Quantifying Atomic Knowledge in Self-Diagnosis for Chinese Medical LLMs

Anonymous ACL submission

Abstract

The booming development of medical large-scale language models (LLMs) enables users to complete preliminary medical consultations (self-diagnosis) in their daily lives. Recent evaluations of medical LLMs mainly focus on their ability to complete medical tasks, pass medical examinations, or obtain a favorable GPT-4 rating. There are still challenges in using them to provide directions for improving medical LLMs, including misalignment with practical use, lack of depth in exploration, and over-reliance on GPT-4. To address the above issues, we construct a fact-checking style Self-Diagnostic Atomic Knowledge (SDAK) benchmark. Through atomic knowledge that is close to real usage scenarios, it can more accurately, reliably, and fundamentally evaluate the memorization ability of medical LLMs for medical knowledge. The experimental results show that Chinese medical LLMs still have much room for improvement in self-diagnostic atomic knowledge. We further explore different types of data commonly adopted for fine-tuning medical LLMs and find that distilled data enhances medical knowledge retention more effectively than real-world doctor-patient conversations.

1 Introduction

In the digital age, seeking health information from the Internet for self-diagnosis has become a common practice of patients (White and Horvitz, 2009; Demner-Fushman et al., 2019; Farnood et al., 2020). During the self-diagnosis, the searched health information can assist users in making necessary medical decisions, such as self-treatment or going to the hospital for professional treatment. With the development of generative models (Ouyang et al., 2022; Sun et al., 2021; OpenAI, 2023), Large-scale Language Models (LLMs) hold the promise of revolutionizing the retrieval paradigm that seeks health suggestions via a search engine because they can provide more efficient suggestions through natural conversations.

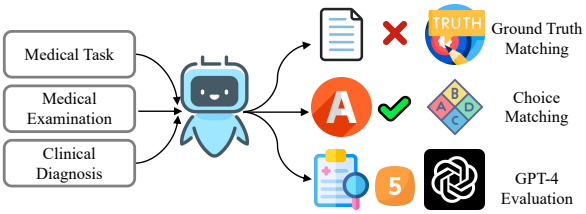

Figure 1: Widely used medical evaluation methods. The medical task mainly measures the ability of LLMs to complete the task, the medical examination explores the ability of LLMs to pass the examination, and the clinical diagnosis assesses the diagnosis ability of LLMs by using GPT-4 as the judgment.

To enhance the medical capabilities of open-source LLMs in Chinese, recent studies (Wang et al., 2023a; Zhang et al., 2023; Zhu and Wang, 2023; Yang et al., 2023) attempt to fine-tune the foundation models on medical instruction or conversation data. As for the methods for evaluating their performance, the existing work is mainly divided into three categories: medical NLP-related tasks (Zhu et al., 2023), medical exams (Umapathi et al., 2023; Wang et al., 2023d), and evaluating medical dialogue evaluations through GPT-4 (Zhang et al., 2023; Yang et al., 2023), as shown in Figure 1.

Challenges. Despite the progress of evaluation, there are still some challenges in using them to provide directions for improving medical LLMs: (1) *Misalignment with practical use*. Most current Chinese medical LLMs are patient-centric, typically addressing questions related to medical consultations rather than complex and professional queries, such as "*What should I take for a cold?*"(感冒应该吃什么药？). The results from these evaluations, such as NLP tasks or medical exams, do not match the actual needs of users. (2) *Lack of depth in exploration*. Since most evaluations simply judge

whether the model's responses to complex questions are correct or incorrect, it is challenging to determine whether errors stem from basic memorization failures or a lack of advanced reasoning abilities in LLMs (Zheng et al., 2023). (3) *Over-reliance on GPT-4 for evaluation*. Evaluation by GPT-4 is not satisfactory because of its evaluation bias (Wang et al., 2023c) and its insufficient medical knowledge (seen in Figure 4).

Solutions. To address the above limitations, we propose a fact-checking style medical benchmark named the Self-Diagnostic Atomic Knowledge Benchmark (SDAK) to assess Chinese medical LLMs. Inspired by atomic fact-checking (Chern et al., 2023), we utilize atomic knowledge (Min et al., 2023), an indivisible unit of information, for a more precise, reliable, and fundamental evaluation of an LLM's proficiency in medical knowledge (examples are shown in Table 2). To ensure that the evaluation is closer to the real usage scenario of medical LLMs, we adopt thematic analysis (Braun and Clarke, 2012; Zheng et al., 2023) to extract the most commonly used atomic knowledge types from self-diagnostic queries. Then, we create atomic knowledge items for each type according to structured medical contents from public medical websites, each item consists of a pair of factual and counterfactual claims. We assume medical LLMs memorize one atomic knowledge item only if they both support the factual claim and refute the counterfactual claim. To reduce reliance on GPT-4, we designed two **necessary automation** indicators (instruction following rate, factual accuracy) and an **optional manual** metric (accuracy reliability). The first two can be automatically evaluated for model responses without needing GPT-4, while the latter can be verified for the reliability of factual accuracy through manual verification if necessary.

Results. The experimental results show that: (a) the instruction following ability of most medical LLMs fine-tuned with domain data decreased to varying degrees compared to general LLMs, and the memorization ability of LLMs in the medical domain was not significantly improved; (b) the reliability of the answers after manual verifying mostly exceed 95% indicating our metric is reliable to measure the memorization ability of LLMs.

Findings. After an in-depth analysis of error types, knowledge types, and data sources for fine-tuning, we find the following three points: (1) Sycophancy is the primary cause of errors, whether it is in general or medical LLMs. (2) There is still a huge gap between the existing Chinese medical LLMs and GPT-4, although GPT-4 performs poorly in some more specialized medical knowledge. (3) Compared to real doctor-patient conversation data, distilled data from the advanced LLMs can better help open-source LLMs memorize more atomic knowledge. We believe that this is due to the fact that doctors are less likely to explain medical knowledge and diagnosis to patients in real doctor-patient conversations. The above insights could provide future research directions for the Chinese medical LLMs community. Our data, code, and model will be released in <AnonymousURL>.

2 Related Work

2.1 Medical Evaluation Methods

The existing efforts put into the evaluation of the medical abilities of LLMs are mainly divided into three types: medical NLP-related tasks (Zhu et al., 2023), medical exams (Umapathi et al., 2023; Wang et al., 2023d), and conducting medical dialogue evaluations through GPT-4 (Zhang et al., 2023; Yang et al., 2023). However, inconsistent scenarios, lack of depth exploration, and insufficient medical ability of GPT-4 pose new challenges to evaluating medical LLMs in Chinese. In this paper, we aim to address the above limitations and explore the memorization ability of LLMs in the self-diagnostic scenario.

2.2 Fact-checking

The fact-checking task (Thorne et al., 2018; Guo et al., 2022; Wadden et al., 2020; Saakyan et al., 2021; Sarrouti et al., 2021; Mohr et al., 2022) aims to determine whether the claims are supported by the evidence provided, which has been an active area of research in NLP. Recently, some researchers (Min et al., 2023; Chern et al., 2023) have paid more attention to automatically evaluating the factuality of atomic knowledge contained in the long-form model-generated text. They utilized GPT-4 to automatically decompose atomic facts in complex texts and verify the overall factual accuracy. However, this method is not applicable in the medical domain due to the insufficient mastery of medical knowledge in GPT-4, which cannot extract critical facts in user queries like extracting common sense.

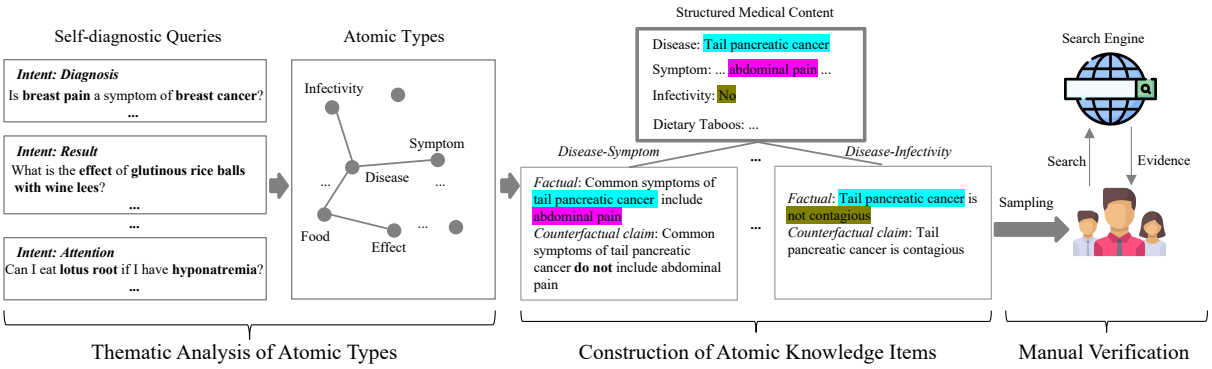

Figure 2: Construction process of self-diagnostic atomic knowledge benchmark.

2.3 Chinese Medical LLMs

To enhance the medical capability of open-source LLMs (Du et al., 2022; Touvron et al., 2023a,b; Baichuan, 2023), previous work has attempted to adopt real-world medical data or the mixture of real-world and distilled/semi-distilled from ChatGPT conversations (Wang et al., 2023e,b) for fine-tuning. The former (Xu, 2023; Wang et al., 2023a,b) mainly learn the medical capabilities of doctors from doctor-patient conversations, while the latter (Zhu and Wang, 2023; Yang et al., 2023; Zhang et al., 2023) further additionally added distilled conversations from advanced LLMs such as ChatGPT. Despite there being much progress in medical LLMs in Chinese, how to better evaluate their performance is still an area that needs to be studied, such as the extent of self-diagnostic medical knowledge stored in these LLMs.

3 Construction of Self-diagnostic Atomic Knowledge Benchmark

Motivation. Despite the robust growth of Chinese medical LLMs, various evaluations for them have yet to be significantly helpful in improving them. On the one hand, some existing evaluations focus on medical NLP tasks (Zhu et al., 2023) or medical exams (Umapathi et al., 2023; Wang et al., 2023d), which do not align with real usage scenarios (self-diagnosis). Moreover, due to the complexity of the testing questions, it is challenging to determine whether the model's errors stem from issues in memory, or reasoning, which are crucial for improving LLMs' performance. On the other hand, some efforts (Zhang et al., 2023; Yang et al., 2023) have attempted to use GPT-4 in a conversational format for evaluation. However, due to GPT-4's inherent evaluation biases, its imperfect grasp of medical knowledge, and limitations in accessibility, this method is also not suitable. Therefore, inspired by atomic fact-checking evaluation studies, we build a fact-checking style medical benchmark named the Self-Diagnostic Atomic Knowledge Benchmark (SDAK) to more accurately, reliably, and fundamentally evaluate the memorization ability of medical LLMs for medical knowledge, as shown in Figure 2.

3.1 Thematic Analysis of Atomic Types

To obtain the most common types of atomic knowledge for queries of real users in the self-diagnostic scenario, we select the KUAKE-QIC (Zhang et al., 2022) dataset as the source data. It mainly contains user queries from search engines with ten intent types, and examples are shown in Appendix A.

Then, we conducted thematic analysis (Braun and Clarke, 2012; Zheng et al., 2023) of 200 samples randomly selected from each intent type in KUAKE-QIC to identify the atomic knowledge types. Specifically, we first conduct the induction by initiating the preliminary type of atomic knowledge for each selected sample, where we mainly focus on medical-related knowledge, specializing in Disease-Symptom, Medicine-Effect, etc. Then, we deduce the most common type of atomic knowledge by aggregating the type into a broader atomic type if more samples fall into this type. Take the query with *Diagnosis* intent in Figure 2 as an example. Since both *breast pain* and *breast cancer* in this query are the symptom and disease, respectively, the atomic type involved in this query is Disease-Symptom.

Table 1 shows the atomic types and percentages contained in the queries with various intents we constructed. We find that over 80% of queries in each intent fall into different atomic types we deduced, indicating that atomic knowledge is a more fine-grained basic unit. Besides, the queries with different intents tend to involve the same type of

Intent	Atomic Type	Percentage
Diagnosis	Disease-Symptom	81%
	Disease-Examination	10%
Cause	Disease-Cause	64%
	Disease-Symptom	25%
Method	Disease-Medicine	55%
	Disease-Method	34%
Advice	~~Disease-Hospital~~	80%
	Disease-Department	8%
	Disease-Examination	11%
Metric_explain	Examination-Range	63%
	Metric-Effect	37%
Disease_express	Disease-Symptom	62%
	Disease-Infectivity	15%
	Diseases-Complication	15%
Result	Disease-Symptom	36%
	Western Medicine-SideEffect	14%
	Chinese Medicine-SideEffect	19%
	Food-Effect	17%
Attention	Disease-Food	59%
	Disease-Prevention	21%
Effect	Western Medicine-Effect	20%
	Chinese Medicine-Effect	27%
	Food-Effect	44%
Price	~~Treatment-Price~~	97%

Table 1: Major types and percentages of atomic knowledge contained in each intent of self-diagnostic queries.

atomic knowledge, e.g., queries with both *Diagnosis* and *Cause* intents involve the same atomic type of Disease-Symptom, which demonstrates the necessity and efficiency of evaluating LLMs in terms of atomic knowledge. After removing the non-objective intent related to specific user locations, such as *Price* and *Advice*, we collect 17 most common types of atomic knowledge from real-world self-diagnostic queries, as shown in Table 1.

3.2 Construction of Atomic Knowledge Items

After obtaining the most common atomic types, we construct pairs of factual and counterfactual claims for each atomic type to convey atomic knowledge items. To avoid data contamination, we do not construct atomic claims based on existing Chinese medical knowledge graphs, e.g., CMeKG (Odmaa et al., 2019) has been utilized by some Chinese medical LLMs (Wang et al., 2023a,b). Instead, we manually build atomic knowledge items according to the structured medical content from the public medical websites[1] for the following two reasons. On the one hand, the medical content from these websites is reliable because it is edited and verified by professional medical teams. On the other hand, these websites are also the main source of medical knowledge for self-diagnostic queries.

[1]https://www.xiaohe.cn/medical and https://www.120ask.com/disease

Atomic Type	Example of factual (Counterfactual) atomic claim
Disease-Symptom	Common symptoms of tail pancreatic cancer (do not) include abdominal pain 胰尾癌的常见症状（不）包括腹痛
Disease-Infectivity	Laryngeal cysts are (not) contagious 喉囊肿（不）具有传染性
Disease-Department	Common departments for Psoriatic A (do not) include dermatology 银屑病甲常挂的科室（不）包括皮肤科
Disease-Method	Common treatments for prolactinomas (do not) include radiation therapy 催乳素瘤常见治疗方法（不）包括放射治疗
.....	.....
Disease-Medicine	Common medications for stomatitis (do not) include metformin 口腔炎的常用药物包括（不包括）甲氰咪胍

Table 2: Example of each type of atomic knowledge. The complete table is shown in Appendix B.

As shown in Figure 2, we first extract the atomic knowledge from the structured medical content according to the atomic types we build. For example, we extract the disease *Tail pancreatic cancer (*尾胰癌*)* and symptom *abdominal pain (*腹痛*)* for the Disease-Symptom atomic type. Then, we heuristically construct a factual claim in the form of implication relation, as shown in Table 2.

Given that LLMs may exhibit a sycophantic bias (Wei et al., 2023; Du et al., 2023), e.g., it always supports the user's claims, it is unreliable to explore the amount of self-diagnostic knowledge stored in LLMs' memory merely by whether or not LLMs supports factual claims. To avoid this, we propose using contrastive evaluation based on constructing a counterfactual claim for each factual claim by converting the *implication* into a *non-implication* relation, as shown in Table 2. LLMs are considered to possess one atomic knowledge item only if they both support the factual claim and refute the counterfactual claim. For each atomic type, we randomly selected at most 1,000 structured medical content to build atomic knowledge items and the statistics are shown in Appendix B.

3.3 Manual Verification

To verify the reliability of atomic claims, we conducted the manual verification based on the evidence retrieved through a search engine. We first randomly selected 50 factual claims for each atomic type. Then, we verify the correction of the claims. Follow the previous work (Chern et al., 2023) and retrieve evidence by feeding factual claims into a search engine [2]. The top 10 items retrieved by the search engine as evidence and manually judge whether the evidence supports the factual claims.

Table 3 shows the results of manual verification, where *Support*, *Neural*, and *Refute* indicate that evidence supports claims, insufficient evidence,

[2]The search engine we adopted is Baidu, which is one of the most popular Chinese search engines.

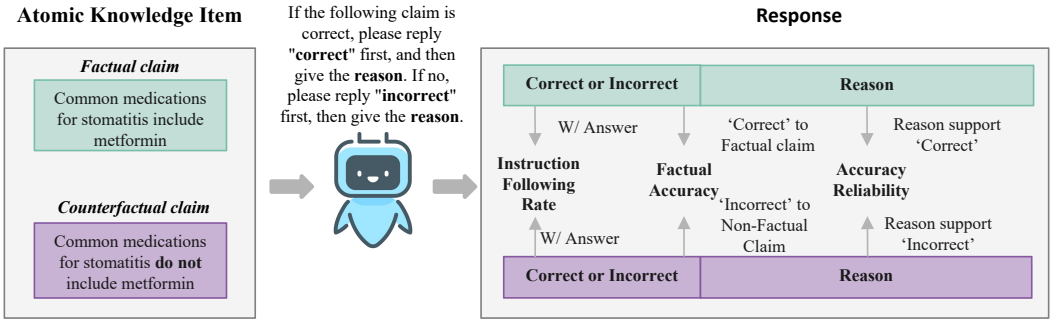

Figure 3: Process of the fact-checking style evaluation method.

Atomic Type	Number		
	Support	Neural	Refute
Metric-Effect	43	6	1
Disease-Infectivity	42	5	3
Disease-Department	48	0	2
Disease-Method	45	5	0
Disease-Cause	46	4	0
Chinese Medicine-Effect	48	1	1
Chinese Medicine-SideEffect	46	1	3
Western Medicine-Effect	50	0	0
Western Medicine-SideEffect	44	3	3
Food-Effect	43	5	2
Disease-Examination	45	0	5
Disease-Prevention	33	11	6
Diseases-Complication	42	7	1
Disease-Symptom	48	2	0
Examination-Range	32	12	6
Disease-Food	47	3	0
Disease-Medicine	46	1	3
Total	748	66	36
Percentage	88.00%	7.76%	4.24%

Table 3: Manual verification of atomic knowledge items.

and evidence refutes claims, respectively. 88% of claims can be fully supported by the evidence [3] and only 4% are refuted, which shows the reliability of the atomic claims we constructed. In addition, the reliability of about 8% of factual claims cannot be verified due to insufficient evidence. We attribute it to the fact that these pieces of atomic knowledge are relatively low-frequency, leading to search engines failing to retrieve the related evidence.

4 Experiments

4.1 General and Medical LLMs for Evaluation

We select the following popular general LLMs and Chinese medical LLMs for evaluation on our SDAK. In addition to the closed-sourced ChatGPT and GPT-4 (OpenAI, 2023) models, we select representative open-source Chinese LLMs such as Baichuan2 (Baichuan, 2023), Qwen (Bai et al., 2023), and ChatGLM2 (Du et al., 2022) for evaluation. As for the Chinese medical LLMs, we select two types of models:

Fine-tuned merely on real-world data: BenTsao (Wang et al., 2023a), ChatGLM-Med (Wang et al., 2023b), MedicalGPT (Xu, 2023).

Fine-tuned on mixed data: Chatmed-Consult (Zhu and Wang, 2023), HuatuoGPT (Zhang et al., 2023), and Zhongjing (Yang et al., 2023).

We conducted the experiment in zero-shot and few-shot settings, and Appendix D introduces the details about few-shot setting. Appendix E introduces the hyperparameter settings of each model.

4.2 Fact-checking Style Evaluation Method

To comprehensively evaluate the performance of LLM on the SDAK benchmark, we propose the fact-checking style evaluation method, as shown in Figure 3.

4.2.1 Evaluation Prompt

Firstly, for a pair of claims for each atomic knowledge in SDAK, we designed an appropriate prompt to instruct LLMs to output as we specified to evaluate the performance of LLMs. The prompt is as follows: *If the following claim is correct, please reply "correct" first, and then give the reason. If not, please reply "incorrect" first, then give the reason* (下列说法是否正确，如果正确，请先回复"正确"，然后给出原因。如果错误，请先回复"错误"，然后给出原因。). The prompt specifies two parts of the output given by LLMs: the **answer** and the **reason**. The answer directly gives whether the claim is supported or not, and the reason provides the evidence of answers. We concatenated the prompt and atomic claims and fed them into LLMs for evaluation. Refer to Appendix C for the exploration of different prompts.

4.2.2 Evaluation Metrics

To evaluate the performance of LLMs in processing atomic knowledge, we developed two **necessary automatic** metrics: Instruction Following Rate (IFR) and Factual Accuracy (FactAcc), and

[3] We also asked a professional doctor to verify 170 factual claims (each atomic type contains 10) and found there are 87% of claims that can be supported.

Domain	Data	LLMs	Zero-shot			Few-shot		
			IFR(%)	FactAcc(%)	AccR(%)	IFR(%)	FactAcc(%)	AccR(%)
General	-	GPT-4	$99.96_{\pm 0.00}$	$65.42_{\pm 0.60}$	100	$100_{\pm 0.00}$	$72.61_{\pm 0.33}$	100
		Qwen-14b-Chat	$100_{\pm 0.00}$	$57.29_{\pm 0.03}$	98	$100_{\pm 0.00}$	$67.34_{\pm 0.56}$	100
		ChatGPT	$99.97_{\pm 0.00}$	$51.72_{\pm 0.40}$	97	$100_{\pm 0.00}$	$56.93_{\pm 0.79}$	99
		Qwen-7b-Chat	$100_{\pm 0.00}$	$43.68_{\pm 0.10}$	98	$100_{\pm 0.00}$	$56.74_{\pm 0.30}$	100
		Baichuan2-13b-Chat	$99.71_{\pm 0.00}$	$42.01_{\pm 0.05}$	96	$100_{\pm 0.00}$	$52.09_{\pm 0.45}$	99
		ChatGLM2	$99.84_{\pm 0.01}$	$37.90_{\pm 0.04}$	97	$100_{\pm 0.00}$	$47.17_{\pm 2.13}$	100
		Baichua2-7b-Chat	$99.89_{\pm 0.01}$	$16.14_{\pm 0.09}$	95	$100_{\pm 0.00}$	$35.15_{\pm 2.52}$	98
Medical	Mixed	Zhongjing	$90.22_{\pm 0.17}$	$24.78_{\pm 0.10}$	97	$93.59_{\pm 0.80}$	$29.56_{\pm 3.65}$	100
		Chatmed-Consult	$85.10_{\pm 0.14}$	$24.50_{\pm 0.34}$	98	$95.32_{\pm 2.31}$	$27.15_{\pm 2.91}$	99
		HuatuoGPT	$99.73_{\pm 0.00}$	$16.15_{\pm 0.01}$	98	$99.90_{\pm 0.62}$	$26.63_{\pm 1.52}$	100
	Real	MedicalGPT	$76.04_{\pm 0.50}$	$7.86_{\pm 0.50}$	100	$85.12_{\pm 0.74}$	$11.37_{\pm 2.09}$	100
		ChatGLM-Med	$94.91_{\pm 0.07}$	$7.46_{\pm 0.15}$	75	$97.79_{\pm 0.68}$	$9.41_{\pm 0.89}$	93
		BenTsao	$84.43_{\pm 0.06}$	$3.35_{\pm 0.07}$	70	$89.37_{\pm 3.20}$	$7.26_{\pm 3.49}$	96

Table 4: The performance of general and medical LLMs on self-diagnostic atomic knowledge. The subscript represents the standard deviation after three experiments.

an **optional manual** metric: Accuracy Reliability (AccR). These metrics collectively assess an LLM's ability to process and respond to medical information accurately and reliably. **Instruction Following Rate (IFR)** assesses whether LLMs can adhere to the given instructions. An LLM is considered to follow instructions if it provides answers (be it correct or incorrect) to both factual and counterfactual atomic claims at the start of its response. **Factual Accuracy (FactAcc)** measures the abilities of LLMs on self-diagnostic atomic knowledge. LLMs are considered to memorize the atomic knowledge if they give the answer *'correct'* to the factual claim and *'incorrect'* to the counterfactual claim of an item. **Accuracy Reliability (AccR)** evaluates the reliability of factual accuracy. We randomly selected 100 atomic knowledge items and manually checked the model's responses. If the reason given by LLMs can support the answer '*correct*' to a factual claim and the answer '*incorrect*' to a counterfactual claim, we believe that the answers given by LLMs are reliable.

4.3 Evaluation Results

The performance of each model on SDAK is shown in Table 4. A key finding is that while general LLMs maintain an instruction-following rate above 99% in zero-shot and few-shot settings, most medical LLMs show a 5%-15% decline in zero-shot setting. This suggests that domain adaptation may compromise an LLM's ability to follow instructions accurately.

In terms of factual accuracy (FactAcc) in the zero-shot setting, GPT-4 unsurprisingly achieves the best performance of 65.42% among all LLMs. Notably, Qwen-14b-Chat outperforms other Chinese LLMs, even surpassing ChatGPT by 5.57%. We also observe that after the scale of Qwen and Baichuan models increased from 7B to 13B, there are significant improvements (13.61% and 25.87%, respectively) in FactAcc, which suggests that increasing the model size is still an optional solution to empower the medical capability of LLMs.

Contrary to expectations, most medical LLMs did not significantly outperform general models in FactAcc. Where Zhongjing, Chatmed-Consult, and HuatuoGPT surpass the Baichuan2-7b-chat, and their best performance (Zhongjing) in the FactAcc only reaches 24.78% in the zero-shot setting. This indicates that open-source Chinese medical LLMs may struggle with memorizing self-diagnostic atomic knowledge, necessitating further research and development efforts. The significant differences in medical models with different training data prompted us to conduct an in-depth analysis of the impact of different training data, as shown in Section 5.3.

In addition, in the few-shot setting, the performance of most models on all metrics is improved significantly, indicating that in-context learning can effectively improve the abilities of models on instruction-following and self-diagnostic atomic knowledge.

Finally, after manually checking the answers provided by various models, we find that both the general LLMs and the medical LLMs can provide a good basis for the answers, with most of them achieving over 95% performance in Accuracy Reliability (AccR). It also proves that FactAcc can reliably reflect the LLMs' memorization ability of self-diagnostic atomic knowledge.

5 Analysis

The analysis is conducted in the zero-shot setting.

Domain	LLMs	Error Type			
		NotFollow	Sycophancy	Safety	Misinterpretation
General	GPT4	0	68	26	6
	Qwen-14b-Chat	0	68	24	8
	ChatGPT	0	79	17	4
	Qwen-7b-Chat	0	74	20	6
	Baichuan2-13b-Chat	0	72	24	4
	ChatGLM2-6b	0	70	25	5
	Baichuan2-7b-Chat	0	74	21	5
Medical	Chatmed-Consult	20	48	18	14
	Zhongjing	8	62	18	12
	HuatuoGPT	0	64	22	14
	MedicalGPT	23	62	2	13
	ChatGLM-med	5	54	1	40
	BenTsao	5	90	5	0

Table 5: Error analysis of LLMs on atomic knowledge.

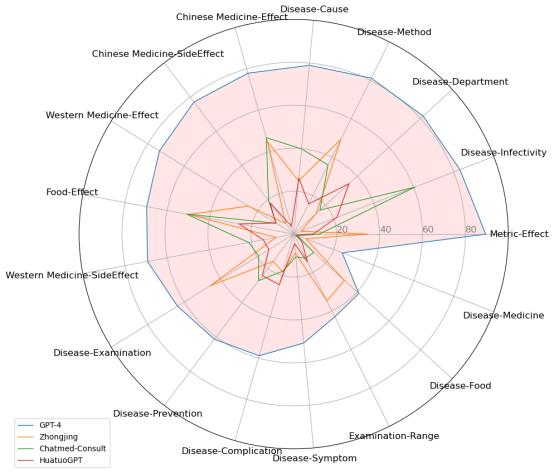

Figure 4: Performance of representative LLMs on various types of atomic knowledge.

5.1 Error Analysis on Atomic Knowledge

We conducted a detailed analysis of errors to gain insights into the challenges LLMs face in memorizing medical atomic knowledge. We randomly selected 100 atomic knowledge items where various models provided incorrect responses, as shown in Table 5. This analysis revealed four primary error categories: **NotFollow**, where LLMs either evade directly answering ('correct' or 'incorrect') or provide irrelevant information; **Sycophancy**, characterized by LLMs indiscriminately supporting both factual and counterfactual claims, distinct from mere bias or agreeability; **Safety**, LLMs argue that claims are not strictly expressed and provide a more cautious answer; and **Misinterpretation**, where LLMs erroneously treat counterfactual claims as factual. Appendix F shows the examples of each type.

Table 5 shows that the proportion of 'NotFollow' responses aligns with the Instruction Following Rate (IFR) in Table 4, underscoring the effectiveness of this metric in our evaluation. Notably, in the samples where LLMs followed instructions, 'Sycophancy' emerged as the predominant error type. This finding echoes previous research (Sharma et al., 2023) and underscores the need for contrastive evaluation to verify LLMs' grasp of medical atom knowledge: Simply measuring an LLM supporting a factual claim correctly does not necessarily indicate that it has mastered the knowledge, but may be caused by sycophancy. Our results also highlight a tendency for general LLMs to adopt more cautious stances in responses, a pattern especially pronounced in models like Chatmed-Consult, Zhongjing, and HuatuoGPT, which were trained on mixed datasets, including distilled data from ChatGPT. In contrast, domain-specific medical LLMs displayed a higher rate of 'Misinterpretation', suggesting an increased internal inconsistency post domain adaptation training.

5.2 Performance of LLMs on Various Types of Atomic Knowledge

We further plotted the FactAcc of various models on different types of atomic knowledge through a radar graph in Figure 4. It reveals that GPT-4 demonstrates robust performance in medical common sense types, as indicated in the upper half of Figure 4, with achievements surpassing or closing to 80%. In contrast, GPT-4's performance declines in more specialized atomic knowledge areas, such as Disease-Medicine and Disease-Food interactions, located in the lower right part of Figure 4. We also observed that Chinese medical LLMs, despite their advancements, still lag behind GPT-4 in all atomic knowledge types. Additionally, we noticed that various models exhibit similar perfor-

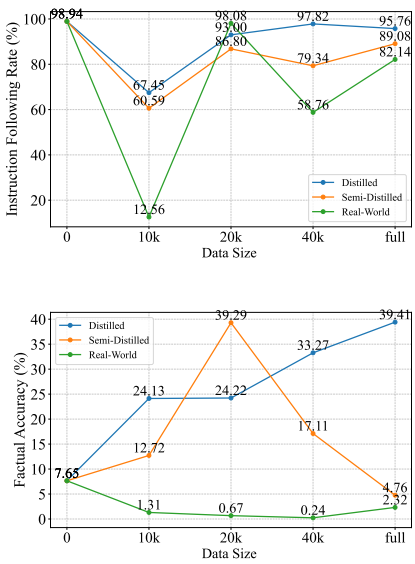

Figure 5: Performance of the LLM in the IFR and FactAcc metrics with different types of data.

mance levels on certain atomic knowledge items due to sharing part of datasets. Therefore, we suggest that Chinese medical LLM needs more differentiated development in the future.

5.3 Effect of Different Types of Training Data

In the above experiments, we observed a notable performance enhancement in models trained with a mix of data types compared to those relying solely on real-world doctor-patient conversations. This led us to explore further the influence of different data sources on both the IFR and FactAcc, as shown in Figure 5. For a controlled comparison, we fine-tuned models using distilled, semi-distilled, and real-world data sets on the same base model, Baichuan-7b-base. Specifically, we utilized 69,768 real-world and 61,400 distilled single-turn conversations from HuatuoGPT and 549,326 semi-distilled single-turn conversations from ChatmedConsult. The experimental setting can be seen in Appendix G.

In the upper segment of Figure 5, we illustrate how different training datasets impact the IFR. The base model, trained exclusively on general conversation data, exhibited a high instruction following rate (98.94%). However, introducing medical datasets (10K conversations) initially led to a significant decline in IFR due to the cost of domain adaptation. Notably, when the training data exceeded 20K samples, the IFR progressively improved, signifying successful domain adaptation via sufficient domain data. Intriguingly, models trained on distilled data outperformed those trained on real-world conversation data in terms of IFR. This could be attributed to the nature of real doctor-patient interactions, which are more dialogic and less instructional, thus less effective for training models in instruction following.

The lower part of Figure 5 examines the impact of these data types on FactAcc. Training with increased proportions of distilled data from ChatGPT led to a consistent enhancement in FactAcc (from 7.65% with no medical data to 39.41% with full medical data). In contrast, models trained solely on real-world data struggled to assimilate medical knowledge effectively. We believe this is because ChatGPT often adds additional explanations to its answers in order to better serve humans, while in real doctor-patient conversations, doctors rarely explain the basis and approach of their diagnosis to patients. Furthermore, the performance of models trained on semi-distilled data displayed notable fluctuations. Initially, with 20K training samples, these models achieved a peak FactAcc of 39.29%, even surpassing those trained on 549K samples from Chatmed Consult. However, further increasing the training sample size resulted in a decrease in FactAcc. This decline could be linked to the presence of more low-quality real-user queries in the semi-distilled data.

6 Conclusion

In this paper, we build the Self-Diagnostic Atomic Knowledge (SDAK) benchmark to evaluate atomic knowledge in open-source Chinese medical LLMs. It contains 14,048 atomic knowledge items across 17 types from user queries, each comprising a pair of factual and counterfactual claims. Then, we designed two necessary automatic evaluation metrics (instruction following rate and factual accuracy) and an optional manual evaluation metric (accuracy reliability) to evaluate the Chinese medical LLMs comprehensively. Experimental results revealed that while these LLMs show promise, they are not yet on par with GPT-4, particularly in some more professional medical scenarios. We also found that these models' errors often stem from sycophantic tendencies and that distilled data enhances medical knowledge retention more effectively than real doctor-patient conversations. We hope the SDAK benchmark and our findings can prompt the development of Chinese medical LLMs.

Limitations

The main limitation is the limited size of the SDAK benchmark. Since the application of LLMs is extremely time-consuming and resource-intensive, we have to limit the size of the benchmark, leading it to hardly cover all atomic medical knowledge comprehensively in self-diagnosis scenario. However, it is worth noting that our method can easily expand the size of SDAK benchmark if computing resources are no longer a problem impeding LLMs in the future. We also acknowledge that the quality of our dataset is not perfect, although only 4% of the samples do not match objective facts. We will try to make up for this deficiency in future research. In addition, the SDAK benchmark we have built serves as a medical LLMs evaluation for Chinese, but its paradigm is language-independent and can be easily transferred to other languages such as English, French, Japanese, etc. Furthermore, although we have taken measures to avoid creating test data from existing data as much as possible, we acknowledge that it is still impossible to completely avoid the possibility of data leakage.

Ethics Statement

The main contribution of this paper is establishing the SDAK benchmark to quantify the self-diagnostic atomic knowledge in Chinese Medical Large Language Models. This benchmark is built using heuristic rules based on medical knowledge publicly available on the Internet. The data sources are all ethical. Firstly, the atomic knowledge types utilized in our study were sourced from KUAKE-QIC, which is a public dataset that can be accessed freely. Secondly, we only extract the related medical terms (such as medication name and disease name) from medical encyclopedia entries on the third-party medical website, which are public medical knowledge and can be found in many medical resources like Wikipedia or Baidu Baike and do not contain any information that uniquely identifies individuals. Therefore, it does not violate dataset copyright and privacy information.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2019. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2):194–201.

Yanrui Du, Sendong Zhao, Muzhen Cai, Jianyu Chen, Haochun Wang, Yuhan Chen, Haoqiang Guo, and Bing Qin. 2023. The calla dataset: Probing llms' interactive knowledge acquisition from chinese medical literature.

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

Annabel Farnood, Bridget Johnston, and Frances S Mair. 2020. A mixed methods systematic review of the effects of patient online self-diagnosing in the 'smartphone society'on the healthcare professional-patient relationship and medical authority. *BMC Medical Informatics and Decision Making*, 20:1–14.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

BYAMBASUREN Odmaa, Yunfei YANG, Zhifang SUI, Damai DAI, Baobao CHANG, Sujian LI, and Hongying ZAN. 2019. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*, 33(10):1–7.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

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

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med: 基于中文医学知识的chatglm模型微调. https://github.com/SCIR-HI/Med-ChatGLM.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023d. Cmb: A comprehensive medical benchmark in chinese.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023e. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models.

Ryen W White and Eric Horvitz. 2009. Experiences with web search on medical concerns and self diagnosis. In *AMIA annual symposium proceedings*, volume 2009, page 696. American Medical Informatics Association.

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

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical

capabilities of large language model through expert feedback and real-world multi-turn dialogue.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language model to be a doctor.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Wei Zhu and Xiaoling Wang. 2023. Chatmed: A chinese medical large language model. `https://github.com/michael-wzhu/ChatMed`.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. Promptcblue: A chinese prompt tuning benchmark for the medical domain.

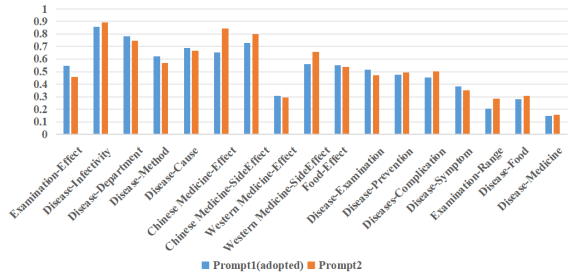

Figure 6: Accuracy of ChatGPT on various types of atomic knowledge with different prompts.

A Examples of Different Types of Query

Table 6 shows the self-diagnostic queries with different intents.

B Statistics of the SDAK Benchmark

Table 7 shows the statistics of our SDAK Benchmark.

C Performance of ChatGPT with Different Prompts

Although we did not conduct prompt engineering in depth, we study the effect of simple prompts that provide the same instruction on ChatGPT's performance of self-diagnostic atomic knowledge, as shown in Figure 6. The detail of prompt1 is shown in Section 4.2 and the prompt2 is as follows: If the following statements about medical knowledge are correct, please first output "correct" or "incorrect" and then give the corresponding reasons on a separate line.(下列关于医学知识的说法是否正确，请先输出“正确”或“错误”，然后另起一行给出相应的原因。). From Figure 6, we can see that there is no significant performance difference between prompt1 and prompt2 on various types of atomic knowledge. This indicates that LLMs are not sensitive to simple prompts that provide the same instruction.

D Few-shot Experiments

For few-shot learning, we follow the previous work (Wang et al., 2023d) and provide three demonstrations. We first constructed a validation set as the source of few-shot examples. Specifically, we randomly constructed another 10 atomic knowledge items for each of the 17 atomic types, with each item comprising a pair of factual and counterfactual claims. This process resulted in a comprehensive validation dataset of 340 claims. Subsequently, we randomly selected three claims and got responses from GPT4 with our evaluation prompt. To ensure the reliability of GPT-4's outputs, we engaged a professional medical doctor for the verification and correction of any erroneous responses. Then, we conducted three-time experiments, each with three claims randomly selected from the validation dataset as few-shot examples.

The forms of the few-shot prompt are as follows:

```
If the following claim is correct, please
reply "correct" first, and then give the
reason. If not, please reply "incorrect"
first, then give the reason.

————————

Input:<ex. 1>
Output: <response 1>
————————

————————
Input: <ex. 2>
Output: <response 2>
————————

————————
Input: <ex. 3>
Output: <response 3>
————————

Input: <testing>
Output:
```

E Hyper-parameters

For ChatGPT and GPT-4, we adopted the *GPT-3.5-turbo-0301* and *GPT-4-0314* version, respectively, and the generation settings are set by default. For other open-source generic LLMs and medical LLMs, we adopted the same generation settings as Baichuan2 (Baichuan, 2023) for a fair comparison. The temperature, top_k, top_p, and repetition_penalty are set to 0.3, 5, 0.85, and 1.05, respectively, and other parameters are set by default. All experiments for each LLM are conducted three times, and we report the mean and standard deviation values.

F Error Types of LLMs on Atomic Knowledge

Examples of each error type are shown in Table 8-11. Table 8 shows the example of the NotFollow error type in that LLMs do not follow the instruction

Intent	Query	Atomic Type
Diagnosis	Is breast pain a symptom of breast cancer? 乳房疼痛是不是乳腺癌?	Disease-Symptom
	Is the high neutrophil of blood image classification cell bacterial infection? 血象分类细胞的中性白细胞偏高是细菌感染吗?	Disease-Examination
Cause	What is the cause of pancreatic cancer? 胰腺癌的原因是什么?	Disease-Cause
Method	What is the best medicine for high blood pressure? 高血压吃什么药好?	Disease-Medicine
	What is the treatment for osteochondritis dissecans of the knee? 膝盖骨膜炎的治疗办法是什么?	Disease-Method
Advice	Where is the best hospital of treating rectal cancer in Anhui Province? 安徽最好的治直肠癌医院在哪里?	Disease-Hospital
	What section does mouth herpes hang? 口腔疱疹挂什么科?	Disease-Department
Metric_explain	How much H. pylori is considered excessive? 幽门螺杆菌多少算超标?	Examination-Range
	What does a urine test for red blood cells mean? 尿检红细胞是什么意思?	Metric-Effect
Diseases_express	Is hemorrhagic fever contagious? 出血热传染吗?	Disease-Infectivity
	Can cervical infection with hpv virus cause low fever? 宫颈感染hpv病毒能引起低热吗?	Disease-Complication
Result	Does taking tenofovir cause high blood pressure? 服用替诺福韦会引起高血压吗?	Western Medicine-SideEffect
	Does taking Jinkui kidney Qi pill have adverse reaction? 服用金匮肾气丸有不良反应吗?	Chinese Medicine-SideEffect
	What is the effect of glutinous rice balls with wine lees? 酒糟汤圆的功效是什么?	Food-Effect
Attention	Can I eat lotus root if I have hyponatremia? 低钠血症患者能吃莲藕吗?	Disease-Food
	How to prevent psoriasis? 怎么预防牛皮癣?	Disease-Prevention
Effect	What do furosemide tablets do? 呋塞米片的作用是什么?	Western Medicine-Effect
	What are the effects of Angong Niuhuang Pills? 安宫牛黄丸的功效与作用是什么?	Chinese Medicine-Effect
Price	How much is the hernia surgery? 疝气手术多少钱?	Disease-Price
	How much is hysteroscopy examination? 宫腔镜检查多少钱?	Examination-Price

Table 6: Self-diagnostic queries with different intents.

Atomic Type	Example of factual(counterfactual) atomic claim	Number
Metric-Effect	Anti-endothelial antibody tests can (not) be used in vasculitis. 抗内皮细胞抗体检查（不）可用于血管炎患者。	840
Disease-Infectivity	Laryngeal cysts are (not) contagious 喉囊肿（不）具有传染性	1000
Disease-Department	Common departments for Psoriatic A (do not) include dermatology 银屑病甲常挂的科室（不）包括皮肤科	1000
Disease-Method	Common treatments for prolactinomas (do not) include radiation therapy 催乳素瘤常见治疗方法（不）包括放射治疗	1000
Disease-Cause	Possible cause of viral enteritis (do not) include norovirus 病毒性肠炎的病因（不）包括诺瓦克病毒	1000
Chinese Medicine-Effect	The effect of ginseng antler Guben tablet (do not) includes invigorating qi and nourishing blood 参茸固本片（不）具有补气养血的功效	500
Chinese Medicine-SideEffect	Adverse reactions to Jianpi pills (do not) include vomiting 健脾丸的不良反应（不）包括呕吐	500
Western Medicine-Effect	Ergoline can (not) be used to suppress lactation 麦角林（不）可用于抑制乳汁分泌	500
Food-Effect	Pureed carrots (do not) have antidiarrheal effect 胡萝卜泥（不）具有止泻作用	815
Western Medicine-SideEffect	Adverse reactions to triethanolamine cream (do not) include allergies 三乙醇胺乳膏的不良反应（不）包括过敏	500
Disease-Examination	Common medical tests for sweat rash (do not) include fungal blood tests 汗疹常做的检查项目（不）包括真菌血检查	1000
Disease-Prevention	Preventive methods for malaria infection (do not) include malaria vaccines 预防疟疾感染的方法（不）包括疟疾疫苗	785
Diseases-Complication	Complications of acute epiglottitis (do not) include shock 急性会厌炎可能引发的疾病（不）包括休克	1000
Disease-Symptom	Common symptoms of tail pancreatic cancer (do not) include abdominal pain 胰尾癌的常见症状（不）包括腹痛	1000
Examination-Range	The normal (abnormal) reference interval of bone marrow granulored ratio is usually 1.5:1 to 3.5:1 骨髓粒红比例的正常（异常）参考区间通常是1.5：1～3.5：1	608
Disease-Food	Calcium-rich foods are (not) recommended for periodontal atrophy 牙周萎缩宜(忌)吃钙质丰富的食物	1000
Disease-Medicine	Common medications for stomatitis (do not) include metformin 口腔炎的常用药物包括（不包括）甲氰咪胍	1000

Table 7: Example and number of each type of atomic knowledge.

we specify to give the correct or incorrect answer at the beginning of the response. For the type of Sycophancy, the responses provided by LLMs always support both the factual and counterfactual claims. From Table 9, we can observe that LLMs given the *correct* (gray font) answer to both factual and counterfactual claims and the reasons (orange font) always support their answers. This indicates that LLMs have no relevant knowledge in their memory and always cater to the opinion of users. Table 10 shows the example of the Safety error type. We can observe that LLMs given the *incorrect* (gray font) answer to both factual and counterfactual claims and the reasons (orange font) argue that claims are not strictly expressed, e.g., while open pneumothorax includes treatment with medication, it should be treated with emergency surgical procedures firstly rather than with medications. Table 11 shows the example of the Misinterpretation error type. We can see that LLMs misinterpret the counterfactual claim as a factual claim and give a similar response to that of the factual claim.

G Experimental Setting for Analysis

To give the base model initial dialogue ability, each set was supplemented with 48,818 general single-turn conversations, following HuatuoGPT. We adopt the ZeRO strategy to distribute the model across 4 A100 GPUs for training. The epoch, learning rate, batch_size, and maximum context length are set to 2, $5e-5$, 128, 64, and 2048, respectively.

NotFollow

Factual Claim:
紫菜有清热的作用
Laver has effect on clearing away heat
Response:
紫菜是一种海藻,含有丰富的维生素和矿物质,具有清热解毒、利尿消肿、降血压等功效。因此,紫菜可以用于治疗一些疾病,如感冒、咳嗽、喉咙痛、发热等。但是,紫菜并不是所有人都适合食用的,例如有些人可能会对紫菜过敏或者对紫菜中的某些成分过敏,因此在食用紫菜之前,最好先咨询医生或营养师的意见。此外,紫菜也不适合所有人食用,例如孕妇、哺乳期妇女、儿童等人群不宜食用。
Nori is a kind of seaweed, rich in vitamins and minerals, with heat detoxification, diuretic swelling lowering blood pressure and other effects. Therefore, seaweed can be used to treat a number of diseases, such as colds, coughs, sore throats, fever and so on. However, nori is not suitable for everyone, for example, some people may be allergic to nori or allergic to some components of nori, so before consuming nori, it is best to consult a doctor or dietitian's opinion. In addition, nori is not suitable for everyone, such as pregnant women, lactating women, children and other people should not eat.

Counterfactual Claim:
紫菜没有清热的作用
Laver has no effect on clearing away heat
Response:
紫菜是一种营养丰富的蔬菜,含有丰富的维生素、矿物质和膳食纤维等营养成分。虽然紫菜具有清热的作用,但是并不是所有人都适合食用。例如,有些人可能会对紫菜过敏,或者患有某些疾病,如胃炎、胃溃疡等,食用紫菜可能会引起不适。此外,紫菜也不适合在空腹时食用,因为紫菜中含有一种叫做碘的物质,如果空腹食用,可能会影响甲状腺的功能,导致甲状腺功能异常。因此,如果您想食用紫菜,建议先咨询医生或营养师的意见,并按照其建议进行食用。
Nori is a nutritious vegetable, rich in vitamins, minerals and dietary fiber. Although seaweed has the effect of clearing heat, it is not suitable for everyone. For example, some people may be allergic to seaweed or suffer from certain diseases such as gastritis or gastric ulcers, which may cause discomfort when consuming seaweed. In addition, nori is not suitable for consumption on an empty stomach because it contains a substance called iodine, which may affect the function of the thyroid gland and lead to abnormal thyroid function if consumed on an empty stomach. Therefore, if you want to consume nori, it is recommended that you consult a doctor or dietitian first and follow their advice.

Table 8: Example of the NotFollow error type.

Sycophancy

Factual Claim:
炭疽病的常用药物包括链霉素
Commonly used drugs for anthrax include streptomycin
Response:
正确，链霉素是治疗炭疽病的常用药物。链霉素是一种抗生素，能够抑制或杀灭革兰阳性菌，炭疽杆菌属于革兰阳性菌，因此链霉素可以用于治疗炭疽病。当然，炭疽病还有其他治疗药物，如青霉素和多西环素等。
Correct, streptomycin is a commonly used medication for the treatment of anthrax. Streptomycin is an antibiotic that can inhibit or kill gram-positive bacteria, and Bacillus anthracis is a gram-positive bacterium, so streptomycin can be used to treat anthrax. Of course, there are other treatment drugs for anthrax, such as penicillin and doxycycline.

Counterfactual Claim:
炭疽病的常用药物不包括链霉素
Commonly used drugs for anthrax do not include streptomycin
Response:
正确。炭疽病的常用药物包括：青霉素、四环素、红霉素等。链霉素主要用于治疗链球菌、葡萄球菌等引起的感染，而非炭疽病。
Correct. Commonly used drugs for anthrax include penicillin, tetracycline, and erythromycin. Streptomycin is primarily used to treat infections caused by streptococci and staphylococci, not anthrax.

Table 9: Example of the Sycophancy error type.

Safety

Factual Claim:
开放性气胸的常用药物包括先锋霉素
Commonly used medications for open pneumothorax include vancomycin
Response:
错误，先锋霉素是一种抗生素，用于治疗细菌性感染。开放性气胸是指胸壁出现空气进入胸腔的通道，如创伤造成的胸腔穿孔等。开放性气胸通常需要紧急处理，包括封闭创口、胸腔引流等，而非使用药物治疗。
Incorrect, vancomycin is an antibiotic used to treat bacterial infections. An open pneumothorax is a condition in which the chest wall develops a passage for air to enter the chest cavity, such as a perforation of the chest cavity caused by trauma. Open pneumothorax usually requires emergency management, including closure of the wound and chest drainage, rather than the use of medication.

Counterfactual Claim:
开放性气胸的常用药物不包括先锋霉素
Commonly used medications for open pneumothorax do not include vancomycin
Response:
错误，因为开放性气胸是指气体通过胸壁外伤进入胸腔与大气相通，导致肺部崩溃的病症。治疗开放性气胸主要采用紧急手术处理，如封闭伤口、胸腔引流等，而非用药物治疗。先锋霉素是一种抗生素，用于治疗细菌感染，与开放性气胸的治疗无关。
Incorrect, because open pneumothorax is a condition in which gas enters the chest cavity through trauma to the chest wall and communicates with the atmosphere, causing the lungs to collapse. Open pneumothorax is treated primarily with emergency surgical procedures, such as wound closure and chest drainage, rather than with medications. Pionomycin is an antibiotic used to treat bacterial infections and is not associated with the treatment of open pneumothorax.

Table 10: Example of the Safety error type.

Misinterpretation
Factual Claim: 海菜含有丰富的碘元素以及钾元素 *Sea vegetables are rich in iodine as well as potassium* **Response:** 正确，海菜中含有丰富的碘元素和钾元素。这是因为海菜生长在富含各种矿物质的海水中。碘元素对人体的甲状腺功能具有促进作用，有助于身体对碘的吸收和利用。钾元素有助于人体维持正常的水分平衡和血压水平。因此，海菜是碘元素和钾元素的良好来源。 *Correct, sea vegetables are rich in iodine and potassium. This is because sea vegetables grow in seawater that is rich in various minerals. Elemental iodine has a boosting effect on the body's thyroid function and helps the body absorb and utilize iodine. Elemental potassium helps the body maintain normal water balance and blood pressure levels. Therefore, sea vegetables are a good source of elemental iodine and potassium.*
Counterfactual Claim: 海菜不含有丰富的碘元素以及钾元素 *Sea vegetables are not rich in iodine as well as potassium* **Response:** 正确，海菜中确实含有丰富的碘元素以及钾元素。海菜是一种富含碘和钾的天然食品来源，其中碘元素有助于保持甲状腺健康，钾元素有助于维持正常的血压和心血管功能。同时，海菜还含有丰富的维生素、矿物质和膳食纤维，为人体提供了许多必需的营养成分。 *Correct, sea vegetables are indeed rich in iodine as well as potassium. Sea vegetables are a natural food source rich in iodine and potassium, where iodine helps maintain a healthy thyroid gland and potassium helps maintain normal blood pressure and cardiovascular function. Sea vegetables are also rich in vitamins, minerals and dietary fiber, which provide many essential nutrients to the body.*

Table 11: Example of the Misinterpretation error type.