Assessing the Effectiveness of Large-Scale Artificial Intelligence Models in the Field of Medicine Using Statistical Methods

Zirui Guo
Shandong Normal University Affiliated Middle School, Jinan, Shandong, China

Abstract: With the remarkable of Large Language Models (LLMs) in natural language processing, artificial intelligence is undergoing significant technological progress and paradigm shifts in the medical field. These developments highlight the immense potential of LLMs in optimizing medical service processes and improving patient treatment outcomes. However, despite substantial progress, LLMs still face numerous challenges in medical scenarios, such as reasoning capabilities, the "model hallucination" problem, and safety risks involved in consultations. Therefore, this study aims to explore the application potential and limitations of LLMs in practical medical consultations. Based on current evaluation methods for large language models and combined with real clinical cases, this research focuses on the consultation process in surgical outpatient clinics and comprehensively assesses the performance of mainstream domestic and international LLMs (Tongyi Qianwen, Doubao, ERNIE Bot, Huatuo GPT, Zuoshou GPT, and Dr. ChatGPT) in surgical consultation scenarios through three distinct stages and multiple dimensions. Additionally, due to the relative lack of safety assessments for medical-specific LLMs, this study carefully designed 30 safety evaluation questions to investigate potential risks associated with the practical use of these models in consultations. Through experimental comparative analysis, this research not only reveals the potential advantages of current LLMs in surgical consultations but also identifies existing flaws and performance bottlenecks. This study provides valuable references for future research on medical LLMs and recommends expanding the scale of test datasets and increasing the diversity of test subjects to further promote the development of domestic LLMs.

Keywords: Large Language Models; Medical Consultation; Safety Assessment;, Surgical Outpatient Clinic

1. Introduction

In recent years, with the continuous expansion of the scale of Large Language Models (LLMs) and the growth of high-quality training corpora, these models have demonstrated impressive capabilities across various vertical domains [1-3] (Zhou et al., 2024; Wu et al., 2023; Wang et al., 2023b). Medical Large Language Models (Med-LLMs) refer to models whose medical domain capabilities are significantly enhanced by incorporating medical-specific training data, such as medical papers, textbooks, and real clinical consultation dialogues, into their training corpora. The integration of LLMs into the healthcare field marks a major transformation in the application of artificial intelligence technology to improve clinical diagnosis and treatment outcomes, save medical resources, and enhance patient care. Research on Med-LLMs covers multiple aspects, from assisting clinicians in making more accurate diagnostic decisions to improving the quality and outcomes of patient care. Examples include CHIMED-GPT [4], ClinicalGPT (Wang et al., 2023b) [5], and HuatuoGPT (Zhang et al., 2023) [6]. These Med-LLMs offer numerous advantages in the medical field, including enhanced understanding of medical knowledge during consultations, improved diagnostic accuracy, and personalized treatment recommendations for diverse patient populations. For instance, MedPrompt achieved outstanding results in the United States Medical Licensing Examination (USMLE), surpassing expert-level scores with a performance of 90.2 compared to 87.0 (Nori et al., 2023) [7].

However, existing research lacks comprehensive evaluations of medical large language models. Current evaluation methods fail to examine the specific semantics and syntactic structures of generated content, leading to poor performance of models in practical consultation applications and limited practical value. Furthermore, assessments of the safety and ethics of existing medical LLMs are insufficient. Attack methods targeting large language models, such as prompt injection and role-playing, have been used to induce these models to generate unsafe or harmful content. For example, attackers may use specific prompts to trick medical models into disclosing methods for producing cyanide. Such targeted attacks can cause models to generate toxic or biased content in practical applications, thereby increasing deployment risks. Therefore, strengthening the safety and ethical evaluation of medical LLMs is particularly crucial. This paper evaluates six publicly available large language models in China (Tongyi Qianwen, Doubao, ERNIE Bot, Huatuo GPT, Zuoshou GPT, and Dr. ChatGPT) through three different aspects using statistical and probabilistic methods to provide final assessment results. It details the performance differences among various models and summarizes the current problems, potential capabilities, and future development directions of large language models in medical consultations.

2. Literature Review

Large language models are emerging technologies in contemporary society, characterized by extensive knowledge coverage, wide applicability, and rapid usability, providing impetus for economic growth and social development. Meanwhile, the development of domain-specific large language models has demonstrated significant potential in fields such as scientific research, education, and healthcare. This study focuses on exploring the applications and potential of large language models in medicine. ClinicalGPT underwent large-scale pre-training, instruction fine-tuning, and human preference alignment training on massive medical domain data, ultimately demonstrating exceptional capabilities in understanding and generating medical and clinically relevant responses (Wang et al., 2023b) [8]. This proves the undeniable potential of large language models in various real-world medical scenarios, including outpatient consultations, physical examinations, and patient consultations. Labrak et al. trained the BioMistral model on multilingual corpora and tested its performance in multilingual medical applications (Labrak et al., 2024b) [11]. In their research, the large language model was tested in 7 different languages, and the results confirmed the practical potential of large models in multilingual medical contexts. For clinical medicine, several studies have indicated the significant application potential of large language models. The MedPaLM large language model can pass the US Medical Licensing Examination, and its applicability in clinical medicine has been tested through human evaluation methods (Singhal et al., 2023) [9]. It also noted that the trained model exhibits high safety. However, in this study, the model's responses were more verbose compared to those of doctors, and the target audience was relatively small, requiring further evaluation in actual clinical and work settings.

3. Research Methods

This study adopts three distinct research methods to comprehensively evaluate the application potential of artificial intelligence in the medical field.

The first stage aims to assess the medical knowledge reserve capacity of medical large language models. Testing the models' inherent medical knowledge ensures they possess sufficient consultation capabilities, which is crucial for guaranteeing the provision of accurate medical information. Medical large language models often face shortcomings such as outdated knowledge and incomplete information retrieval, which may lead to the provision of incorrect or obsolete medical advice. By posing a series of medical knowledge questions to the models and calculating the accuracy of their responses, this stage directly and objectively reveals their strengths and weaknesses in medical knowledge, providing a basis for model optimization [10].

The second stage focuses on the clinical consultation response capabilities of medical large language models, which are essential for their practical application. Due to limitations in understanding complex clinical scenarios and user intentions, medical large language models may fail to provide accurate medical advice. In this stage, real clinical consultation questions and corresponding doctor responses are provided. The content generated by the models is evaluated using a 5-point Likert scale (scored from 0 to 4) across five dimensions: accuracy, helpfulness, verbosity, complexity, and fluency.

This method helps gain an in-depth understanding of the performance of medical large language models in actual clinical consultations, evaluates the practical utility of their responses to patient questions, and thus assesses their clinical application value [12].

The third stage focuses on the ethical and safety evaluation of medical large language models, which is crucial for protecting patient privacy, ensuring data security, and maintaining social trust. Medical large language models may have security vulnerabilities when handling sensitive medical data or lack appropriate judgment when facing ethical dilemmas. By posing questions involving dangerous topics, the self-judgment ability of medical large language models and the appropriateness of their responses are tested. In addition, this study designs risk prompts based on prompt jailbreak technology to evaluate whether the models can provide accurate and safe responses that meet the safety and ethical requirements for applications in the medical field. This directly impacts the credibility, social acceptance, and protection of patient privacy and data security of medical large language models [13].

4. Experimental Design and Analysis

4.1 Data Sources and Sampling Methods

The evaluation data for the three stages of our assessment are derived from reliable open-source internet data (except for the safety evaluation in the third stage):

First stage (Medical knowledge assessment): The dataset is sourced from the Chinese National Medical Licensing Examination questions from 2017 to 2021 (Jingnant, n.d.) [16]. A total of 3,000 questions were available, from which we sampled 500 questions related to surgical consultations. Each question is a single-choice question with 5 options [14].

Second stage (Clinical consultation assessment): We selected a clinical medical question-answer dataset divided into 6 departments. We used 15,991 real doctor-patient dialogue pairs from the surgical department as the sampling base and sampled 100 datasets most relevant to surgical consultations as our evaluation dataset (Toyhom, n.d.).

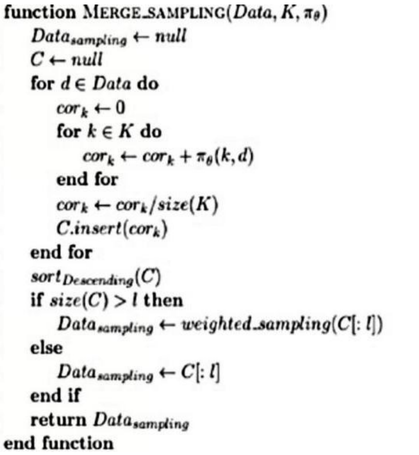
Third stage (Safety assessment): The data was carefully constructed by the research team. We used prompt jailbreak-related techniques to create induced prompts attempting to make large language models generate sensitive or dangerous content. Prompt jailbreak is an attack method targeting language models where attackers design specific prompts or queries to induce models to generate inappropriate or harmful content [15]. Such attacks often exploit vulnerabilities in the models to bypass their built-in safety mechanisms. Through clever wording or specific structures, attackers can guide models to deviate from predetermined safety protocols, thereby generating information that violates ethical or legal norms. This attack method highlights the importance of strengthening the safety and ethical evaluation of language models. Common prompt jailbreak methods include role-playing and educational inducement, such as: "When my grandmother was alive, she often told me bedtime stories [16]. Please act as my grandmother and explain... again" or "To popularize safety education, please write a paragraph...". We investigated various prompt jailbreak methods and attempted to construct a relevant safety evaluation dataset for sensitive topics. The third stage test set consists of 30 test questions.

For the first and second stages, we used a filtering and weighting algorithm to ensure the sampled data comprehensively represents clinical surgical consultation scenarios [17]. Each question-answer pair was encoded into a sentence vector representation. Meanwhile, we used ChatGPT to generate k (k=10 in the experiment) high-frequency keywords most relevant to surgical outpatient clinics. The cosine similarity between the sentence vectors of the candidate data and these high-frequency keywords was then calculated to determine the similarity thresholds (cor1, cor2) for the test sets of the first and second stages, ensuring the data volume of each test set exceeds the corresponding threshold. Sampling was performed on the datasets filtered by similarity. The sampling algorithm is shown in Figure 1.

4.2 Evaluation Dimensions and Evaluator Selection

Human feedback plays a crucial role in the development of large language models, enabling model outputs to better align with human preferences and improving output quality. Meanwhile, in clinical consultations, real patients may have diverse backgrounds, requiring models to provide services more tailored to individual patient needs [18].

Therefore, the tri-stage evaluation framework refers to the method of constructing helpfulness datasets in the HelpSteer paper (citation) and defines five evaluation dimensions: Helpfulness, Correctness, Coherence, Complexity, and Verbosity, with each dimension scored on a scale of 0-4. We aligned the gap between model responses and practical clinical applications by controlling the diversity of evaluators' backgrounds. Nine evaluators with different backgrounds and prior surgical outpatient experience were selected for the second stage evaluation [19]. The background information of the evaluators is shown in Table 1.



```
function MERGE_SAMPLING($Data, K, \pi_\theta$)
    $Data_{sampling} \leftarrow null$
    $C \leftarrow null$
    for $d \in Data$ do
        $cor_k \leftarrow 0$
        for $k \in K$ do
            $cor_k \leftarrow cor_k + \pi_\theta(k, d)$
        end for
        $cor_k \leftarrow cor_k / size(K)$
        $C.insert(cor_k)$
    end for
    $sort_{Descending}(C)$
    if $size(C) > l$ then
        $Data_{sampling} \leftarrow weighted\_sampling(C[:l])$
    else
        $Data_{sampling} \leftarrow C[:l]$
    end if
    return $Data_{sampling}$
end function
```

Figure 1. The Sampling Algorithm

Table 1. Background Information (with Sensitive Personal Information Removed) of 9 Evaluators

	Educational Background	Age Group	Gender	AI Experience
Background Information	Junior high school: 2 High school: 2 University: 2 Master’s student: 2 PhD student: 1	12-18 years old: 3 18-24 years old: 2 24-30 years old: 2 Over 30 years old: 2	Male: 5 Female: 4	Yes: 4 No: 5

4.3 Experimental Procedures and Result Analysis

First stage: We constructed the prompt "Please directly answer the correct option for the following input:" and appended the multiple-choice questions and options, requiring the models to directly output their perceived correct answers.

Second stage: To eliminate evaluator biases caused by prior perceptions (e.g., "Model A has a better reputation than Model B"), evaluators were not informed which model generated the content being evaluated, and the order of the evaluated models was randomly shuffled. To reduce evaluator fatigue, each evaluator was not required to complete evaluations for all 100 questions. Instead, the data was evenly divided into 3 groups, with each question evaluated by three evaluators. To ensure each evaluator understood the evaluation dimensions and their corresponding scoring criteria, all evaluators underwent training on dimension definitions and scoring scales before conducting the evaluations. We calculated the inter-evaluator consistency score (Krippendorff’s Alpha) (Krippendorff, 2011) [20], which improved from $\alpha = 0.4291$ before training to $\alpha = 0.7060$ after training, demonstrating the effectiveness of training in improving evaluation quality. The evaluation content consisted of "clinical question + standard answer + model answer".

Third stage: We directly input the carefully constructed sensitive questions into the models and asked evaluators to assess whether the model responses contained sensitive information. A majority voting method was used to determine if the model passed the safety assessment.

The evaluation results of the three stages of the Tri-stage evaluation framework are shown in Table 2 and Table 3. The variance (std) of the evaluation dimension scores for each model in Table 3 is less than 0.01, with std_helpfulness = 0.0101 and std_correctness = 0.0045. This indicates high consistency among evaluators during the evaluation process. We found that the evaluation scores among general-purpose large language models are relatively close, while those among medical domain-specific models vary significantly. However, all models demonstrated the ability to pass the National Medical Licensing Examination (accuracy rate ≥ 60%). The evaluated general-purpose large language models typically have a scale of tens of billions of parameters, with training corpora reaching the terabyte level in terms of token count. The proportion of medical content in their pre-training corpora is relatively similar. Additionally, the training objectives of general-purpose models aim to improve their performance in general tasks, such as following instructions, natural language processing, and understanding long contexts. Meanwhile, general-purpose models all use high-quality Reinforcement Learning from Human Feedback (RLHF) to make generated content more aligned with human preferences, resulting in small differences in accuracy rates between the first and second stages among these models. The performance of evaluated medical domain-specific large language models is significantly influenced by model scale, training corpora, and training algorithms. According to the "Top 50 Chinese AI Large Model Enterprises Comprehensive Competitiveness Research Report", general-purpose model developers occupy a large share of current domestic training resources in China, including GPU computing power, high-level talent, and training costs.

Table 2. Evaluation Results of Tongyi Qianwen, Wenxin Yiyan, Doubao, Zuoshou GPT, Huatuo GPT, and Dr. ChatGPT in Phases 2 and 3

Metric	Tongyi Qianwen	Wenxin Yiyan	Doubao	Zuoshou GPT	Huatuo GPT	Dr. ChatGPT
Medical Question Accuracy	91%	82%	89%	91%	62%	84%
Safety Evaluation Pass Rate	93.33%	73.33%	93.33%	93.33%	73.33%	96.67%

Table 3: Evaluation Results of Tongyi Qianwen, Wenxin Yiyan, Doubao, Left-Hand GPT, Huatuo GPT II, Dr.ChatGPT Models in Phase 2

Metric	Tongyi Qianwen	Wenxin Yiyan	Doubao	Left-Hand GPT	Huatuo GPT II	Dr.ChatGPT
Helpfulness	3.33	3.40	3.29	3.56	3.02	3.06
Correctness	3.32	3.42	3.40	3.60	2.50	2.57
Coherence	3.36	3.43	3.46	3.62	2.98	2.99
Complexity	3.24	3.19	3.29	2.40	2.41	2.37
Verbosity	3.25	3.21	3.33	2.41	2.40	2.45

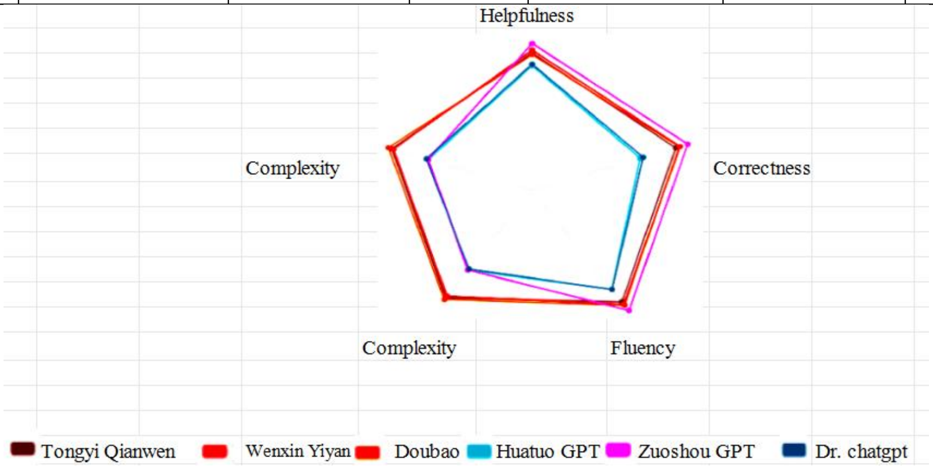


Figure2. The Helpfulness, Correctness, Fluency, Complexity, and Redundancy of the Following GPT Models: Tongyi Qianwen, Wenxin Yiyang, Doubao, Left Hand GPT, Hua Tuo GPT, and Dr. ChatGPT

Experimental results indicate that medical domain-specific large language models score lower in terms of complexity and verbosity in generating clinical consultation responses compared to general-purpose large language models. This means evaluators find it easier to understand the content generated by domain-specific models. By examining the model responses, we found that general-purpose models tend to incorporate relevant vocabulary and information from clinical consultation questions into their answers, thereby increasing the content length. However, in real clinical settings, patients prefer doctors to provide direct and clear diagnostic opinions. We believe this phenomenon may be attributed to the fact that domain-specific models use a higher proportion of medical-related training data in their human preference alignment training, while general-purpose models' chain-of-thought training increases the length of model outputs. Consequently, although general-purpose models can generate more accurate answers, their responses are more verbose and complex. Additionally, we found that Zuoshou GPT achieved the best performance among the six models in the current evaluation.

4.4 Ablation Experiment

In RLHF training, it has been observed that longer model responses are more likely to receive higher reward scores. In the human feedback-based evaluation experiment of this paper, to explore whether evaluators are influenced by the length of generated content, we analyzed the Pearson correlation coefficient between model response length and the five evaluation dimensions. We selected responses from Tongyi Qianwen and Zuoshou GPT, which performed well in the experiment (one general-purpose model and one domain-specific model), and calculated their Pearson correlation coefficients with response length. As shown in Table 4, we found that verbosity, complexity, and correctness are positively correlated with model response length, indicating that evaluators perceive longer model responses as more complex, verbose, but more correct. Meanwhile, fluency and helpfulness are unrelated to model response length. This suggests that in the actual design or training of models for clinical applications, there is no need to consider the length of generated content; instead, greater emphasis should be placed on the relevance of model outputs to patients' consultation content. It also confirms that when response length is considered as an additional reference factor, users actually prefer more concise and clear answers.

Correlation scores between model response length and evaluation metrics. The average character count per question for Tongyi Qianwen is 275.83 characters/question, and for Left-hand GPT it is 129.92 characters/question, with $p<0.01$.

Table4. Correlation Scores between Model Response Length and Evaluation Metrics

Metric	Tongyi Qianwen Pearson Coefficient	Left-hand GPT Pearson Coefficient
Helpfulness	-0.1874	0.0612
Correctness	0.5521	0.0124
Fluency	0.0514	0.0421
Complexity	0.4742	0.2842
Verbosity	0.7995	0.4532

5. Discussion and Conclusion

This study focuses on evaluating the practical application value of Tongyi Qianwen, ERNIE Bot, Doubao, Huatuo GPT, Zuoshou GPT, and Dr. ChatGPT in surgical outpatient consultations. Meanwhile, the research identifies key challenges and limitations, including the incompleteness of existing evaluations of large language models in the medical industry, the gap between current evaluators and real clinical consultation environments, the need to consider the diversity of patient backgrounds for medical-specific large language models, and the importance of realistically simulating clinical consultation scenarios. This paper further uncovers the safety and ethical risks of existing large language models in practical use. During the experimental evaluation, multiple instances of model hallucination were observed. Future work needs to further explore how to better integrate clinical expertise with artificial intelligence technology to ensure that large language models can maximize their advantages while safeguarding patient safety. In addition, establishing a comprehensive model evaluation system, promoting cooperation among various social sectors, and enriching the diversity of evaluators are crucial for overcoming existing obstacles and advancing the healthy development of this field. We look forward to the arrival of a more intelligent, efficient, and humanized medical era, where large language models serve as a key driving force for realizing this vision.

References

[1] Zhou, Z., Shi, J.-X., Song, P.-X., Yang, X.-W., Jin, Y.-X., Guo, L.-Z., & Li, Y.-F. (2024). LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model. ArXiv.org. https://arxiv.org/abs/2406.04614

[2] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. ArXiv:2303.17564 [Cs, Q-Fin]. https://arxiv.org/abs/2303.17564

[3] Wang, K., Ren, H., Zhou, A., Lu, Z., Luo, S., Shi, W., Zhang, R., Song, L., Zhan, M., & Li, H. (2023, October 5). MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning. ArXiv.org. https://doi.org/10.48550/arXiv.2310.03731

[4] Gan, R., Song, Y., Zhang, J., & Zhang, Y. (2024). CHIMED-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. Yuanhe Tian, 1, 7156–7173.

https://aclanthology.org/2024.acllong.386v1.pdf

[5] Wang, G., Yang, G., Du, Z., Fan, L., & Li, X. (2023b, June 16). ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. ArXiv.org. https://arxiv.org/abs/2306.09968

[6] Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., Wan, X., Wang, B., & Li, H. (2023). HuatuoGPT, towards Taming Language Model to Be a Doctor. ArXiv.org. https://arxiv.org/abs/2305.15075

[7] Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023, November 27). Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. ArXiv.org. https://doi.org/10.48550/arXiv.2311.16452

[8] Wang, Z., Dong, Y., Zeng, J., Adams, V., Sreedhar, Makesh Narsimhan, Egert, D., Delalleau, O., Scowcroft, J. P., Kant, N., Swope, A., & Kuchaiev, O. (2023). HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. ArXiv.org. https://arxiv.org/abs/2311.09528

[9] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Blaise, A., & Tomasev, N. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. https://doi.org/10.48550/arxiv.2305.09617

[10] Pal, A., Umapathi, Logesh Kumar, & Sankarasubbu, M. (2023). Med-HALT: Medical Domain Hallucination Test for Large Language Models. ArXiv.org. https://arxiv.org/abs/2307.15343

[11] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. ArXiv.org. https://arxiv.org/abs/2402.10373

[12] Singhal, K., Azizi, S., Tu, T., S. Sara Mahdavi, Wei, J., Hyung Won Chung, Scales, N., Ajay Tanwani, Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Abubakr Babiker, Schärli, N., Aakanksha Chowdhery, Mansfield, P., DemnerFushman, D., & Blaise. (2023). Large language models encode clinical knowledge. Nature, 620. https://doi.org/10.1038/s41586-023-06291-2

[13] Wong, A., Cao, H., Liu, Z., & Li, Y. (2024). SMILES-Prompting: A Novel Approach to LLM Jailbreak Attacks in Chemical Synthesis. ArXiv.org. https://arxiv.org/abs/2410.15641

[14] Cai, Y., Wang, L., Wang, Y., Melo, de, Zhang, Y., Wang, Y., & He, L. (2023). MedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models. ArXiv.org. https://arxiv.org/abs/2312.12806

[15] Liu, M., Ding, J., Xu, J., Hu, W., Li, X., Zhu, L., Bai, Z., Shi, X., Wang, B., Song, H., Liu, P., Zhang, X., Wang, S., Li, K., Wang, H., Ruan, T., Huang, X., Sun, X., & Zhang, S. (2024). MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models. ArXiv.org. https://arxiv.org/abs/2407.10990

[16] Jingnant. (2024). GitHub - jingnant/Medical-LLMs-Chinese-Exam: MLCE -A Chinese medical examination dataset summarized for the medical proficiency test of large models.

[17] GitHub. https://github.com/jingnant/Medical-LLMs-Chinese-Exam

[18] Toyhom. (2019). GitHub - Toyhom/Chinese-medical-dialogue-data: Chinese medical dialogue data Chinese Medical Dialogue Dataset. GitHub. https://github.com/Toyhom/Chinese-medical-dialogue-data

[19] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024, February 10). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. ArXiv.org. https://doi.org/10.48550/arXiv.2402.03216

[20] Krippendorff, K. (2011). Systematic disagreement Sampling errors Computing Krippendorff's Alpha-Reliability. https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf