 3.1-8b (Grattafiori et al., 2024)	28.38	26.15	38.29	11.61	14.65	22.22	31.48	33.33	48.77
Qwen-2.5-7b (Yang et al., 2025)	37.95	35.38	43.08	19.44	14.39	26.01	40.12	43.21	57.41
Qwen-2.5-14b (Yang et al., 2025)	35.56	36.24	46.67	18.69	20.71	27.27	32.72	44.44	64.20
Qwen-2.5-32b (Yang et al., 2025)	39.83	42.22	51.62	16.41	16.41	31.82	34.57	37.04	67.28
Advanced Proprietary Models									
DeepSeek-V3 (Liu et al., 2024)	46.84	52.65	**66.50**	18.94	22.73	39.65	45.68	51.23	65.43
GPT-4o (Hurst et al., 2024)	39.32	49.40	59.66	20.45	26.26	41.16	39.51	52.47	68.52
Medical Specialized Models									
HuatuoGPT-o1-7B (Chen et al., 2024)	31.97	36.41	41.20	19.19	12.63	23.48	43.21	38.89	58.64
Baichuan-M1-14B (Wang et al., 2025)	30.94	45.81	54.53	20.96	21.97	35.61	34.57	45.06	63.58
Reasoning-focused Models									
DeepSeek-R1 (Guo et al., 2025)	**53.60**	**56.24**	65.64	37.37	34.34	45.71	**64.20**	**64.20**	**73.46**
o1 (Jaech et al., 2024)	52.31	48.55	64.44	**39.90**	**36.11**	**46.46**	57.41	56.79	72.22

Table 2: Performance across different models and prompting strategies for equation-based, rule-based, and semantic-based calculators. Bold denotes the best performance. D: Direct prompting, ZC: Zero-shot CoT, OC: One-shot CoT.



	Equation -based	Rule -based	Semantic -based	Overall
#Indicators	37	20	12	69
#Instances	585	396	162	1143
Avg. L of Note	1495.3	258	209.7	884.4
Min Attr.	1	1	–	1
Max Attr.	7	31	–	31
Avg. Attr.	2.8	9.8	–	5.6

Figure 3: Overview of the proposed CMedCalc-Bench dataset. The top figure shows the distribution of calculators across clinical departments. The table below summarizes key dataset statistics, including the number of indicators, instances, average note length, and attribute complexity across three calculator types.

cluding `GPT-4o` (Hurst et al., 2024) and `DeepSeek-v3` (Liu et al., 2024); (4) Reasoning-focused LLMs including `o1` (Jaech et al., 2024) and `DeepSeek-R1` (Guo et al., 2025).

Following Khandekar et al. (2024), we similarly investigate three prompting strategies: (1) *Zero-shot Direct Prompting*: the model directly outputs answers without explanations; (2) *Zero-shot Chain-of-Thought (CoT) Prompting*: the model first generates step-by-step reasoning (Wei et al., 2022) before producing the final answer; (3) *One-shot CoT Prompting*: the model is provided with a manually curated exemplar consisting of a patient note, the calculation task name, and the expected output with explanatory steps and final answer value.

We adopt accuracy as the evaluation metric. For equation-based calculations, we enforce exact-match requirements for clinical date-related tasks (*e.g.*, estimated due dates) but permit a ±5% tolerance for other numerical outputs. In contrast, rule-based and semantic-based tasks maintain strict exact-match criteria across all evaluation instances.

4.2 Main Results

The main results are shown in Table 2, from which we can obtain the following observations:

(1) Performance stratification across model types. Reasoning-oriented architectures achieve the best overall performance, with `DeepSeek-V3` leading on equation-based tasks, `o1` on rule-based tasks, and `DeepSeek-R1` on semantic tasks. Proprietary models form the second tier, while open-source foundation models rank third, where larger scales generally yield stronger results in line with scaling laws. Domain-specialized models outperform same-scale open-source models, highlighting the advantage of medical knowledge integration.

(2) Prompting strategies also affect outcomes. One-shot COT prompting consistently improves accuracy, with additional gains from external medical demonstrations. Zero-shot COT shows mixed out-

Model	Knowledge Acquisition	Parameter Extraction	Unit Conversion	Calculation /Comprehension
Equation-based				
Qwen-2.5-32B (Yang et al., 2025)	55.56 / 62.05	54.19 / 60.51	53.85 / 60.51	42.22 / 51.62
DeepSeek-V3 (Liu et al., 2024)	61.37 / 71.79	59.49 / 71.11	59.32 / 71.11	52.65 / 66.50
GPT-4o (Hurst et al., 2024)	60.00 / 69.05	58.12 / 68.38	57.09 / 68.21	49.40 / 59.66
o1 (Jaech et al., 2024)	67.86 / 75.38	65.47 / 74.52	64.79 / 73.68	48.55 / 64.44
DeepSeek-R1 (Guo et al., 2025)	68.55 / 74.02	66.15 / 73.68	65.64 / 73.68	56.24 / 65.64
Rule-based				
Qwen-2.5-32B (Yang et al., 2025)	32.07 / 42.42	29.55 / 41.67	-	16.41 / 31.82
DeepSeek-V3 (Liu et al., 2024)	40.66 / 46.97	39.14 / 44.94	-	22.73 / 39.65
GPT-4o (Hurst et al., 2024)	47.47 / 47.22	45.71 / 46.46	-	26.26 / 41.16
o1 (Jaech et al., 2024)	53.03 / 53.79	50.00 / 52.78	-	36.11 / 46.46
DeepSeek-R1 (Guo et al., 2025)	52.78 / 52.78	51.76 / 51.52	-	34.34 / 45.71
Semantic-based				
Qwen-2.5-32B (Yang et al., 2025)	55.56 / 76.54	-	-	37.04 / 67.28
DeepSeek-V3 (Liu et al., 2024)	66.67 / 75.31	-	-	51.23 / 65.43
GPT-4o (Hurst et al., 2024)	70.99 / 75.93	-	-	52.47 / 68.52
o1 (Jaech et al., 2024)	75.31 / 85.19	-	-	56.79 / 72.22
DeepSeek-R1 (Guo et al., 2025)	75.93 / 83.95	-	-	64.20 / 73.46

Table 3: Fine-grained performance (Zero-shot / One-shot) of models.

comes: reasoning-focused models degrade on rule-based and semantic tasks, likely due to excessive reasoning traces, whereas other models improve, indicating explicit reasoning steps are especially useful for advanced proprietary models.

(3) Task-level analysis highlights gaps. Overall, LLMs achieve the highest performance on semantic-based tasks, moderate performance on equation-based problems, and the lowest on rule-based calculations. The strong results on semantic-based tasks suggest that LLMs already possess adequate medical knowledge, while persistent errors on rule-based tasks expose weaknesses in handling medical scales and operational rules.

5 Discussion

5.1 Fine-Grained Analysis

Unlike prior benchmarks that primarily evaluate final outputs, CMedCalc-Bench introduces a four-stage framework that explicitly examines intermediate reasoning steps to identify where errors arise. The framework consists of: (1) *Knowledge Acquisition*, assessing the ability to recall and contextualize equations or guidelines; (2) *Parameter Extraction*, measuring precision in identifying variables from patient notes; (3) *Unit Conversion*, testing accuracy in numerical standardization; and (4) *Calculation/Comprehension*, evaluating the correctness of the final output or classification. For equation-based calculators, all four stages are applied sequentially; rule-based calculators omit Unit Conversion; and semantic calculators conclude after Knowledge Acquisition and direct classification. Crucially, errors propagate across stages: a failure in an early step invalidates subsequent operations.

Inspired by Arora et al. (2025), we employ `GPT-4o` (Hurst et al., 2024) to evaluate CoT outputs according to this four-stage framework. Each output is scored in a binary fashion (1 if the stage is satisfied, 0 otherwise), and stage-level accuracies are then computed. After manual verification confirmed the reliability of this procedure, `GPT-4o` was adopted for full evaluation across the dataset.

The fine-grained performance across reasoning steps is shown in Table 3, leading to three main observations: ***(1) One-shot exemplars effectively bridge knowledge gaps.*** Across models and task types, the one-shot COT setting consistently surpasses the zero-shot setting. This confirms that in-context exemplars can provide the necessary background equations or classification criteria, compensating for deficits in specialized medical knowledge. ***(2) Models excel at preliminary reasoning.*** Accuracy is generally high in Knowledge Acquisition, especially for equation- and semantic-based tasks. The subsequent steps of Parameter Extraction and Unit Conversion show only marginal drops, indicating that LLMs can reliably identify and extract key information. ***(3) Final calculation and comprehension remain the bottleneck.***

Model	D	ZC	OC
Open-Source Foundation Models			
Qwen-2.5-32B (Yang et al., 2025)	45.64	75.81	56.11
Advanced Proprietary Models			
DeepSeek-V3 (Liu et al., 2024)	49.88	66.33	55.11
GPT-4o (Hurst et al., 2024)	50.12	52.62	47.63
Medical Specialized Models			
HuatuoGPT-o1-7B (Chen et al., 2024)	28.18	53.87	21.20
Baichuan-M1-14B (Wang et al., 2025)	5.73	58.60	34.66
Reasoning-focused Models			
DeepSeek-R1 (Guo et al., 2025)	58.60	55.11	50.87
o1 (Jaech et al., 2024)	32.42	40.40	44.39

Table 4: Faithful reasoning performance of models under different prompting strategies.

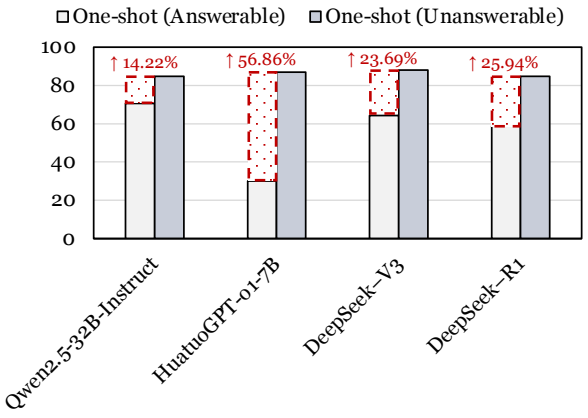


Figure 4: Refusal accuracy under different demonstrations on the uncomputable subtest.

The most substantial decline occurs at the Calculation/Comprehension stage, where errors in multi-parameter computations or misjudgments of classification criteria lead to failure. This highlights computational precision and classification accuracy as the critical challenges for current LLMs.

5.2 Faithful Reasoning

In real-world clinical notes, missing or contradictory attributes frequently hinder medical calculators. Forcing models to output results in such cases risks clinically misleading conclusions. To address this, we perform a faithful reasoning analysis that evaluates model behavior on uncomputable inputs. Concretely, we constructed a dedicated test set of 400 uncomputable cases. These span *equation-based* (200), *rule-based* (100), and *semantic-based* (100) tasks. Each instance contains either absent parameters or internal contradictions and is paired with an expert-authored rationale for refusal.

As shown the results in Table 4, `DeepSeek-R1` achieves the highest refusal rate (58.60%) under direct prompting, reflecting stronger intrinsic safeguards. Zero-shot CoT further improves refusal accuracy, suggesting that explicit reasoning helps expose missing information. In contrast, one-shot CoT consistently reduces refusal rates, even though it enhances accuracy on valid tasks. Overall, the results indicate a tension between accuracy-oriented prompting and reliable refusal. Methods that improve task accuracy can simultaneously weaken a model's ability to abstain when necessary.

5.3 Demonstration Effect

We further explore how exemplar choice shapes refusal behavior. Specifically, we compared one-shot prompts with an answerable exemplar versus an unanswerable exemplar on a dedicated test set of 400 uncomputable cases. As shown in Figure 4, all models achieved markedly higher refusal accuracy with unanswerable demonstrations. For instance, DeepSeek-V3 improved from 64.34% to 88.03%, while HuatuoGPT-o1 rose from 30.17% to 87.03%. These results suggest that answerable exemplars implicitly encourage over-generation, whereas unanswerable exemplars provide a stronger inductive bias toward abstention.

5.4 Case Study

To illustrate the failure modes of LLMs in medical calculations, we highlight representative error cases from our fine-grained evaluation on GPT-4o in Figure 5. The model often falters at the outset, committing basic errors like recalling a flawed Winters formula, misinterpreting "no murmurs" as anything but a Levine Grade 0, or overlooking critical lab values during extraction. Such early missteps expose a fragile reasoning chain, where one slip can derail the entire clinical assessment.

The most serious failures appear in the final comprehension stage. For instance, the model may correctly detect a gallop rhythm but misclassify it as Killip Grade I, thereby underestimating the severity of heart failure. This is not a trivial arithmetic slip but a gap in clinical interpretation, underscoring the models' shallow grasp of what symptoms signify.

6 Conclusion

In this paper, we presented CMedCalc-Bench, a new fine-grained benchmark for evaluating Chinese medical calculations in LLMs. CMedCalc-Bench introduces a four-stage evaluation protocol consisting of *knowledge acquisition, parameter extraction,*

	Equation-based	Rule-based	Semantic-based
Error in Knowledge Acquisition	LLM: Winters公式为：ΔPaCO2=1.2×HCO3-+15±2。将患者的HCO3-值代入公式中，我们得到： ΔPaCO2=1.2×16.1+15=19.32+15=34.32。因此，预期的代偿性PaCO2范围是34.32±2，即[32.32,36.32]mmHg。 The Winters formula is: ΔPaCO2 = 1.2 × HCO3- + 15 ± 2. Substituting the patient's HCO3- value into the formula, we get: ΔPaCO2 = 1.2 × 16.1 + 15 = 19.32 + 15 = 34.32. Therefore, the expected compensatory PaCO2 range is 34.32 ± 2, i.e., [32.32, 36.32] mmHg. Ground-truth: Winters公式为ΔPaCO2(mmHg)=1.5×[HCO3-]+8±2。由于本例HCO3-约为16.1mmol/L，计算为：1.5×16.1+8=24.15+8=32.15mmHg。结合±2的范围，期望PaCO2大约是30.15–34.15mmHg The Winters formula is ΔPaCO2 (mmHg) = 1.5 × [HCO3-] + 8 ± 2. Given that in this case HCO3- is approximately 16.1 mmol/L, the calculation is: 1.5 × 16.1 + 8 = 24.15 + 8 = 32.15 mmHg. Considering the ±2 range, the expected PaCO2 is approximately 30.15–34.15 mmHg.	LLM: 无法准确计算PSI评分 Unable to accurately calculate PSI score Ground-truth: 1.**年龄计算**：PSI评分中女性年龄按（实际年龄-10）计算，75岁女性得分为75-10=65分。[...] 4.**其他指标评估**：呼吸频率、血压、BUN、血糖等均未达加分阈值（如BUN=27＜30，血糖=180＜250）。5.**总分计算**：65（年龄）+10（养老院）+10（心衰）+10（胸腔积液）=95分。6.**风险分级**：95分对应PSIIV级（91-130分）， Age Calculation: In the PSI score, the age for females is calculated as (actual age - 10). A 75-year-old female scores 75 - 10 = 65 points. [...]. Other Indicators Assessment: Respiratory rate, blood pressure, BUN, blood glucose, etc., do not reach the threshold for additional points (e.g., BUN = 27 < 30, blood glucose = 180 < 250). Total Score Calculation: 65 (age) + 10 (nursing home) + 10 (heart failure) + 10 (pleural effusion) = 95 points. Risk Classification: 95 points correspond to PSI Class IV (91-130 points).	Patient Notes: 外伤性骨折，心脏各瓣膜区未闻及杂音 [...] Traumatic fracture, no murmurs heard in any cardiac valve area [...] LLM: 病历中没有提供心脏杂音的相关信息，因此无法应用Levine6级分级法进行评估。 The medical record does not provide information related to heart murmurs, so the Levine 6-grade scale cannot be applied for assessment. Ground-truth: 根据Levine6级分级法评估心脏杂音需基于听诊结果：1.查体明确记录『心脏各瓣膜区未闻及杂音』，表明未检测到任何病理性心音；2.患者主诉及现病史均为外伤性骨折，无心血管症状（如心悸、胸痛）；3.辅助检查（X线）及生命体征（BP、心律）均未提示心脏结构或功能异常。综上，无心脏杂音存在的客观证据。分为0级。 Assessment of heart murmurs using the Levine 6-grade scale requires auscultation results: The physical examination clearly records “no murmurs heard in any cardiac valve area,” indicating no pathological heart sounds were detected. [...] Classified as grade 0.
Error in Parameter Extraction	LLM: 根据提供的病历摘要，我们可以看到患者的血清钾(K)水平为1.8mmol/L，钠(Na)水平为152mmol/L，氯(Cl)水平为143mmol/L，血糖(Glu)为9.7mmol/L，BUN为0。血浆渗透压 = 2×(152+1.8)+9.7/18+0/2.8 = 308.78 Based on the provided medical record summary, we can see the patient's serum potassium (K) level is 1.8 mmol/L, sodium (Na) level is 152 mmol/L, chloride (Cl) level is 143 mmol/L, blood glucose (Glu) is 9.7 mmol/L, and BUN is 0. Plasma osmolality = 2 × (152 + 1.8) + 9.7/18 + 0/2.8 = 308.78 Ground-truth: 已知Na=152mmol/L、K=1.8mmol/L、快速血糖=9.7mmol/L（约合9.7×18=174.6mg/dL）、BUN=10.9mmol/L（换算成mg/dL约为10.9/0.357≈30.53mg/dL）。将各数值代入：2×(152+1.8)=307.6，血糖折算项174.6/18≈9.7，BUN折算项30.53/2.8≈10.9，合计307.6+9.7+10.9≈328.2mOsm/kg。 Given Na = 152 mmol/L, K = 1.8 mmol/L, fast blood glucose = 9.7 mmol/L (approximately 9.7 × 18 = 174.6 mg/dL), BUN = 10.9 mmol/L (converted to mg/dL approximately 10.9/0.357 ≈ 30.53 mg/dL). Substituting the values: 2 × (152 + 1.8) = 307.6, glucose conversion term 174.6/18 ≈ 9.7, BUN conversion term 30.53/2.8 ≈ 10.9, totaling 307.6 + 9.7 + 10.9 ≈ 328.2 mOsm/kg.	Patient Notes: 70岁男性患者因头痛、气促就诊。现病史：未控制的高血压（血压164/113 mmHg），既往卒中病史。[...] A 70-year-old male patient presents with headache and shortness of breath. Current medical history: uncontrolled hypertension (blood pressure 164/113 mmHg), history of stroke. [...] LLM: 根据病历信息，患者高血压未控制，年龄超过65岁，因此得分为2分。 Based on the medical record information, the patient's hypertension is uncontrolled, and he is over 65 years old, thus scoring 2 points. Ground-truth: 患者年龄70岁（≥65岁加1分）。存在高血压加1分（总分2）。卒中史加1分（总分3）。无INR波动、出血史及药物使用因素。最终HAS-BLED总评分为3分。 The patient is 70 years old (≥65 years old adds 1 point). Presence of hypertension adds 1 point (total score 2). History of stroke adds 1 point (total score 3). No INR fluctuations, history of bleeding, or medication use factors. The final HAS-BLED total score is 3 points.	–
Error in Calculation /Comprehension	LLM: 需要根据患者的年龄和体位来估算动脉血氧分压(PaO2)的预测值。患者年龄为38岁。 卧位PaO2=103.5-0.42×38 = 89.34 坐位PaO2=104.2-0.27×38 = 93.78 It is necessary to estimate the predicted value of arterial oxygen partial pressure (PaO2) based on the patient's age and position. The patient is 38 years old. Supine PaO2 = 103.5 - 0.42 × 38 = 89.34 Sitting PaO2 = 104.2 - 0.27 × 38 = 93.78 Ground-truth: {'卧位PaO2': 87.5, '坐位PaO2': 93.9} {'Supine PaO2': 87.5, 'Sitting PaO2': 93.9}	Patient Notes: 61岁男性患者因复杂病史定期随访就诊。现病史：高血压控制良好（血压118/61 mmHg）[...] A 61-year-old male patient presents for regular follow-up due to a complex medical history. Current medical history: well-controlled hypertension (blood pressure 118/61 mmHg) [...] LLM: -H(高血压)：患者血压控制良好，评分为0 -H (Hypertension): The patient's blood pressure is well-controlled, scoring 0 Ground-truth: 存在高血压加1分（1分） Presence of hypertension adds 1 point (1 point)	LLM: 查体发现心率102次/分（窦性心动过速），存在奔马律（S3心音提示心功能不全），初步判断患者的Killip心功能分级为1级。 Physical examination revealed a heart rate of 102 beats per minute (sinus tachycardia) and the presence of a gallop rhythm (S3 heart sound indicating heart failure, so it can be preliminarily judged that the patient's Killip classification is Grade 1. Ground-truth: 根据查体发现奔马律（S3心音），但无肺部啰音、颈静脉怒张及外周水肿，符合Killip II级标准：存在心力衰竭体征但未达肺水肿。 Based on the physical examination findings of a gallop rhythm (S3 heart sound), but without pulmonary rales, jugular venous distention, or peripheral edema, it meets the criteria for Killip Class II: signs of heart failure are present but do not reach pulmonary edema.
Error in Unit Conversion	LLM: 1umol/L=0.0113mg/dL 969umol/L=969*0.0113mg/dL=10.90mg/dL Ground-truth: 969umol/L = 10.95	–	–

Figure 5: Case Study.

unit conversion, and final calculation. Experiments over 69 calculators and 1,143 patient cases shows that existing LLMs consistently underperform, due to compounded errors stemming from incorrect entity extraction and flawed computational logic. We hope CMedCalc-Bench can support future work

on improving Chinese medical calculation.

Limitations

While CMedCalc-Bench advances the evaluation of Chinese medical calculations, several limitations remain. First, the current four-stage protocol is restricted to text-only inputs, leaving untested multimodal reasoning over imaging, waveform, or speech data that play an increasing role in clinical decision-making. Second, automatic grading of chain-of-thought outputs relies on GPT-4o; as with other LLM-based evaluations, this introduces noise and potential bias toward the model's own reasoning style. Finally, CMedCalc-Bench is currently limited to mainland Chinese clinical language. Expanding to additional languages and regional variants will broaden its coverage and help address disparities in health-care evaluation resources.

Ethics Statement

To construct our benchmark, we exclusively curated data from publicly available sources, including published case report articles and anonymized, clinician-authored patient vignettes. No identifiable personal health information (PHI) was collected, used, or disclosed in the process. Therefore, our study fully complies with privacy and data protection standards. The benchmark is developed solely for the purpose of evaluating the medical reasoning and calculation abilities of LLMs in a controlled research setting. It is not intended for direct clinical use, medical diagnosis, or decision-making. All outputs from LLMs evaluated with this dataset should be interpreted with caution and should not replace professional medical advice.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alex J Goodell, Simon N Chu, Dara Rouholiman, and Larry F Chu. 2023. Augmentation of chatgpt with clinician-informed tools improves performance on medical calculation tasks. *medRxiv*, pages 2023–12.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tim A Green, Stevan Whitt, Jeffery L Belden, Sanda Erdelez, and Chi-Ren Shyu. 2019. Medical calculators: prevalence, and barriers to use. *Computer Methods and Programs in Biomedicine*, 179:105002.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and 1 others. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, and 1 others. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, and 1 others. 2025. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, and 1 others. 2024. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang, and Xiaofan Zhang. 2025a. MeNTi: Bridging medical calculator and LLM agent with nested tool calling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5097–5116, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025b. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.