Should I Believe in What Medical AI Says? A Chinese Benchmark for Medication Based on Knowledge and Reasoning

Yue Wu, Yangmin Huang*, Qianyun Du*, Lixian Lai,
Zhiyang He, Jiaxue Hu, Xiaodong Tao
Xunfei Healthcare Technology Co., Ltd.
{yuewu21, ymhuang9, qydu, lxlai2, zyhe, jxhu2, xdtao}@iflytek.com

Abstract

Large language models (LLMs) show potential in healthcare but often generate hallucinations, especially when handling unfamiliar information. In medication, a systematic benchmark to evaluate model capabilities is lacking, which is critical given the high-risk nature of medical information. This paper introduces a Chinese benchmark aimed at assessing models in medication tasks, focusing on knowledge and reasoning across six datasets: indication, dosage and administration, contraindicated population, mechanisms of action, drug recommendation, and drug interaction. We evaluate eight closed-source and five open-source models to identify knowledge boundaries, providing the first systematic analysis of limitations and risks in proprietary medical models.

1 Introduction

Large language models (LLMs) have made significant strides in various domains, including medication, where they provide information and recommendations related to medical treatments (Singhal et al., 2022; Nori et al., 2023). However, a significant challenge remains: these models are prone to generating hallucinations and confidently providing incorrect or incomplete information, especially in cases where they lack adequate knowledge (Stefansson and Johansson, 2021; Shukla et al., 2022). In the context of medication and drug usage, such hallucinations can lead to critical errors, particularly in high-risk situations like identifying contraindicated populations or recommending unsafe drug combinations. Despite the progress made in medical AI, a notable gap remains in the development of systematic benchmarks to evaluate the full range of a model's capabilities in medication applications.

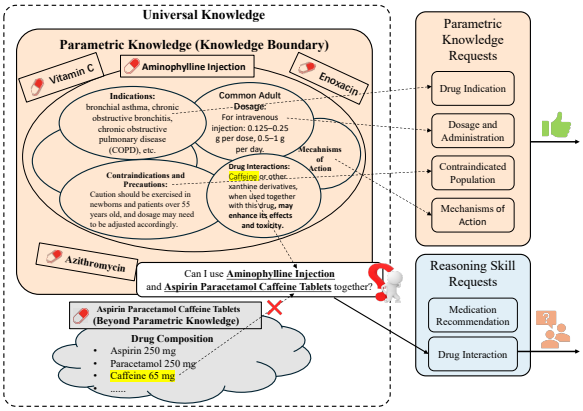


Figure 1: Our benchmark involves four datasets that directly examine model parametric knowledge and two datasets that examine model reasoning ability.

Considering that many existing public benchmarks are constructed by collecting data from online sources and may be susceptible to data leakage risks, we construct a Chinese benchmark, **ChiDrug**, from scratch, based on authoritative drug brochures. This benchmark is specifically designed to evaluate LLMs' knowledge and reasoning capabilities in the medication domain. As shown in Figure 1, our benchmark is structured into two key subdimensions: **parametric knowledge** and **reasoning capability**. We construct six diverse datasets that cover crucial aspects of drug information, including dosage and administration, indication, contraindicated population, mechanisms of action, medication recommendation, and drug interaction.

To evaluate the capabilities of existing models, we apply our benchmark to eight closed-source and five open-source models. Our work also explores various methods for expressing knowledge boundaries, providing insights into the potential risks of overconfident but inaccurate AI-generated responses.

Our contributions include: (1) This benchmark

*Corresponding authors.
Our data and code are available at: https://github.com/LeoCoder33/ChiDrug-benchmark

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1155–1164

serves as the first systematic tool for analyzing the capabilities of LLMs in the field of medicine across various dimensions. (2) To the best of our knowledge, this is the first work to systematically conduct knowledge boundary analysis on medical models in the medication domain, providing a comprehensive overview of their performance in real-world medical applications.

2 Related Work

2.1 Chinese Benchmark in Medication

Evaluating the capabilities of Large Language Models (LLMs) in the medical field necessitates specialised benchmarks, particularly when addressing Chinese medical texts. Recent efforts have led to the development of several Chinese-specific medical benchmarks, focusing on various domains such as clinical question answering, knowledge recall, and medication recommendations (Singhal et al., 2023; Liu et al., 2024; Wang et al., 2024; Yue et al., 2024).

MedExpQA (Alonso et al., 2024) proposes a multilingual benchmark evaluating models on medical question-answering tasks, including drug-related and clinical guideline questions. DialMed (He et al., 2022) focuses on dialogue-based medication recommendations, testing models on handling patient symptom queries and drug interactions. However, existing datasets do not have a dedicated benchmark built in the field of medication in Chinese to evaluate the model's ability in this area.

2.2 Abstention in LLMs

The ability of Large Language Models (LLMs) to refrain from providing answers when uncertain—is crucial for enhancing model reliability and safety. Studies have explored various methods to improve this capability (Wen et al., 2024):

Currently, methods to guide models in refusing to answer include: **Calibration-Based Methods**: After the model provides an answer, continue by asking, "Are you sure about your answer?" to verify its confidence (Tian et al., 2023). **Training-Based Methods**: Construct a training set containing both questions the model can answer and those it cannot, training the model to refuse to answer questions with unfamiliar knowledge (Slobodkin et al., 2023; Zhang et al., 2023; Stengel-Eskin et al., 2024). **Consistency-Based Methods**: Perform multiple samplings and calculate the consistency score of the model's responses to assess reliability (Kuhn et al., 2022; Feng et al., 2024a). **Token Probability Methods**: Ensemble the probability of each token generated by the model to determine the uncertainty of the response (Liang et al., 2024; Malinin and Gales, 2021).

3 Dataset

ChiDrug is designed to assess models' parametric knowledge and reasoning ability in handling critical medication-related tasks. Below, we outline the dataset construction process and the verification procedures used to ensure the quality and reliability of the data. The entire benchmark construction process is shown in Figure 2.

3.1 Dataset Construction

We began by collecting official drug brochures for existing medications from the internet[1]. We organized this information into a table that includes details on 8,000 drugs, encompassing their generic names, ingredients, specifications, indications, dosages, contraindications, drug interactions, adverse reactions, and mechanisms of action. This structured dataset served as the foundation for developing questions that evaluate the model's parametric knowledge in four areas: **Indication**, **Dosage and Administration**, **Contraindicated Population**, and **Mechanism of Action**. We extracted the relevant sections from each drug brochure and utilized Spark[2] to generate multiple-choice question stems and answer options. In constructing these questions, we ensured that the incorrect options did not overlap with the correct ones (the left part of Figure 2).

The second step involved constructing questions for **Medication Recommendation** (the middle part of Figure 2). We collected doctor-patient dialogues from the existing DIALMED dataset (He et al., 2022), where Spark transformed these dialogues into question formats, using the doctor's recommended medication as the correct option. To generate distractor options that could confuse the model, we first used Spark to extract the patient's symptoms and demographic information, then searched the drug brochures for medications that treat the same symptoms but are not suitable for the patient's demographic group, thereby creating incorrect options (e.g., "symptom in indication and de-

[1] https://drugs.dxy.cn
[2] https://xinghuo.xfyun.cn

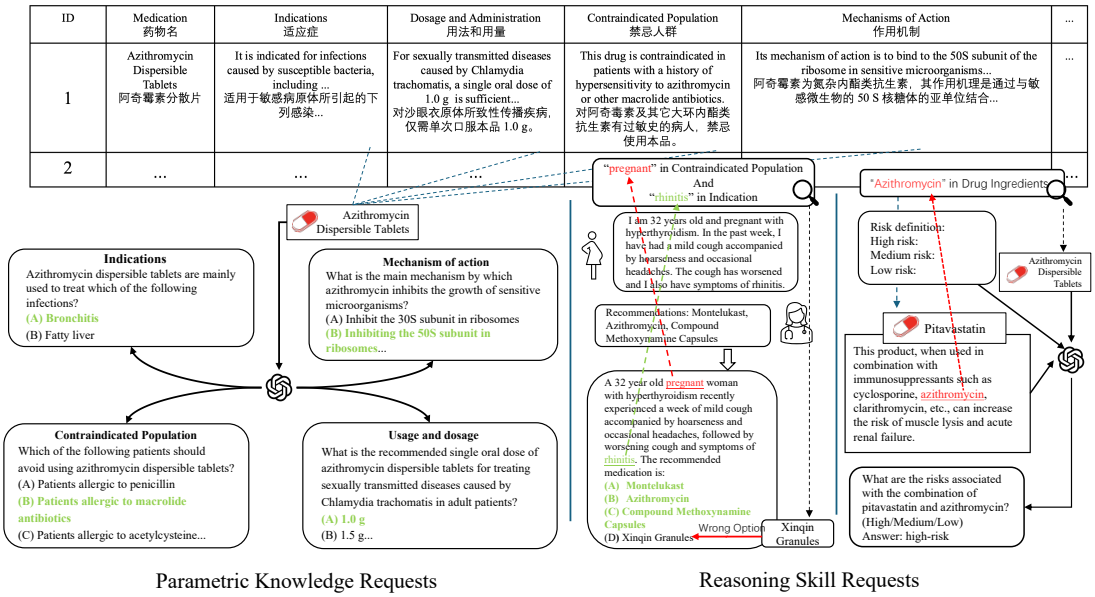


Figure 2: Overview of our benchmark construction process

mographic in contraindicated population").

In the third step, we constructed a dataset for **Drug Interaction** (the right part of Figure 2). First, doctors defined three risk levels for drug interaction (high, medium, and low). We then randomly selected a drug from the brochures and identified its combination guidelines. From there, we extracted the ingredients involved in drug interactions and further searched for medications that contained the same ingredients. Finally, we input the two drugs and the interaction documentation into Spark to generate the appropriate risk level as the correct answer.

3.2 Verification

Since we automatically generated the questions for the dataset, we implemented a double-check process to ensure the questions were reasonable. Each question was tested by three large models (GPT-4[3], Qwen-max[4], ERNIE bot[5]). We gave these models the question, options, and document sources and asked them to check the following: (1) If the question is clear, well-phrased, and free of ambiguity. (2) If the answer is correct. (3) If the answer is unique. A question was considered valid only if all three models agreed it was correct. During dataset construction, we initially generated 7,100 questions. In the LLM verification stage, Spark-generated questions had an acceptance rate of 79.32%, resulting in the removal of 1,468 questions.

Additionally, we hire doctors with licensed qualifications to examine all the datasets we construct (Appendix C).

Ultimately, we construct a benchmark dataset comprising a total of 5,243 samples, encompassing the following categories: Indication (705), Dosage and Administration (651), Contraindicated Population (659), Mechanism of Action (773), Medication Recommendation (838), and Drug Interaction (1,617). More details are provided in Appendix B.

4 Experiment

In this section, we evaluate the performance of large language models (LLMs) on our benchmark. We assess both closed-source and open-source models, using our benchmark to examine their capabilities in handling medication-related queries and their ability to identify knowledge gaps and overconfidence. Table 1 presents the results for the model ability, while the second table focuses on the methods to express the knowledge boundaries in seven different methods.

4.1 Model Performance Evaluation

We selected models with strong Chinese language capabilities, including GPT4o (Hurst et al., 2024), Claude3.5-Sonnet[6], Qwen-max[7], Doubao[8], GLM4

[3] https://chatgpt.com
[4] https://tongyi.aliyun.com/qianwen
[5] https://yiyan.baidu.com
[6] https://claude.ai
[7] https://tongyi.aliyun.com/qianwen
[8] https://www.doubao.com/chat

Close-source Models							
	Dosage and Administration	Indication	Contraindicated Population	Mechanism of Action	Medication Recommendation	Drug Interaction	Avg.
XiaoYi	**81.1**	**77.87**	66.71	92.85	**65.31**	**63.27**	**73.52**
GPT4o	66.41	73.65	69.35	92.13	59.79	59.93	**70.21**
ERNIE	67.64	65.3	57.97	92.76	51.43	38.59	62.28
Qwen-max	69.02	72.13	68.19	93.28	61.22	54.73	69.76
Doubao	71.32	71.24	54.17	92.77	63.25	55.35	68.02
GLM4	71.32	75.71	**71.02**	**94.16**	59.79	54.92	71.15
Claude3.5	54.59	74.53	70.29	89.92	54.06	60.73	67.24
Baichuan4	62.14	69.97	69.24	90.35	52.98	52.81	66.25
Open-source Models							
Bencao	28.92	19.88	12.2	40.71	16.23	**38.28**	26.04
MedGLM	38.92	13.21	8.75	44.86	20.17	34.59	26.75
MedicalGPT	33.51	10.14	3.18	49.41	13.84	30.98	23.51
ChiMedical	33.51	16.04	14.32	38.54	24.71	36.05	27.20
HuatuoGPT2	**55.83**	**47.03**	**18.66**	**77.16**	**25.18**	25.60	**41.58**

Table 1: This table presents the performance of 8 closed-source models and 5 open-source models across various medication-related tasks. Bold indicates the best performance, while underlining denotes the second-best.

	Dosage and Administration		Indication		Contraindicated Population		Mechanism of Action		Medication Recommendation		Drug Interaction		Avg.	
	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc
Baseline	55.83	55.83	47.03	47.03	18.66	18.66	77.16	77.16	25.18	25.18	25.60	25.60	41.58	41.58
Post-calibration	55.94	**57.91**	47.01	47.62	18.05	18.17	77.21	77.37	25.23	26.12	24.76	25.67	41.37	42.12
IDK	54.26	54.68	47.18	47.18	17.90	17.30	78.68	80.28	24.47	24.66	24.55	26.64	41.17	41.79
LNS	53.99	50.46	50.78	50.85	18.18	23.03	79.97	71.07	28.22	**26.21**	24.94	23.58	42.68	40.87
Probing	66.11	43.90	66.11	51.86	22.83	27.43	85.24	80.01	29.82	19.80	31.31	30.54	50.24	42.26
R-tuning	53.90	57.61	53.66	**56.00**	22.12	**31.00**	81.35	83.00	19.80	12.45	30.34	31.68	43.53	45.29
Self-Consistency	**66.85**	46.20	68.03	48.46	15.24	10.67	**89.37**	78.60	**30.15**	20.84	20.90	30.00	48.42	39.13
Semantic Entropy	64.79	55.32	**70.28**	52.71	**28.11**	26.95	86.24	**83.09**	29.22	24.95	**38.56**	**31.76**	**52.87**	**45.79**

Table 2: This table displays the performance of 7 different methods on the models' ability to detect knowledge boundaries and manage uncertainty.

(GLM et al., 2024), Baichuan4[9], XiaoYi[10], and ERNIE Bot[11], for evaluation of closed-source models. For open-source models, we chose Bencao (Wang et al., 2023), MedGLM (Haochun Wang, 2023), MedicalGPT (Xu, 2023), ChiMedical (Tian et al., 2024), and HuatuoGPT2 (Chen et al., 2024) for evaluation.

The results summarised in Table 1 show that the closed-source models generally outperformed the open-source models across all dimensions, with XiaoYi leading in overall performance, followed closely by GPT4o and ERNIE Bot. In Open-source models, Bencao and MedicalGPT demonstrated lower performance, particularly in complex tasks like Contraindicated Populations and Drug Interactions, while HuatuoGPT2 generally outperformed other models. We will provide a more detailed analysis of each model in the Appendix A.

4.2 Methods for Knowledge Boundary Detection

4.2.1 Task Definition

A formal definition of knowledge boundary detection can be briefly stated as follows:

Input: A model M and a query q.

Output: A response r where

$$r = \begin{cases} M(q), & \text{within parametric knowledge} \\ U, & \text{beyond parametric knowledge} \end{cases} \quad (1)$$

where $M(q)$ is the model's generated answer, and U is an explicit uncertainty expression or abstention (e.g., "I don't know", a confidence score, or an alternative uncertainty marker).

4.2.2 Methods

In this subsection, we apply seven methods to explore their impact on expressing uncertainty or abstention, using HuatuoGPT2 as the backbone.

Post-Calibration (Tian et al., 2023): Enhances model confidence by prompting it to verbalize its certainty after providing an answer.

IDK (I Don't Know): Similar to non-of-the-above (NOTA) in (Feng et al., 2024b), we incorporate an additional "I don't know" option and instruct the model to abstain from answering.

LNS (Malinin and Gales, 2021): Utilizes probabilistic ensemble-based techniques to assess uncertainty in structured prediction tasks, aiding in more reliable outputs.

Probing (Slobodkin et al., 2023): Analyzes internal model representations to understand how they encode information about answerability, helping detect overconfidence and hallucinations.

[9] https://www.baichuan-ai.com
[10] https://chatdr.iflyhealth.com
[11] https://yiyan.baidu.com

R-tuning (Zhang et al., 2023): Instructs models to explicitly state when they lack knowledge, reducing the generation of hallucinated information.

Self-Consistency (Kuhn et al., 2023): Enhances reasoning by generating multiple reasoning paths and selecting the most consistent answer, improving response reliability.

Semantic Entropy (Feng et al., 2024a): Estimates uncertainty in natural language generation by considering linguistic invariances, allowing models to assess the reliability of their outputs better.

4.2.3 Evaluation Metrics

Given the definition as follows:

C = the number of correct answers.

A = the number of total answered questions (excluding abstentions).

A_{correct} = the number of correct abstentions, i.e., questions the model correctly refused to answer because the answer was unknown or uncertain.

N = the total number of questions.

In this experiment, two evaluation metrics are used:

Precision: Measure the proportion of correct answers out of the total predictions made, without abstaining.

$$\text{Precision} = \frac{C}{A} \tag{2}$$

Abstain Accuracy (Feng et al., 2024a): Evaluates the proportion of correct answers and correct abstention due to uncertainty.

$$\text{Abstain-Acc} = \frac{C + A_{\text{correct}}}{N} \tag{3}$$

Results are shown in Table 2. Post-calibration and IDK cannot achieve good results in HuatuoGPT2 due to its weak instruction capabilities. Self-Consistency improved accuracy in complex tasks like Medication Recommendations. Probing refined uncertainty estimations with varying effectiveness. R-tuning reduced hallucinations but sometimes sacrificed performance on complex tasks, while LNS showed mixed results, improving Medication Recommendations but hindering performance on Drug Interaction. Overall, Semantic Entropy has achieved good results in both metrics, and we further analyze the effectiveness of this method on multiple models in Appendix D.

5 Conclusion

We present ChiDrug, a benchmark designed to evaluate LLMs (Large Language Models) in medication-related tasks, with an emphasis on their knowledge and reasoning abilities. Both GLM4 and XiaoYi performed exceptionally well; however, even these advanced models exhibited gaps in drug knowledge. Our work highlights the need for effective methods to align the knowledge boundaries of LLMs, particularly for high-risk tasks.

6 Limitations

This study primarily focuses on Chinese medical texts, which may limit its generalizability. The benchmark doesn't fully capture the complexities of real-world medical decision-making. Additionally, model generalization to new knowledge, handling uncertainty, and reliance on high-quality, up-to-date data are ongoing challenges for AI in healthcare.

7 Ethical considerations

The medication dictionary we constructed is entirely sourced from DingXiangYuan[12], a public medical website. The data on DingXiangYuan was also collected from the China National Medical Products Administration, and the website has a statement allowing non-commercial citations. Documents within our dictionary do not contain private information, so there is no risk of privacy leakage. All the drug information we collect has obtained the national drug approval certificate and is free from copyright issues, in accordance with the regulations of the Chinese government.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024a. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b.

[12] https://drugs.dxy.cn

Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *Preprint*, arXiv:2402.00367.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Sendong Zhao Bing Qin Ting Liu Haochun Wang, Chi Liu. 2023. Chatglm-med: chatglm. https://github.com/SCIR-HI/Med-ChatGLM.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. Dialmed: A dataset for dialogue-based medication recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 721–733.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, et al. 2024. Internal consistency and self-feedback in large language models: A survey. *arXiv preprint arXiv:2407.14507*.

Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Preprint*, arXiv:2407.10990.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

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

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625.

Elis Stefansson and Karl H. Johansson. 2021. Computing complexity-aware plans using kolmogorov complexity. *Preprint*, arXiv:2109.10303.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. Chimed-gpt: A chinese medical large language model with full training regime

and better alignment to human preferences. *Preprint*, arXiv:2311.06025.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2024. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Wenjing Yue, Xiaoling Wang, Wei Zhu, Ming Guan, Huanran Zheng, Pengfei Wang, Changzhi Sun, and Xin Ma. 2024. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine. *Preprint*, arXiv:2406.01126.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

A Model Performance Analysis

A.1 Visualization of Model Performance

In this section, we present radar chart visualizations to highlight the performance of both closed-source and open-source models across different medication-related tasks. As shown in Figure 3 and Figure 4, the radar charts provide a clear, comparative view of how various models handle tasks such as Indication, Dosage and Administration, Contraindicated Population, and Mechanisms of Action. Notably, models such as GLM4 and XiaoYi stand out for their excellent performance, with XiaoYi leading the closed-source models and GLM4 demonstrating remarkable consistency. On the other hand, HuatuoGPT2 significantly outperforms the other open-source models. These findings underscore the importance of model selection in high-stakes domains like healthcare, where the quality of responses directly impacts patient safety.

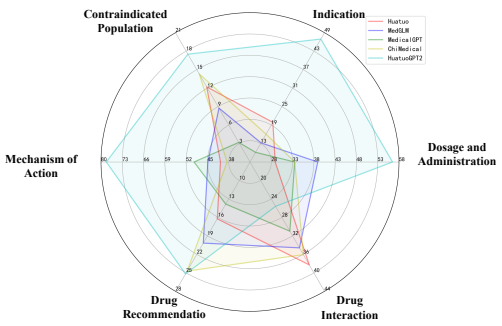


Figure 3: Radar Chart Representation of Open-Source Models Performance.

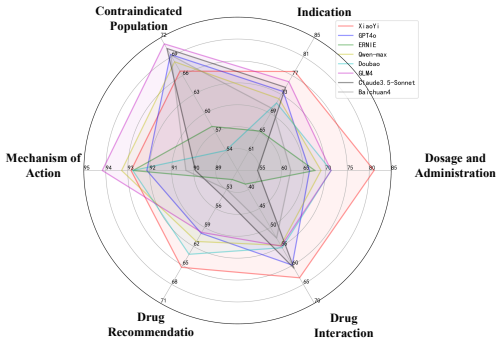


Figure 4: Radar Chart Representation of Close-Source Models Performance.

A.2 Knowledge Mastery Assessment of Common Drugs

To further evaluate model capabilities, we focus on a subset of 282 commonly used drugs. For each drug, we constructed questions about Indication, Dosage and Administration, Contraindicated Population, and Mechanism of Action, drawing from the benchmark dataset. The knowledge boundary of the models was then assessed by visualizing their performance on these tasks, as shown in the radar charts for GLM4, XiaoYi, and GPT4o, and the results are presented in Table 3.

Figure 5 illustrates the knowledge boundary performance of three models across 282 drugs, where each drug has 4 different sub-task questions. The radial score axis (0-4) represents the total number of correct answers per drug: Since each drug has 4 questions, a score of 4.0 means the model answered all 4 questions correctly.

Orange regions indicate that the model had only one chance to answer all 4 questions.

Yellow regions indicate that the model had up to 5 attempts per drug to answer correctly. If it answered any question correctly in any of these attempts, it was awarded 1 point.

The yellow regions that are not covered by orange represent cases where the model initially gave incorrect answers but later recovered and answered correctly within the 5 attempts. These visualizations emphasize that while some models show robustness in their knowledge, significant gaps remain in certain drug-related tasks.

The results indicate that GLM4 and XiaoYi exhibit stronger consistency in answering these questions correctly across the four tasks compared to GPT4o. However, there were cases where even the most advanced models struggled to demonstrate comprehensive knowledge across all aspects of these drugs. This highlights a key issue—despite their advanced capabilities, large models still fall short in areas of medication-related knowledge.

	Dosage and Administration	Indication	Contraindicated Population	Mechanisms of Action	Avg.
XiaoYi	**82.27**	**78.01**	63.12	92.20	**78.90**
GPT4o	67.02	74.11	70.2	92.91	76.24
ERNIE	69.15	64.89	58.51	91.13	70.92
Qwen-max	67.73	74.11	71.63	92.55	76.51
Doubao	75.89	74.47	64.54	92.55	76.86
GLM4	74.11	75.89	**71.99**	**93.97**	78.99
Claude3.5	67.09	67.73	54.26	93.46	70.64
Baichuan4	59.93	70.21	51.06	91.49	68.17

Table 3: Performance of various models on Common Drugs

A.3 Performance on Reasoning Models

While the GPT family is known for its strong reasoning capabilities, our results reveal nuanced performance differences, as shown in Table 4. In particular, OpenAI's o1 performs worse than GPT-4o, suggesting that strong reasoning ability alone does not guarantee superior performance, especially in knowledge-intensive domains like medicine. We argue that a model's performance ceiling in such domains is also closely tied to its parametric knowledge.

For instance, DeepSeek-R1 excels in tasks such as Drug Indication and Contraindicated Population, and this strength naturally extends to better performance in Medication Recommendation, which relies heavily on knowledge of drug usage constraints. In contrast, o1 performs worse in Indication, which correlates with its weaker performance in the Medication Recommendation task.

Furthermore, although GPT-4o is not explicitly trained on complex chain-of-thought (CoT) reasoning datasets, it demonstrates competitive reasoning ability. However, models like o1 and o3-mini, which are optimised for reasoning in code and mathematics, do not show a clear advantage in our medical reasoning benchmark, highlighting the limits imposed by insufficient medical knowledge.

To further validate that reasoning ability becomes the primary bottleneck when knowledge is sufficient, we constructed knowledge-complete prompts for Medication Recommendation and Drug Interaction tasks. These prompts explicitly included all necessary domain knowledge required to answer the questions.

As the results illustrated in Table 5 confirm, our hypothesis is supported: when provided with sufficient knowledge, models with strong reasoning capabilities, such as o1 and o3-mini, outperform GPT-4o, which lacks comparable reasoning-specific training. This highlights the importance of not only enriching models with knowledge but also enhancing their reasoning mechanisms—especially in professional domains like medicine.

In summary, our benchmark suggests that advancing domain-specific reasoning is a critical frontier for LLM development, and we hope our work offers meaningful insight for future research in this direction.

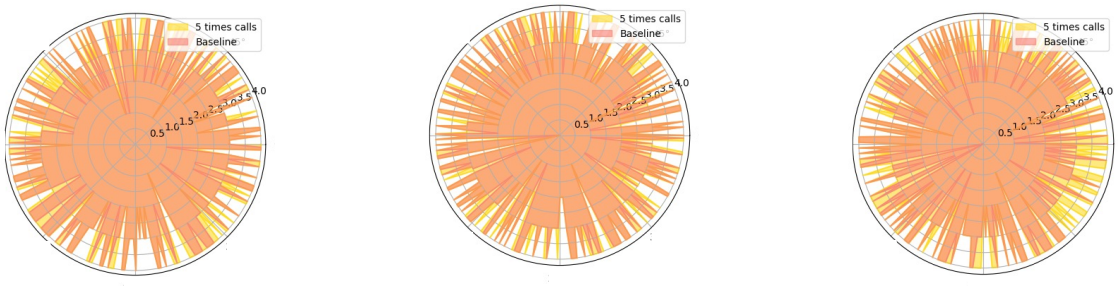
B Dataset Statistics

This section presents key statistics of our benchmark dataset across six task categories. As shown in Table 6, each task varies in terms of average input length, the number of unique drugs, and the number of associated diseases. Notably, the Drug Interaction task contains the longest samples and the largest set of drugs, reflecting its complexity. These statistics highlight the diversity and richness of our dataset, which is crucial for evaluating both the parametric knowledge and reasoning capabilities of LLMs in the medication domain.

C Expert Review of Datasets

During the manual verification phase, we hire 10 doctors with licensed qualifications to examine all the datasets we construct. We divided the 10 doctors into two groups, each consisting of five doctors. One group was responsible for reviewing the questions related to Drug Indication, Dosage and Administration, Contraindicated Population, and Mechanisms of Action, while the other group reviewed the questions for Medication Recommendation and Drug Interaction. After the initial review, the groups conducted a cross-check to ensure accuracy.

For data points that the doctors identified as problematic, we directly archived them and excluded



XiaoYi GLM4 GPT4o

Figure 5: Knowledge boundary chart for GLM4, XiaoYi, and GPT4o across 282 common drugs. The orange area indicates that the model answered correctly once, while the yellow area indicates 5 times opportunities to answer correctly. The orange areas in the yellow-covered region represent cases where a model made an error in a single attempt but was able to recover after multiple tries.

Task	Dosage and Administration	Indication	Contraindicated Population	Mechanism of Action	Medication Recommendation	Drug Interaction	Average
DeepSeek-R1	70.81	77.06	75.69	91.98	61.59	62.75	73.31
OpenAI o1	56.69	74.80	58.31	92.93	61.53	45.10	64.89
OpenAI o3-mini	68.52	70.69	72.57	94.15	51.77	64.79	70.42
GPT-4o	66.41	73.65	69.35	92.13	59.79	59.93	70.21

Table 4: Zero-shot accuracy of reasoning models across medical knowledge tasks.

Model	**Medication Recommendation**	**Drug Interaction**
OpenAI o1	84.73	92.02
OpenAI o3-mini	80.79	91.34
GPT-4o	79.83	88.62

Table 5: Accuracy on reasoning tasks with knowledge-complete prompts.

Task	Average Length	Drugs Involved	Diseases Involved
Drug Indication	86.21	605	1061
Dosage and Administration	108.36	581	188
Contraindicated Population	94.45	607	636
Mechanisms of Action	108.25	791	224
Medication Recommendation	114.61	940	830
Drug Interaction	479.97	1293	None

Table 6: Task-wise statistics of the ChiDrug benchmark.

them from the final dataset. As a result, the accept rate of quesiton is 93.09%. The cost of hiring an doctor to label a single sample was 2 RMB (approximately 0.26 USD).

D Semantic Entropy (SE) Method for Knowledge Boundary Expression

In this section, we explore the Semantic Entropy (SE) method used to detect knowledge boundaries, as introduced in Section 4.2. The SE method is particularly noteworthy for its effectiveness in expressing model uncertainty and improving response reliability, as demonstrated in our experiments. We applied this method to HuatuoGPT2 and XiaoYi, observing that it significantly enhanced the models' performance on challenging tasks, such as Medication Recommendations and Drug Interactions.

As shown in Table 7, the SE method proved to be robust and consistent across different model architectures and sizes. It improved both Precision and Abstain Accuracy, regardless of the model's scale. This reinforces the notion that SE is an effective tool for managing uncertainty, making it an essential method for enhancing the reliability of models in real-world medical applications.

E Case Study

In Figure 6, we present a case to illustrate the practical section of the ChiDrug.

Model	Method	Dosage and Administration		Indication		Contraindicated Population		Mechanisms of Action		Medication Recommendation		Drug Interaction	
		Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc	Precision	A-Acc
HuatuoGPT2	w/o SE	55.83	55.83	47.03	47.03	18.66	18.66	77.16	77.16	25.18	25.18	25.60	25.60
	SE	64.79	55.32	70.28	52.71	28.11	26.95	86.24	83.09	29.22	24.95	38.56	31.72
XiaoYi	w/o SE	81.1	81.1	77.87	77.87	66.71	66.71	92.85	92.85	59.31	59.31	63.27	3.27
	SE	83.42	84.46	94.01	94.15	80.14	83.28	91.71	92.1	67.71	75.71	63.85	69.02

Table 7: Application of the SE method on HuatuoGPT2 and XiaoYi models, showcasing the performance improvements achieved through the SE method. This method enhances precision and uncertainty handling, effectively reducing hallucinations.

Tasks	Question	Answer
Dosage and Administration	复方锌布颗粒剂的推荐服用方式为多少包，多少次一天？ (A)儿童5岁以下一次2包，一天3次 (B)成人一次2包，一天3次 (C)成人一次1包，一天3次 (D)6～14岁儿童一次1包，一天2次 (E)6～14岁儿童一次1包，一天3次	B 用法用量： 口服。3～5岁儿童，一次半包；6～14岁儿童，一次1包；成人，一次2包，一日3次。
Indication	复方锌布颗粒剂主要用于缓解以下哪些症状？ (A)普通感冒引起的发热 (B)急性肠胃炎 (C)普通感冒引起的四肢酸痛 (D)普通感冒引起的打喷嚏	ACD 适应症： 用于缓解普通感冒或流行性感冒引起的发热、头痛、四肢酸痛、鼻塞、流涕、打喷嚏等症状。
Contraindicated Population	复方锌布颗粒剂不适用于以下哪些人群？ (A)心脏病患者 (B)哺乳期妇女 (C)对阿司匹林过敏的哮喘患者 (D)高血压患者	BC 禁忌： 1.对其他非甾体抗炎药过敏者禁用。2.孕妇及哺乳期妇女禁用。3.对阿司匹林过敏的哮喘患者禁用。
Mechanism of Action	关于复方锌布颗粒剂各组分的主要药理作用是： (A)布洛芬具有抗炎作用，葡萄糖酸锌促进蛋白质合成，马来酸氯苯那敏为解热镇痛药 (B)布洛芬具有解热镇痛作用，葡萄糖酸锌能增强免疫功能，马来酸氯苯那敏为抗组胺药 (C)布洛芬为抗组胺药，葡萄糖酸锌具有解热功能，马来酸氯苯那敏具有镇痛作用 (D)布洛芬为解热镇痛药，葡萄糖酸锌参与多种酶的合成与激活，马来酸氯苯那敏为抗组胺药	D 作用机制： 布洛芬能抑制前列腺素合成，具有解热镇痛作用；葡萄糖酸锌中锌离子能参与多种酶的合成与激活，有增强吞噬细胞的吞噬能力的作用；马来酸氯苯那敏为抗组胺药，能减轻由感冒或流感引起的鼻塞、流涕、打喷嚏等症状。
Drug Recommendation	回答以下不定项选择题（可能包含1个或多个正确选项）： 一位孕晚期患者因为感冒出现咳嗽、喉咙里有异物感以及扁桃体发炎，可以考虑推荐的药物是： (A)双黄连口服液 (B)热毒宁注射液 (C)贝美前列素滴眼液 (D)银芩胶囊	A 双黄连口服液的适应症： 疏风解表，清热解毒。用于外感风热所致的感冒，症见发热、咳嗽、咽痛。
Drug Interaction	注射用降纤酶与抗纤溶药联用的风险等级是？	高风险 注射用降纤酶 使用本品应避免与水杨酸类药物（如：阿司匹林）合用。抗凝血药可加强本品作用，引起意外出血；抗纤溶药可抵消本品作用，禁止联用。

Figure 6: Partial cases of ChiDrug on 6 sub datasets.