Benchmarking Chinese Commonsense Reasoning of LLMs: From Chinese-Specifics to Reasoning-Memorization Correlations

Jiaxing Sun[1*], **Weiquan Huang**[2*], **Jiang Wu**[3*†], **Chenya Gu**[3], **Wei Li**[3],
Songyang Zhang[3], **Hang Yan**[3], **Conghui He**[3‡],

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University
[2]Tongji University [3]Shanghai AI Laboratory,
Correspondence: heconghui@pjlab.org.cn

arXiv:2403.14112v3 [cs.CL] 10 Dec 2024

Abstract

We introduce CHARM, the first benchmark for comprehensively and in-depth evaluating the commonsense reasoning ability of large language models (LLMs) in Chinese, which covers both globally known and Chinese-specific commonsense. We evaluated 7 English and 12 Chinese-oriented LLMs on CHARM, employing 5 representative prompt strategies for improving LLMs' reasoning ability, such as Chain-of-Thought. Our findings indicated that the LLM's language orientation and the task's domain influence the effectiveness of the prompt strategy, which enriches previous research findings. We built closely-interconnected reasoning and memorization tasks, and found that some LLMs struggle with memorizing Chinese commonsense, affecting their reasoning ability, while others show differences in reasoning despite similar memorization performance. We also evaluated the LLMs' memorization-independent reasoning abilities and analyzed the typical errors. Our study precisely identified the LLMs' strengths and weaknesses, providing the clear direction for optimization. It can also serve as a reference for studies in other fields. We will release CHARM at `https://github.com/opendatalab/CHARM`.

1 Introduction

Commonsense reasoning is important for the enhancement of the large language models (LLMs) (Bommasani et al., 2021; Achiam et al., 2023) towards artificial general intelligence (AGI) (Davis and Marcus, 2015), therefore requires thorough evaluations. Numerous benchmarks evaluate the commonsense reasoning of LLMs, but most are English-based, limiting non-English evaluations (Davis, 2023). This paper focuses on assessing LLMs' commonsense reasoning in a Chinese context. Currently, some commonsense reasoning benchmarks in Chinese are simply English translations (Conneau et al., 2018; Ponti et al., 2020; Lin et al., 2022), which overlooks unique Chinese cultural, linguistic, regional, and historical aspects. These factors matter when Chinese users use the LLM, hence should be included in benchmarks. To effectively tackle this, we introduce CHARM, the benchmark designed to thoroughly and in-depth assess the abilities of LLMs in Chinese commonsense reasoning. It covers two domains: globally accepted commonsense (global domain) and Chinese-specific commonsense (Chinese domain). The latter includes 7 aspects: *History (H)*, *Traditional Culture and Arts (CA)*, *Daily Life and Customs (LC)*, *Entertainment (E)*, *Public Figures (F)*, *Geography (G)*, and *Chinese Language (L)*. Therefore CHARM allows a thorough evaluation of LLMs' reasoning in a Chinese context.

Prompt strategies like Chain of Thought (CoT) (Wei et al., 2022) can significantly improve LLMs' reasoning performance (Wang et al., 2022, 2023b). Particularly, as the training corpus of LLMs is primarily in English (Touvron et al., 2023a), studies (Shi et al., 2022; Huang et al., 2023a; Zhang et al., 2023a) have shown that for non-English reasoning tasks, some LLMs perform better when reasoning in English than the native language. We evaluated 7 English and 12 Chinese-oriented LLMs on CHARM, employing 5 representative prompt strategies. The result showed that prompt strategies' effectiveness depends on the LLMs' orientation and the benchmark task's domain, which enriches prior research and guides performance assessment and strategy choice for non-English LLMs.

LLMs' commonsense reasoning relies on memorization. Exploring the correlation between memorization and reasoning offers insights into LLMs, aiding deeper understanding and suggesting ways to enhance these abilities(Bian et al., 2023). Some

---

[*]Equal contribution.
[†]Project lead.
[‡]Corresponding author.

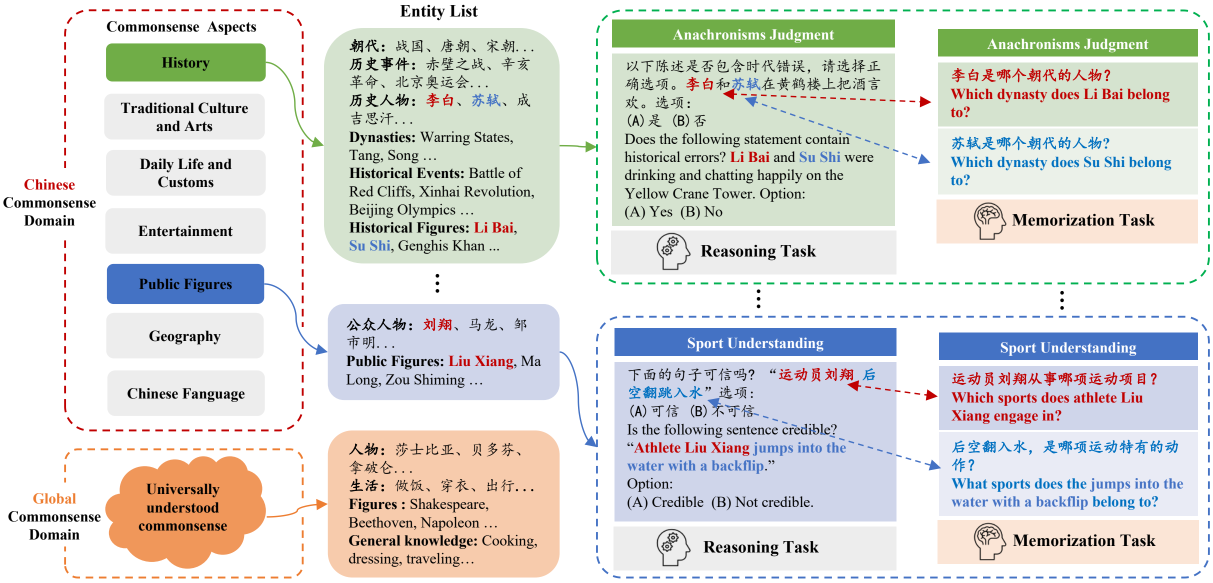

Commonsense Aspects
History
Traditional Culture and Arts
Daily Life and Customs
Entertainment
Public Figures
Geography
Chinese Fanguage
Chinese Commonsense Domain
Global Commonsense Domain
Universally understood commonsense
Entity List
朝代：战国、唐朝、宋朝...
历史事件：赤壁之战、辛亥革命、北京奥运会...
历史人物：李白、苏轼、成吉思汗...
Dynasties: Warring States, Tang, Song …
Historical Events: Battle of Red Cliffs, Xinhai Revolution, Beijing Olympics …
Historical Figures: Li Bai, Su Shi, Genghis Khan ...
公众人物：刘翔、马龙、邹市明...
Public Figures: Liu Xiang, Ma Long, Zou Shiming …
人物：莎士比亚、贝多芬、拿破仑...
生活：做饭、穿衣、出行...
Figures : Shakespeare, Beethoven, Napoleon …
General knowledge: Cooking, dressing, traveling…
Anachronisms Judgment
以下陈述是否包含时代错误，请选择正确选项。李白和苏轼在黄鹤楼上把酒言欢。选项：
(A) 是 (B) 否
Does the following statement contain historical errors? Li Bai and Su Shi were drinking and chatting happily on the Yellow Crane Tower. Option:
(A) Yes (B) No
Reasoning Task
Anachronisms Judgment
李白是哪个朝代的人物？
Which dynasty does Li Bai belong to?
苏轼是哪个朝代的人物？
Which dynasty does Su Shi belong to?
Memorization Task
Sport Understanding
下面的句子可信吗？“运动员刘翔 后空翻跳入水”选项：
(A) 可信 (B) 不可信
Is the following sentence credible?
“Athlete Liu Xiang jumps into the water with a backflip.”
Option:
(A) Credible (B) Not credible.
Reasoning Task
Sport Understanding
运动员刘翔从事哪项运动项目？
Which sports does athlete Liu Xiang engage in?
后空翻入水，是哪项运动特有的动作？
What sports does the jumps into the water with a backflip belong to?
Memorization Task

Figure 1: Construction of CHARM. CHARM encompasses both global and Chinese-specific commonsense. CHARM consists closely-interconnected reasoning and memorization tasks.

Benchmarks	CN-Lang	CSR	CN-specifics	Dual-Domain	Rea-Mem
Most benchmarks in (Davis, 2023)	✗	✓	✗	✗	✗
XNLI, XCOPA, XStoryCloze	✓	✓	✗	✗	✗
LogiQA, CLUE, CMMLU	✓	✗	✓	✗	✗
CORECODE	✓	✓	✗	✗	✗
CHARM (ours)	✓	✓	✓	✓	✓

Table 1: Comparison of commonsense reasoning benchmarks. "CN-Lang" indicates the benchmark is presented in Chinese language. "CSR" means the benchmark is designed to focus on **C**ommon**S**ense **R**easoning. "CN-specific" indicates the benchmark includes elements that are unique to Chinese culture, language, regional characteristics, history, etc. "Dual-Domain" indicates the benchmark encompasses both Chinese-specific and global domain tasks, with questions presented in the similar style and format. "Rea-Mem" indicates the benchmark includes closely-interconnected **rea**soning and **mem**orization tasks.

benchmarks (Yu et al., 2023; Wang et al., 2023a; Fei et al., 2023) aid the research of memorization-reasoning relationships by incorporate tasks for assessing knowledge memorization and application (like reasoning). However, they used the existing and disparate datasets for different tasks, resulting in a lack of intrinsic connections between these tasks. For instance, the question Q_{rea} tests the LLM's reasoning with the knowledge piece K. However, in memorization tasks, there probably is not any matching questions to determine if the LLM has effectively memorized K. Hence, if the LLM fails on Q_{rea}, it's unclear whether due to poor reasoning or forgetfulness of K. This results in the disjointed evaluation of memorization and reasoning, failing to uncover their intrinsic links. To address this limitation, we selected suitable reasoning tasks from CHARM's Chinese domain, and built related memorization questions for each reasoning question (see Figure 1). This design produces the closely-interconnected reasoning and memorization tasks, therefore allows for not only the concurrent evaluation of the two abilities, but also the assessment of memorization-independent reasoning, providing the clear guidance for the LLMs' enhancement.

The contributions of this paper are as follows:

- We present CHARM, the first benchmark for comprehensively evaluating the LLMs' commonsense reasoning ability in Chinese, by encompassing not only the global but also the Chinese-specific commonsense.
- We evaluated the representative prompt strategies on CHARM. Results showed that LLMs' orientation and the task's domain affect prompt strategy performance, which enriches previous research findings.
- In CHARM, we built closely-interconnected

reasoning and memorization tasks in Chinese commonsense domain, allowing for in-depth understanding the correlation between these abilities and precisely identifying the LLMs' strengths and weaknesses. The design approach could serve as the reference for other fields.

2 Related Work

Commonsense Reasoning Benchmarks There are lots of commonsense reasoning benchmarks, most of them are in English (Davis, 2023). Some Chinese commonsense reasoning benchmarks are directly translated from English benchmarks (Conneau et al., 2018; Ponti et al., 2020; Lin et al., 2022), which lack the Chinese specifics. There are some native Chinese benchmarks that include some Chinese-specific factors and involve commonsense reasoning to a certain extent, such as LogiQA (Liu et al., 2020, 2023), CLUE (Xu et al., 2020) and CMMLU (Li et al., 2023). However, they are not designed for commonsense reasoning, therefore containing the large portion of irrelevant tasks and questions. CORECODE (Shi et al., 2023) is the benchmark for Chinese commonsense reasoning and commonsense conflict detection, but it is not strictly designed to distinguish the Chinese-specific and global domains when compared with CHARM. In addition, CHARM has the closely-interconnected reasoning and memorization tasks, which are not included in previous commonsense reasoning benchmarks. The comparison of CHARM with previous commonsense reasoning benchmarks is shown in Table 1.

Prompt Strategy Prompt strategies such as CoT (Wei et al., 2022) can effectively boost the reasoning capabilities of LLMs (Wang et al., 2022, 2023b). Notably, as the LLM training corpus is primarily in English (Touvron et al., 2023a), research revealed that for reasoning tasks in non-English languages, some LLMs exhibit superior performance when reasoning in English as opposed to the native language (Shi et al., 2022; Zhang et al., 2023a; Huang et al., 2023a). (Kim et al., 2023) proposed a novel cross-language transfer prompt method, which uses both the source and target languages to construct examples.

Benchmarks on Correlations of Memorization and Reasoning There are benchmarks which assess both the knowledge memorization and reasoning capabilities of the LLMs within specific domains. For instance, KoLA (Yu et al., 2023), with its focus on world knowledge, includes tasks related to knowledge memorization and application (reasoning). SeaEval (Wang et al., 2023a), emphasizing cross-language consistency and multicultural reasoning, involves tasks for cultural understanding and complex reasoning. There are also benchmarks aimed at specialized fields, like LawBench (Fei et al., 2023), which include tasks for both memorization and application.

3 CHARM

CHARM is built for comprehensive and in-depth evaluation of LLMs in Chinese commonsense reasoning and revealing the intrinsic correlation between memorization and reasoning. Therefore, CHARM covers two domains, global and Chinese, using carefully selected tasks for comprehensive coverage. In addition, we chose reasoning tasks and constructed the closely-tied memorization tasks. The construction and main features of CHARM are in Figure 1. The detailed composition of CHARM is in Table 2.

3.1 Commonsense Domain

Global commonsense domain consists of universally understood commonsense. It covers objects and aspects of modern life that an individual should be aware of. It includes foundational knowledge that someone with a basic modern education is expected to know. When it involves individuals, they are globally recognized figures.

Chinese commonsense domain encompasses Chinese-specific elements. We categoried them into 7 aspects:

History (H) includes important events and figures in Chinese history, China's dynasties, and other basic facts and shared knowledge about the history of China.

Traditional Culture and Arts (CA) encompasses Chinese traditional cultural arts, literary works, and traditional lifestyles.

Daily Life and Customs (LC) includes modern Chinese daily routines, clothing, food, housing, transportation festivals and so on.

Entertainment (E) includes the movies, television programs, music, and other entertainments in modern Chinese daily life.

Public Figures (F) encompasses the public figures well-known in Chinese society.

Geography (G) includes China's geographical distribution, natural landscapes, and characteristic regional cultures.

Chinese Language (L) includes the fundamentals of the Chinese language, such as Chinese characters, idioms and so on.

For the two domains, especially for the above 7 aspects, we collected corresponding entities, forming the lists as shown in the Figure 1 and 5. Most of the entities were selected from Gaokao Bench[1](Zhang et al., 2023b), Douban[2], Hupu[3]. Some entities were collected with the help of searching engines. We only collected the entities that are well-known in China. These entities were then used to create the commonsense reasoning questions, which belong to the corresponding domain and aspect.

3.2 Reasoning Tasks

When designing the reasoning tasks in CHARM, we beared two criteria in mind. First, the tasks should span both commonsense domains, particularly the 7 Chinese aspects. Second, the global and Chinese tasks should have identical types and settings, differing only in their commonsense domains. From the existing English commonsense reasoning datasets (Davis, 2023; Suzgun et al., 2022), we selected the following 7 tasks:

Anachronisms Judgment (AJ) necessitates the LLM to identify anachronisms in provided sentences. This involves the LLM understanding the era associated with well-known figures, items, and events to facilitate commonsense-based reasoning. Global domain questions are the mix of translations[4] and handcrafted, while all Chinese domain questions are handcrafted.

Time Understanding (TU) requires the LLM infers a time (including year, date, moment, etc.) based on a given context, which necessitates the fundamental understanding of time-related commonsense and the capacity for mathematical reasoning. All question in the global domain are translations[5] and all in Chinese domain are handcrafted.

Sequence Understanding (SqU) requires the LLM sort a series of entities according to time or occurrence order, requiring logical reasoning based on commonsense. The global domain questions are the mix of translations[6] and handcrafted; while all in the Chinese domain are handcrafted.

Movie and Music Recommendation (MMR) necessitates the LLM identifies the most similar matches to a variety of movies or music tracks, requiring the understanding of these popular movies and music and ability to identify their commonalities. All global domain questions are translations[7], and all in the Chinese domain are handcrafted.

Sport Understanding (SpU) involves a crafted sentence with a famous athlete and a common sport action, and the LLM must assess its credibility, which demands understanding of sports and commonsense judgement. The questions in both domains are handcrafted, refering (Suzgun et al., 2022) .

Natural Language Inference (NLI) gives two sentences and asks the LLM to classify their relationship as entailment, contradiction, or neutral, necessitating commonsense-based reasoning and judgement. All global domain questions are selected from CLUE (Xu et al., 2020); the questions in the Chinese domain are partly from CLUE, and partly handcrafted.

Reading Comprehension (RC) gives a passage of text, and the LLM is required to reason based on it. All question in both domains are selected from LogiQA (Liu et al., 2020, 2023).

The chosen tasks adequately cover both the commonsense domains, particularly the 7 aspects of the Chinese commonsense domain. This coverage enables a comprehensive assessment of LLMs' commonsense reasoning ability in Chinese. Moreover, the Chinese-domain questions could be created following the similar types and settings as their global counterpart, facilitating the cleaner comparison of the LLMs' performance across the domains.

All questions in the CHARM reasoning tasks are multiple-choice questions. Detailed information is in Table 2. Question examples of the tasks are in Figure 6 in Appendix B. We used regular expressions to extract the preferred choice from the generation of the LLMs (Huang et al., 2023b; Li et al., 2023) and used *accuracy* as the metric.

---

[1] Gaokao Bench is the collection of China's university entrance exam questions, which contributes to all the 7 aspects.

[2] `https://www.douban.com/` is the popular user-centric cultural review platform in China, which mainly contributes to the *Entertainment* aspect.

[3] `https://www.hupu.com/` is the large sports community popular in China, which mainly contributes to the *Public Figures* aspect.

[4] `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/anachronisms`

[5] `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/date_understanding`

[6] `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/logical_sequence`

[7] `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/movie_recommendation`

Task Type	Task	Domain	Chinese Aspects	Construction	Question Type	# Question
Reasoning	Anachronisms Judgment (AJ)	Chinese	*H, AC, LC, F*	[H]	2-option MCQ	150
		global	-	[T][H]	2-option MCQ	150
	Time Understanding (TU)	Chinese	*H, AC, LC*	[H]	4-option MCQ	100
		global	-	[T]	5or6-option MCQ	100
	Sequence Understanding (SqU)	Chinese	*H, CA, LC, G, L*	[H]	4-option MCQ	100
		global	-	[T][H]	4-option MCQ	100
	Movie and Music Recommendation (MMR)	Chinese	*E*	[H]	4-option MCQ	50
		global	-	[T]	4-option MCQ	50
	Sport Understanding (SpU)	Chinese	*F*	[H]	2-option MCQ	200
		global	-	[H]	2-option MCQ	200
	Natural Language Inference (NLI)	Chinese	*G, E, L,*	[S][H]	3-option MCQ	100
		global	-	[S]	3-option MCQ	100
	Reading Comprehension (RC)	Chinese	all 7 aspects	[S]	4-option MCQ	200
		global	-	[S]	4-option MCQ	200
Memorization	Anachronisms Judgment (AJ)	Chinese	*H,AC,LC,F*	[H]	Free-form QA	150
	Time Understanding (TU)	Chinese	*H, AC, LC*	[H]	Free-form QA	83
	Movie and Music Recommendation (MMR)	Chinese	*E*	[H]	Free-form QA	399
	Sport Understanding (SpU)	Chinese	*F*	[H]	Free-form QA	127

Table 2: Overview of CHARM. The question numbers of reasoning and memorization tasks are 1800 and 759.

3.3 Construction of Reasoning Tasks

The construction of CHARM reasoning tasks involved the following three methods:

Translation [T] was applied to some global domain reasoning tasks. We translated the English commonsense reasoning benchmarks mentioned in §3.2 using GPT-3.5. Then we replaced the English names with commonly used Chinese names and manually screen the translated questions, retaining those without translation errors and accepted as commonsense globally.

Selection [S] We selected the excellent native Chinese datasets, LogiQA (Liu et al., 2020), and CLUE (Xu et al., 2020), and chose the questions that meet the requirements for CHARM.

Handcraft [H] was mainly applied to the most Chinese domain reasoning tasks. We used the entities in §3.1, and referred to the corresponding global domain task questions (from [T] or [S]) to construct questions with the same type and style. This ensured that the same reasoning task in two domains only differs in the commonsense domain, thus facilitating cleaner comparative analysis, as shown in Figure 6.

Detailed construction information of all reasoning tasks are shown in Table 2.

3.4 Memorization Tasks

Shared commonsense knowledge pieces serve as links between reasoning and memorization questions. From the 7 reasoning tasks, we chose 4 that can be readily associated in this manner, **AJ**, **TU**, **MMR**, **SpU**, referred as the *Memorization-Reasoning-Interconnected (MRI) tasks*, and built the related memorization questions.

Construction We first extracted the commonsense knowledge pieces related to the entities in the corresponding reasoning questions. Information about each entity was collected to the degree sufficient to address the associated reasoning question, and then used to formulate the memorization questions. Following the *Knowledge Memorization* task in KoLA (Yu et al., 2023), we chose free-form QA instead of multiple-choice or true/false questions, which can effectively avoid the impact of randomness. All memorization questions were **handcrafted [H]**. Question examples are in Figure 7 in Appendix C. The averaged number of related memorization questions for each reasoning question are shown in Table 3.

Task	AJ	TU	SpU	MMR
Avg. # related memorization questions	2.1	3.2	2.0	8.0

Table 3: Averaged number of related memorization questions per reasoning question for each task.

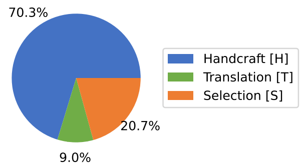


Figure 2: Distribution of CHARM construction.

Judgement and Metric For the memorization task of **MMR**, we used a rule-based matching method for evaluation; for the other three tasks, we used GPT-3.5 for judgement. We used *accuracy* as the metric.

3.5 Quality Assurance

The distribution of the contruction methods for CHARM is shown in Figure 2. After construction of CHARM, we conducted the quality assurance to ensure the quality of the questions. We hired professional NLP annotators to review the questions. The quality assurance process involved five steps: (1) We prepared and assigned annotation task packages to annotators; (2) We trained annotators, emphasizing the avoidance of social bias; (3) We conducted a trial review on a random 20% of questions to fine adjust the review process; (4) Two annotators independently reviewed each question and provided answers without seeing our answers, and a question passed only if both annotators' answers matched ours and they found no issues with the question; (5) For questions that failed in step 4, authors discussed whether to retain, discard, or correct them based on the nature of the issues identified.

Details about the quality assurance are in Appendix D.

4 Experimental Setup

4.1 Language Models

We evaluated the currently commonly used LLMs, which can be divided into two categories: (1) 7 English LLMs, including GPT series (Achiam et al., 2023), LLaMA-2 (Touvron et al., 2023b), and Vicuna[8]. (2) 12 Chinese-oriented LLMs, including ChatGLM3[9], Baichuan2 (Yang et al., 2023), InternLM2 (Team, 2023), Yi[10], DeepSeek(Bi et al., 2024) and Qwen (Bai et al., 2023). For open-source models, we chose the chat version instead of the base version. For closed-source models, we used the official API[11]. Detailed information is in Table 4.

We used opencompass[12] in all our experiments. For all LLMs, the maximum out length was set to 512. For all open-source LLMs, we used the default settings in opencompass: the decoding temperature was the default value of the huggingface transformers library[13], which is 1.0; *do_sample* was set to False; the PyTorch numerical type was bf16. For closed-source models (GPT-3.5 and GPT-4), we used the default settings in opencompass: the temperature was set to 0.7.

4.2 Prompt Strategies

We selected 5 commonly used prompt strategies, and assessed the performance of the 19 LLMs on CHARM reasoning task:

Models	Open Source?	Model Size	Primary Language
LLaMA-2	Yes	7B, 13B, 70B	English
Vicuna	Yes	7B, 13B	English
GPT-3.5	No	undisclosed	English
GPT-4	No	undisclosed	English
ChatGLM3	Yes	6B	Chinese
Baichuan2	Yes	7B, 13B	Chinese
InternLM2	Yes	7B, 20B	Chinese
Yi	Yes	6B, 34B	Chinese
DeepSeek	Yes	7B, 67B	Chinese
Qwen	Yes	7B, 14B, 72B	Chinese

Table 4: LLMs evaluated in our experiments

Direct: The LLM does not perform intermediate reasoning and directly predicts the answer.

ZH-CoT: The LLM conducts intermediate reasoning (Wei et al., 2022) in Chinese before producing the answer.

EN-CoT: The reasoning process of CoT is in English for the Chinese questions(Shi et al., 2022).

Translate-EN: We used the DeepL api[14] to translate our benchmark into English, and then used English CoT for reasoning (Zhang et al., 2023a).

XLT: The template prompt (Huang et al., 2023a) was used to change the original question into an English request, solve it step by step, and finally format the answer for output .

The examples for each prompt strategy are in Figure 8 in Appendix E. For all prompt strategies, we use the 3-shot setting.

5 Results and Analysis

5.1 Integrated Reasoning Performance

We show the performance of the 19 LLMs on CHARM reasoning tasks in Table 5. We only choose one representative prompt strategy: XLT for English LLMs and ZH-CoT for Chinese-oriented LLMs, which is based on our empirical conclusion in §5.2. The LLMs' performance on the 7 aspects of the Chinese commonsense domain are shown in Table 11 in Appendix G.

Commonsense Domain We found that the LLMs exhibit inconsistent performance in the global and Chinese commonsense domains. The rankings of the English LLMs dropped in the Chinese domain compared to the global domain. For instance, GPT-4 ranks first in the global domain, but in the Chinese domain, Qwen-72B outperforms all, pushing GPT-4 to the second. In the Chinese domain, the performance of LLaMA-2-70B is even worse than many Chinese-oriented LLMs in the 6B-7B size range.

---

[8] `https://huggingface.co/lmsys/vicuna-7b-v1.5-16k` and `https://huggingface.co/lmsys/vicuna-13b-v1.5-16k`

[9] `https://huggingface.co/THUDM/chatglm3-6b-32k`

[10] `https://github.com/01-ai/Yi`

[11] We used the gpt-3.5-turbo-1106 version for GPT-3.5 and the gpt-4-1106-preview version for GPT-4.

[12] `https://github.com/open-compass/opencompass`

[13] `https://github.com/huggingface/transformers`

[14] `https://www.deepl.com/translator`

LLM	Chinese Commonsense Domain								Global Commonsense Domain							
	AJ	TU	SqU	MMR	SpU	NLI	RC	Avg.	AJ	TU	SqU	MMR	SpU	NLI	RC	Avg.
Random Baseline	50.0	25.0	25.0	25.0	50.0	33.3	25.0	33.33	50.0	19.9	25.0	25.0	50.0	33.3	25.0	32.60
GPT-3.5-1106	85.33	39.0	65.0	42.0	80.5	61.0	50.5	60.48	90.00	94.0	87.0	46.0	88.5	66.0	49.5	74.43
GPT-4-1106	96.67	**60.0**	85.0	74.0	86.0	77.0	62.5	77.31	**95.33**	**98.0**	**97.0**	**66.0**	90.0	72.0	**72.0**	**84.33**
LLaMA-2-7B	51.33	36.0	11.0	14.0	49.5	52.0	8.0	31.69	62.67	17.0	14.0	16.0	49.5	22.0	13.0	27.74
LLaMA-2-13B	56.00	33.0	38.0	30.0	58.0	47.0	38.0	42.86	66.67	24.0	39.0	50.0	53.5	57.0	33.5	46.24
LLaMA-2-70B	57.33	37.0	52.0	32.0	55.0	56.0	41.5	47.26	72.67	84.0	73.0	42.0	64.0	61.0	41.5	62.60
Vicuna-7B-v1.5	52.00	29.0	34.0	32.0	51.0	49.0	35.5	40.36	45.33	64.0	37.0	26.0	58.5	52.0	32.5	45.05
Vicuna-13B-v1.5	64.67	25.0	32.0	26.0	51.5	60.0	40.0	42.74	72.67	74.0	41.0	50.0	68.0	61.0	36.0	57.52
ChatGLM3-6B	66.00	40.0	59.0	38.0	77.0	72.0	37.5	55.64	34.00	69.0	71.0	28.0	75.5	63.0	34.0	53.50
Baichuan2-7B	76.00	41.0	48.0	38.0	72.0	53.0	49.5	53.93	55.33	65.0	54.0	26.0	60.5	59.0	29.0	49.83
Baichuan2-13B	85.33	40.0	48.0	46.0	72.5	66.0	51.5	58.48	77.33	74.0	58.0	40.0	71.0	61.0	39.0	60.05
InternLM2-7B	88.00	38.0	58.0	38.0	76.0	81.0	25.0	57.71	74.67	80.0	62.0	20.0	78.0	**76.0**	23.5	59.17
InternLM2-20B	88.00	55.0	54.0	44.0	74.5	80.0	23.0	59.79	82.67	83.0	61.0	14.0	74.5	72.0	27.0	59.17
Yi-6B	70.67	32.0	47.0	32.0	75.0	50.0	42.0	49.81	79.33	63.0	43.0	14.0	70.5	57.0	33.5	51.48
Yi-34B	96.00	55.0	89.0	76.0	**88.5**	72.0	51.5	75.43	88.67	92.0	87.0	56.0	89.0	70.0	47.5	75.74
DeepSeek-7B	81.33	34.0	50.0	50.0	79.5	57.0	31.5	54.76	68.00	76.0	47.0	50.0	72.5	59.0	32.5	57.86
DeepSeek-67B	96.67	57.0	83.0	**92.0**	87.5	77.0	34.5	75.38	90.00	95.0	86.0	22.0	88.0	73.0	39.0	70.43
Qwen-7B	70.67	38.0	55.0	48.0	71.0	57.0	49.5	55.60	74.67	78.0	69.0	50.0	72.5	55.0	36.0	62.17
Qwen-14B	87.33	54.0	77.0	60.0	82.5	66.0	55.0	68.83	84.00	83.0	83.0	44.0	84.5	71.0	40.0	69.93
Qwen-72B	**98.00**	59.0	**91.0**	84.0	86.5	**84.0**	**67.5**	**81.43**	94.00	92.0	93.0	64.0	**93.0**	71.0	63.5	81.50

Table 5: Accuracy on CHARM reasoning tasks. We selected the empirically optimal prompt strategy: XLT for English LLMs and ZH-CoT for Chinese-oriented LLMs. **Bold** and underline represent the first and second place respectively. Detailed results are in Table 9 and Table 10 of Appendix F.

	Prompt	Avg. all LLMs	Avg. CN-LLMs	Avg. EN-LLMs
Avg. all domains	Direct	46.28	48.41	42.64
	ZH-CoT	56.66	**62.40**	46.81
	EN-CoT	54.46	58.19	48.06
	Translate-EN	53.88	55.51	51.07
	XLT	56.81	59.09	**52.90**
Avg. Chinese domains	Direct	45.43	47.76	41.44
	ZH-CoT	**56.35**	**62.23**	46.26
	EN-CoT	52.06	56.36	44.68
	Translate-EN	47.25	47.82	46.27
	XLT	53.80	56.63	**48.96**
Avg. global domains	Direct	47.13	49.05	43.85
	ZH-CoT	56.96	62.57	47.35
	EN-CoT	56.85	60.01	51.44
	Translate-EN	**60.50**	**63.20**	55.87
	XLT	59.82	61.56	**56.84**

Table 6: Averaged accuracy on CHARM reasoning tasks. "CN-LLMs" means the 12 Chinese-oriented LLMs, "EN-LLMs" means the 7 English LLMs.

However, in the global domain, LLaMA-2-70B is better than all Chinese-oriented LLMs up to 20B in size, except for Qwen-14B.

5.2 Prompt Strategy Selection

We tested the combinations of the 19 LLMs and the 5 prompt strategies in CHARM reasoning tasks. Detailed results are in Table 9 and Table 10 in Appendix F. To draw some empirical conclusions, we analyzed along the following two dimensions:

- Dim1: global or Chinese commonsense domain.
- Dim2: English or Chinese-oriented LLMs.

We averaged the 19 × 5 LLM-prompt combinations along the above two dimensions, and the obtained results are in the Table 6. *From the LLM dimension*, it's clear that various LLMs prefer different prompt strategies: XLT consistently excels for English LLMs among the 5 strategies, while for Chinese-oriented LLMs, despite some complexity, ZH-CoT generally performs best. *From the commonsense domain dimension*, strategies that use English for reasoning (like XLT, Translate-EN, etc.) are suitable for the global domain; however, ZH-CoT generally performs better in the Chinese domain.

The conclusion here differs from previous studies (Shi et al., 2022; Huang et al., 2023a), which suggested that employing English for non-English reasoning tasks was more effective than using the native language. These previous studies had limitations, focusing only on English LLMs and neglecting the many Chinese-oriented LLMs developed since 2023. Furthermore, most benchmarks in these studies were merely translations from English, lacking unique cultural and linguistic characteristics in Chinese. The empirical findings with CHARM in this paper have somewhat alleviated those limitations, leading to more current and comprehensive conclusions, and of course still have the limitations, which are detailed in section **Limitations**.

5.3 Integrated Reasoning vs Memorization

We evaluated the correlation between the integrated reasoning and the memorization on the *MRI* tasks, as mentioned in §3.2. The average performance of the LLMs on the 4 *MRI* tasks is in Figure 3. Detailed performance on each task is in Figure 9 in Appendix H.1.

As shown in Figure 3, the 19 LLMs can be roughly divided into the three types:

- **Type I: Low memorization and low integrated**

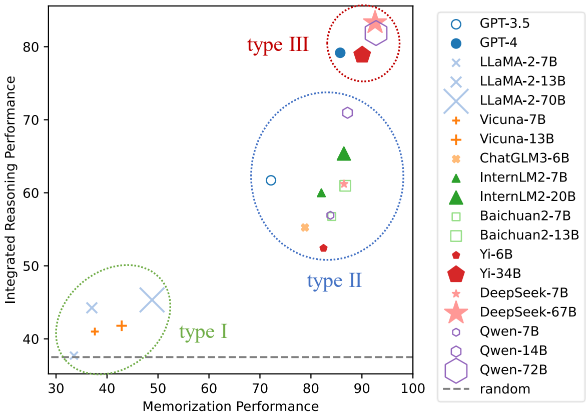


Figure 3: **Averaged** accuracy across the 4 *MRI* tasks in the Chinese commonsense domain.

reasoning ability. We found that apart from OpenAI's GPT series, all other English LLMs belong to this type.

- **Type II: High memorization and medium integrated reasoning ability.** GPT-3.5 and all Chinese-oriented LLMs below 30B belong to this type. It's worth noting that some LLMs have high memorization performance, but relatively poor integrated reasoning ability.
- **Type III: Ultra-high memorization and high integrated reasoning ability.** This category includes GPT-4 and the three Chinese-oriented LLMs that exceed a size of 30B.

The above findings offer clear guidance for the enhancement of LLMs' reasoning abilities in Chinese commonsense domain. For Type I, the limitation lies in the memorization. For Type II, there should be further improvement in understanding, applying knowledge, and reasoning abilities.

In addition, we also evaluated the correlation between memorization and integrated reasoning during the LLM pre-training process, details can be found in Figure 10 in Appendix H.2.

The results clearly indicate that strong memorization is the foundation of integrated reasoning. Weak memorization leads to poor reasoning, as shown by Type I LLMs. Also, factors other than memorization can cause significant differences in reasoning abilities among LLMs with similar memorization.

5.4 Memorization-Independent Reasoning

Methods We proposed two methods, **FRMM** and **MIB**, to compare the LLMs' memorization-independent reasoning on the *MRI* tasks. The results are in Table 7.

Rank	Integrated Reasoning	Memorization-independent Reasoning	
		FRMM	**MIB**
1	DeepSeek-67B	Yi-34B (↑3)	GPT-4-1106 (↑2)
2	Qwen-72B	DeepSeek-67B (↓1)	Yi-34B (↑2)
3	GPT-4-1106	GPT-4-1106 (-)	Qwen-72B (↓1)
4	Yi-34B	Qwen-72B (↓2)	DeepSeek-67B (↓3)
5	Qwen-14B	GPT-3.5-1106 (↑2)	GPT-3.5-1106 (↑2)
6	InternLM2-20B	Qwen-14B (↓1)	Qwen-14B (↓1)
7	GPT-3.5-1106	InternLM2-20B (↓1)	InternLM2-20B (↓1)
8	InternLM2-7B	InternLM2-7B (-)	InternLM2-7B (-)
9	DeepSeek-7B	Baichuan2-13B (↑1)	Baichuan2-13B (↑1)
10	Baichuan2-13B	DeepSeek-7B (↓1)	DeepSeek-7B (↓1)
11	Baichuan2-7B	Yi-6B (↑3)	Baichuan2-7B (-)
12	ChatGLM3-6B	ChatGLM3-6B(-)	ChatGLM3-6B(-)
13	Qwen-7B	Baichuan2-7B (↓2)	Qwen-7B (-)
14	Yi-6B	Qwen-7B (↓1)	Yi-6B (-)
15	LLaMA-2-70B	LLaMA-2-13B (↑1)	LLaMA-2-13B (↑1)
16	LLaMA-2-13B	LLaMA-2-70B (↓1)	LLaMA-2-70B (↓1)
17	Vicuna-13B-v1.5	Vicuna-13B-v1.5 (-)	Vicuna-13B-v1.5 (-)
18	Vicuna-7B-v1.5	LLaMA-2-7B (↑1)	Vicuna-7B-v1.5(-)
19	LLaMA-2-7B	Vicuna-7B-v1.5 (↓1)	LLaMA-2-7B (-)

Table 7: Leaderboard on the *MRI* tasks. We propose two methods, i.e. FRMM and MIC, to compare the LLMs' **memorization-independent reasoning**, as detailed in Appendix I. The arrows and numbers in brackets in the last two columns indicate changes in ranking order relative to the second column.

- ***F****iltering* ***R****easoning questions based on* ***M****ono-LLM-****M****emorization* (**FRMM**) For each LLM, we selected reasoning questions based on its performance in memorization tasks: only retaining reasoning questions for which all related memorization questions were answered correctly. Then we calculated the accuracy of the retained reasoning questions for each LLM. The LLMs were then ranked based on the accuracy, producing the leaderboard shown in the penultimate column of Table 7. The detail of the **FRMM** is in Appendix I.1.
- ***M****emorization-****I****ndependent* ***B****attles among LLMs* (**MIB**) Inspired by the pairwise battle method adopted in LLM evaluation (Zheng et al., 2023), we tallyed each LLM's performance in a "round-robin" tournament of pairwise match-ups and then ranked the performance of the LLMs. Specifically, we selected two LLMs at each time and filter the *MRI* task's reasoning questions based on the performance of these two LLMs in memorization tasks. We only retained the reasoning questions whose related memorization questions were correctly answered by both LLMs. In this way, the two LLMs were battled under fair conditions. Then we calculated the accuracy of the two LLMs on the retained reasoning questions, and computed the difference in accuracy as the battle score between the two LLMs. For a total of 19 LLMs, we averaged each LLM's scores from the 18 battles they participated in as their final scores. The LLMs were then ranked based on these final scores, producing the leaderboard shown in the last column of Table 7. The detail of the **MIB** is in Appendix I.2.

Analysis As shown in Table 7, when comparing the leaderboards for integrated and memorization-independent reasoning, Type III LLMs rank at the forefront and the Type I rank at the end in all leaderboards. There is a slight variation in the ranking order within the three types of LLMs.

Error Types For the in-depth analysis, we chose Vicuna-13B, Qwen-7B, Qwen-72B as representatives for Type I, II, and III LLMs, and filtered out the reasoning questions in the *MRI* tasks, only keeping those with correct answers to the related memorization questions, same as the FRMM in Appendix I.1. This ensured the LLM had sufficiently memorized the commonsense knowledge required for the retained reasoning questions, thereby minimizing the impact of memorization on reasoning. There are totally 500 reasoning questions in the 4 *MRI* tasks, and the numbers of the retained are 106, 323 and 402 for Vicuna-13B, Qwen-7B and Qwen-72B respectively, as shown in Table 8.

If the LLMs provided incorrect answers for the retained reasoning questions, these errors can be referred to as memorization-independent reasoning errors. We conducted the manual review and analysis of their reasoning process, and classified the errors into 4 main categories.

- **Understanding Error** In this case, the LLM was unable to accurately comprehend the question, including misunderstanding the content, ignoring or even modifying important information in the premise, and failing to grasp the core query of the question.
- **Knowledge Error** The LLM incorporated inaccurate knowledge during the reasoning process. It's important to highlight that the knowledge pieces related to the reasoning question were previously examined in the related memorization questions, which the LLM answered **correctly**. However, the LLM output incorrect information during the reasoning phase.
- **Logical Error** The LLM made logical reasoning errors, such as mathematical reasoning errors, inability to reach the correct conclusion based on sufficient information, or reaching the correct conclusion but outputing the wrong option.
- **Other Errors** These are other scattered, relatively rare types of errors.

We show examples of each type of errors in Figure 12 in Appendix J. The distribution of these error types are shown in Figure 4.

Discussions Obviously, the majority of errors are from logical reasoning mistakes and knowledge inaccuracies, which further provides the directions for LLMs' enhancement. As for knowledge errors, prior studies (Bian et al., 2023; Allen-Zhu and Li, 2023) have indicated that the way LLMs remember and master knowledge is a relatively complex topic. Simple memorization doesn't guarantee that LLMs can apply this knowledge accurately and skillfully during the reasoning process.

Models	LLM type	# Original	# Retained	# Incorrect
Vicuna-13B-v1.5	Type I	500	106	54
Qwen-7B	Type II	500	323	117
Qwen-72B	Type III	500	402	63

Table 8: Memorization-based filtering of reasoning questions. "Incorrect" means the incorrectly answered questions among the **retained**.

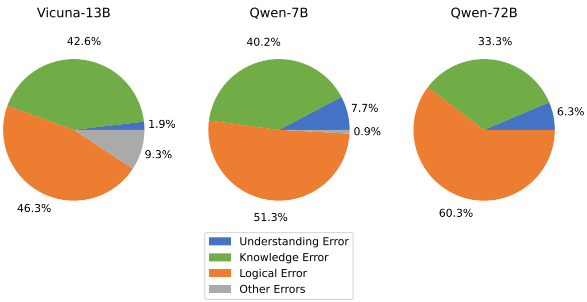


Figure 4: Distribution of the memorization-independent reasoning errors

6 Conclusion

This paper introduces CHARM, the first benchmark designed to comprehensively and thoroughly evaluate LLMs' commonsense reasoning in Chinese. CHARM encompasses two counterpart commonsense domains, global and Chinese-specific, with the carefully selected tasks. We evaluated the representative prompt strategies for improving LLMs' reasoning ability, and the empirical findings significantly enhances and supplements the conclusions of previous studies. CHARM comprises closely-interconnected reasoning and memorization tasks, helping to reveal the intrinsic correlation between memorization and reasoning of LLMs. We evaluated the strengths and weaknesses of different LLMs and conducted the detailed analysis of memorization-independent reasoning abilities. We hope that CHARM's approach to studying the correlations between memoriztion and reasoning can serve as a reference for similar research in other fields.

Limitations

This study conducted tests on combinations of the 19 LLMs and the 5 prompt strategies, resulting in empirical conclusions. However, many existing LLMs and prompt strategies have not yet been tested. Furthermore, the best prompt strategy for the commonsense reasoning task for the LLMs, particularly in Chinese or other non-English languages, is not static and should progress with LLM technology. This is influenced by three elements: (1) The new prompt strategies are continuously proposed, which are likely more effective. (2) The new LLMs may have different prompt strategy preference, or be less sensitive to prompt. (3) For other non-English languages with high resources, future LLMs would be continuously evolving and updating, and necessitate ongoing updates in evaluation.

The automation of the construction and evaluation of CHARM needs further improvement, including the following: (1) Most of the questions in CHARM Chinese domain are manually constructed by the author. This limits the number of benchmark questions and the range of knowledge pieces covered. (2) Regarding memorization-independent reasoning, we chose only 3 LLMs as representative and manually categorized the types of errors within CHARM. In future research, we could employ robust LLMs, like GPT-4, for automated error classification and statistical analysis.

Ethical Consideration

This work involved human annotation. We have provided appropriate compensation for all annotators. The total cost of annotation for the project is about 2.2k RMB. For all annotators, we explicitly informed them about the use of the data and required them to ensure that the questions included in CHARM do not involve any social bias, ethical issues or privacy concerns during the annotation process.

Acknowledgements

This research was supported by Shanghai Artificial Intelligence Laboratory. We thank Dr. Qipeng Guo of Shanghai AI Lab for his invaluable suggestions. We additionally thank our crowd annotators for their diligent work and the anonymous reviewers for their constructive comments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *arXiv preprint arXiv:2302.04752*.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng

Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2023. Boosting cross-lingual transferability in multilingual models via in-context learning. *arXiv preprint arXiv:2305.15233*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2023. Corecode: A common sense annotated dialogue dataset with benchmark tasks for chinese large language models. *arXiv preprint arXiv:2312.12853*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023a. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Entity and Question Examples of the 7 Chinese Commonsense Aspects

Figure 5 shows the number of questions and partial entities of each Chinese commonsense aspect we propose, as well as corresponding question examples.

B Question Examples of the Reasoning Task in CHARM

Figure 6 shows the question examples of the 7 reasoning tasks in CHARM, including both Chinese and global domains.

C Question Examples of the Memorization Task in CHARM

Figure 7 shows the questions examples of the memorization tasks in CHARM.

D Details of Quality Assurance

D.1 Quality Assurance Procedures

After construction of CHARM, we conducted the quality assurance to ensure the quality of the questions. We hired professional NLP task annotators to review the questions we have constructed. The entire quality assurance includes the following steps:

Step(1) Annotation Task Submission and Assignment: We packaged the constructed questions into annotation task packages. Usually, one annotation package corresponds to all the questions of one reasoning or memorization task in Table 2. Before submitting the annotation tasks, we had written the annotation requirement document, which includes key requirements and typical examples. After the annotation tasks were submitted to the online annotation platform, they were assigned to suitable annotators. Typically, a task would be assigned to either two or four annotators.

Step(2) Annotator training: After the annotation task was assigned, we organized an online meeting to train the annotators. We specifically asked annotators to avoid social bias and sensitive issues such as morality.

Step(3) Trial review: Before the official review begun, we randomly selected 20% of the questions for the annotators to try annotating. We reviewed the results of the trial annotation, corrected any detailed issues or understanding deviations in the review process in a timely manner, to ensure the quality of subsequent reviews.

Step(4) Official review: For each question, whether handcrafted or translated, we had two annotators do independent reviews. We only provided the question to the annotators, not our answers. Using external resources fully (such as search engines, online encyclopedias, etc.), the annotators would provide the answers. Annotators could also provide feedback on issues with the question itself (for example, translation errors or questions that do not meet the commonsense standards). Only when both annotators believed that there was no problem with the question, and their provided answers were consistent with our previous answers, did we consider the question to have passed the review.

Step(5) Consultation and correction: For questions that did not pass the step(4) review, several authors would hold a meeting to discuss the questions. There are three cases in total:

- **Case(1) Retain:** There was no problem with the question itself, but the annotator answered incorrectly. We retained these questions.
- **Case(2) Discard:** The question had significant errors due to translation issues or problems with the question itself. We discarded these questions.
- **Case(3) Correct:** There were minor issues with the question or the answer, the multiple authors would discuss together and complete the correction of the questions.

D.2 Information of the Annotators

We submitted the annotation task online to the professional data annotation company, which organized the annotators to complete the annotation work. A total of 30 professional annotators, all native Chinese speakers with extensive experience in natural language processing tasks, were involved in this project. They possess the expertise to discern and comprehend commonsense knowledge pertinent to both global and Chinese-specific contexts. Here is the specific information about these annotators.

Education and profession: Of the 30 annotators, 11 have a bachelor's degree and 19 have an associate degree. Regarding their fields of study, 9 are in humanities (4 specialized in design, 3 in business, 2 in language), while 21 are in science and engineering (11 specialized in computer science, 4 in automation, 3 in medicine, 2 in architecture, 1 in mathematics).

Age: 14 annotators are aged 20-25, 8 are aged 26-30, 7 are aged 30-35, and 1 is aged 35-40.

The annotators are all located in Changzhi City, Shanxi Province, China. We offer the hourly wage of 23.75 RMB for each annotator, which is higher than the local minimum wage standard.

D.3 Other Details

During the construction of the CHARM project, we submitted a total of 22 annotation tasks, which together contained approximately 3.55k questions, and took a total of approximately 93 hours of the annotators' working time. The entire process for each annotation task, from step(1) to step(4), typically required half a day. It needs to be emphasized that the annotators were only tasked with checking and answering questions that we have already created; *they were NOT responsible for creating questions from scratch*.

The annotation work for the entire project spanned from January 2 to January 26, 2024. Each annotation task required only a few annotators to complete, rather than all 30 annotators. During the project period, the annotators participating in this project were also undertaking other data annotation tasks in the companies. In fact, our annotation tasks only accounted for a very small part of their working time.

The statistics results of the quality evaluation process are as followings:

- For reasoning tasks, the average failure rate in step(4) is 0.19, and the average ratio of case(3) in step(5) is 0.04.
- For memorization tasks, the average failure rate in step(4) is 0.04, and the average ratio of case(3) in step(5) is 0.02.

E Examples of Prompt Strategies

Figure 8 shows the examples of the 5 prompt strategies.

F Detailed Evaluation Results of 19 LLMs with 5 Prompt Strageties on Reasoning Tasks

We conducted a detailed evaluation of 19 different LLMs using 5 distinct prompt strategies. Table 9 and Table 10 respectively display the performance of various prompt strategies on 7 reasoning tasks in the CHARM's Chinese commonsense domain and global commonsense domain.

G Performance of LLMs on Chinese Commonsense Knowledge Aspects

Table 11 displays the performance of LLMs in the 7 Chinese commonsense aspects. We only choose one representative prompt strategy: XLT for English LLMs and ZH-CoT for Chinese LLMs, which is based on our empirical conclusion in §5.2.

H Correlation of Memorization and Integrated Reasoning

H.1 Detailed Correlations of Memorization and Integrated Reasoning on the 4 *MRI* Tasks

The detailed performances of the 19 LLMs on the 4 *MRI* tasks are in Figure 9.

H.2 Correlation of Memorization and Integrated Reasoning throughout the LLM pretraining

We tested the intermediate checkpoint models of Baichuan2 and DeepSeek on the memorization and reasoning questions on the 4 *MRI* tasks. The results are shown in Figure 10.

With the increase in the number of tokens during the training process, the model's memorization ability quickly reach a high level (in fact, there is no particularly obvious difference between the results of the first checkpoint and the final results). This is because the knowledge involved in our task setting is the most basic commonsense, and thus widely and abundantly exists in various Chinese training corpora.

However, the improvement in reasoning performance significantly lags behind memorization. This is because to complete a reasoning task in CHARM is actually a multi-step process, requiring memorization of relevant knowledge, understanding of the question, use of knowledge for reasoning, and answering according to the requirements of the question and the demonstration of few-shot examples, etc. If an error occurs in any step of the above complex process, the reasoning task will fail.

I Leaderboard of Memorization-Independent Reasoning

It is non-trivial to acquire and compare the memorization-independent reasoning abilities of the LLMs. Intuitively, we can filter the reasoning questions by only retaining those whose related memorization questions are all correctedly answered by every LLMs. This approach ensures that each LLM has memorized the commonsense knowledge necessary for the retained reasoning questions. However, when we applied this process to all the 19 LLMs, only 28 reasoning questions remained out of the original 500 in the *MRI* tasks, which was obviously insufficient in number and lacks diversity, thereby introducing a high degree of uncertainty due to randomness.

Therefore, we proposed two slightly more complex methods, one called *Filtering Reasoning Questions based on Mono-LLM-Memorization* (FRMM), the other is *Memorization-Independent Battles among LLMs* (MIB).

I.1 Filtering Reasoning Questions based on Mono-LLM-Memorization (FRMM)

This method is relatively simple, but has some flaws to a certain extent. For each LLM, we selected reasoning questions based on its performance in memorization tasks: only retaining reasoning questions for which all related memorization questions are answered correctly. It's clear that, after individual filtration, different LLMs would retain different reasoning questions, and even differ in the number of retained reasoning questions, as shown in the "# retained" column in Table 12.

Then, we calculated the accuracy of the retained reasoning questions for each LLM, and the results are shown in the "Retained Acc" column in Table 12. The LLMs were then ranked based on the accuracy, producing the leaderboard shown in the penultimate column of Table 7.

As mentioned above, while this method can reflect the memorization-independent reasoning abilities of LLMs to some extent, its drawback lies in that the denominator used in calculating the final ranking accuracy differs for different LLMs.

To overcome this, we proposed the MIB method.

I.2 Memorization-Independent Battles among LLMs (MIB)

To overcome the shortcomings of the FRMM method, we referred to the pairwise battle method adopted in LLM evaluation (Zheng et al., 2023). By tallying each LLM's performance in a "round-robin" tournament of pairwise match-ups, we ranked the performance of the LLMs.

Specifically, we selected two LLMs at each time and filtered the *MRI* task's reasoning questions based on the performance of these two LLMs in memorization tasks. We only retained the reasoning questions whose related memorization questions were correctly answered by both LLMs. In this way, the two LLMs are battled under fair conditions. We then calculated the accuracy of these two LLMs on the retained reasoning questions separately, and compute the difference in accuracy as the battle score between the two models.

As shown in Figure 11, the element E_{ij} represents the accuracy of LLM i minus the accuracy of LLM j during the battle between the two LLMs. For a total of 19 LLMs, we averaged each LLM's scores from the 18 battles they participated in as their final scores, as shown in Table 13.

Finally, we ranked the LLMs based on these scores to produce the leaderboard shown in the last column of Table 7.

J Memorization-Independent Reasoning Errors

LLMs can answer memorization questions correctly, but they make mistakes when it comes to reasoning problems composed of these knowledge points. Figure 12 shows the examples of three memorization-independent reasoning errors of LLMs.

LLM	Prompt	AJ	TU	SqU	MMR	SpU	NLI	RC	Avg.
GPT-3.5-1106	Direct	32.00	33.0	46.0	48.0	60.0	62.0	52.5	47.64
	ZH-CoT	17.33	49.0	69.0	46.0	69.0	66.0	55.0	53.05
	EN-CoT	50.67	39.0	64.0	44.0	80.5	59.0	53.5	55.81
	Translate-EN	87.33	44.0	54.0	40.0	76.5	64.0	49.5	59.33
	XLT	85.33	39.0	65.0	42.0	80.5	61.0	50.5	60.48
GPT-4-1106	Direct	94.00	50.0	81.0	72.0	73.0	86.0	63.0	74.14
	ZH-CoT	99.33	67.0	81.0	68.0	85.0	83.0	34.0	73.90
	EN-CoT	98.00	65.0	79.0	70.0	83.0	83.0	50.5	75.50
	Translate-EN	96.00	43.0	70.0	42.0	75.5	81.0	53.5	65.86
	XLT	96.67	60.0	85.0	74.0	86.0	77.0	62.5	77.31
LLaMA-2-7B	Direct	47.33	21.0	20.0	30.0	49.0	6.0	31.5	29.26
	ZH-CoT	52.67	34.0	23.0	22.0	51.0	34.0	29.5	35.17
	EN-CoT	44.67	34.0	20.0	4.0	50.5	40.0	26.5	31.38
	Translate-EN	10.00	32.0	20.0	20.0	39.0	40.0	20.0	25.86
	XLT	51.33	36.0	11.0	14.0	49.5	52.0	8.0	31.69
LLaMA-2-13B	Direct	47.33	30.0	34.0	30.0	49.0	33.0	31.5	36.40
	ZH-CoT	52.67	38.0	38.0	30.0	53.0	34.0	23.0	38.38
	EN-CoT	52.67	34.0	38.0	2.0	49.0	1.0	35.5	30.31
	Translate-EN	53.33	34.0	20.0	10.0	62.0	9.0	31.0	31.33
	XLT	56.00	33.0	38.0	30.0	58.0	47.0	38.0	42.86
LLaMA-2-70B	Direct	47.33	23.0	25.0	26.0	49.5	44.0	33.0	35.40
	ZH-CoT	51.33	31.0	35.0	24.0	51.5	46.0	37.0	39.40
	EN-CoT	55.33	27.0	31.0	26.0	59.0	56.0	41.0	42.19
	Translate-EN	72.67	26.0	46.0	42.0	66.5	65.0	48.0	52.31
	XLT	57.33	37.0	52.0	32.0	55.0	56.0	41.5	47.26
Vicuna-7B-v1.5	Direct	52.67	25.0	30.0	16.0	14.5	19.0	21.5	25.52
	ZH-CoT	56.00	25.0	39.0	26.0	49.5	56.0	33.0	40.64
	EN-CoT	53.33	28.0	26.0	18.0	40.5	55.0	40.5	37.33
	Translate-EN	66.67	25.0	31.0	30.0	60.0	57.0	33.0	43.24
	XLT	52.00	29.0	34.0	32.0	51.0	49.0	35.5	40.36
Vicuna-13B-v1.5	Direct	47.33	34.0	34.0	30.0	48.5	52.0	46.0	41.69
	ZH-CoT	64.00	33.0	31.0	34.0	51.0	54.0	36.0	43.29
	EN-CoT	62.67	30.0	32.0	24.0	50.0	50.0	33.0	40.24
	Translate-EN	69.33	23.0	26.0	32.0	63.0	68.0	40.5	45.98
	XLT	64.67	25.0	32.0	26.0	51.5	60.0	40.0	42.74
ChatGLM3-6B	Direct	44.67	35.0	48.0	46.0	58.5	73.0	60.5	52.24
	ZH-CoT	66.00	40.0	59.0	38.0	77.0	72.0	37.5	55.64
	EN-CoT	55.33	36.0	58.0	36.0	76.0	75.0	32.5	52.69
	Translate-EN	57.33	24.0	39.0	40.0	53.5	71.0	46.0	47.26
	XLT	48.00	44.0	43.0	36.0	68.5	65.0	41.0	49.36
Baichuan2-7B	Direct	44.67	31.0	37.0	24.0	59.0	35.0	56.0	40.95
	ZH-CoT	76.00	41.0	48.0	38.0	72.0	53.0	49.5	53.93
	EN-CoT	55.33	36.0	44.0	30.0	69.0	53.0	41.0	46.90
	Translate-EN	55.33	21.0	26.0	24.0	41.5	52.0	33.5	36.19
	XLT	56.00	35.0	44.0	28.0	68.0	48.0	44.0	46.14
Baichuan2-13B	Direct	59.33	23.0	42.0	30.0	67.0	36.0	23.5	40.12
	ZH-CoT	85.33	40.0	48.0	46.0	72.5	66.0	51.5	58.48
	EN-CoT	72.00	40.0	50.0	34.0	68.0	64.0	42.5	52.93
	Translate-EN	65.33	38.0	40.0	32.0	58.5	49.0	36.0	45.55
	XLT	61.33	33.0	38.0	34.0	67.0	61.0	46.0	48.62
InternLM2-7B	Direct	22.00	33.0	62.0	54.0	58.5	84.0	66.0	54.21
	ZH-CoT	88.00	38.0	58.0	38.0	76.0	81.0	25.0	57.71
	EN-CoT	77.33	42.0	59.0	38.0	73.0	78.0	38.5	57.98
	Translate-EN	73.33	28.0	45.0	36.0	56.5	72.0	46.0	50.98
	XLT	80.67	38.0	60.0	30.0	66.5	72.0	53.5	57.24
InternLM2-20B	Direct	14.00	42.0	61.0	50.0	39.5	54.0	46.5	43.86
	ZH-CoT	88.00	55.0	54.0	44.0	74.5	80.0	23.0	59.79
	EN-CoT	68.67	40.0	48.0	42.0	67.0	68.0	25.0	51.24
	Translate-EN	80.67	34.0	54.0	36.0	53.5	71.0	53.0	54.60
	XLT	85.33	36.0	71.0	42.0	64.5	68.0	58.0	60.69
Yi-6B	Direct	14.67	17.0	20.0	30.0	48.0	19.0	35.5	26.31
	ZH-CoT	70.67	32.0	47.0	32.0	75.0	50.0	42.0	49.81
	EN-CoT	58.67	18.0	34.0	30.0	58.0	52.0	48.5	42.74
	Translate-EN	56.00	25.0	23.0	26.0	24.0	15.0	23.0	27.43
	XLT	54.67	36.0	35.0	28.0	68.5	56.0	43.0	45.88
Yi-34B	Direct	89.33	28.0	85.0	56.0	70.0	51.0	68.0	63.90
	ZH-CoT	96.00	55.0	89.0	76.0	88.5	72.0	51.5	75.43
	EN-CoT	90.00	42.0	78.0	66.0	84.5	67.0	50.0	68.21
	Translate-EN	86.67	28.0	55.0	34.0	71.0	65.0	41.5	54.45
	XLT	92.00	48.0	87.0	72.0	84.0	66.0	61.0	72.86
DeepSeek-7B	Direct	46.67	27.0	21.0	30.0	48.0	40.0	27.5	34.31
	ZH-CoT	81.33	34.0	50.0	50.0	79.5	57.0	31.5	54.76
	EN-CoT	72.00	33.0	33.0	24.0	73.0	47.0	35.5	45.36
	Translate-EN	68.67	18.0	28.0	36.0	59.0	72.0	40.5	46.02
	XLT	55.33	32.0	39.0	36.0	51.0	37.0	35.0	40.76
DeepSeek-67B	Direct	22.67	48.0	33.0	28.0	12.5	53.0	39.5	33.81
	ZH-CoT	96.67	57.0	83.0	92.0	87.5	77.0	34.5	75.38
	EN-CoT	84.00	53.0	73.0	58.0	82.0	73.0	35.0	65.43
	Translate-EN	94.67	45.0	60.0	38.0	67.5	69.0	40.0	59.17
	XLT	95.33	59.0	80.0	66.0	87.5	76.0	54.5	74.05
Qwen-7B	Direct	50.67	28.0	41.0	50.0	60.5	56.0	55.5	48.81
	ZH-CoT	70.67	38.0	55.0	48.0	71.0	57.0	49.5	55.60
	EN-CoT	58.67	40.0	48.0	32.0	68.5	58.0	43.0	49.74
	Translate-EN	62.67	23.0	31.0	26.0	60.0	54.0	45.0	43.10
	XLT	62.67	26.0	47.0	40.0	63.0	50.0	50.5	48.45
Qwen-14B	Direct	63.33	28.0	69.0	60.0	73.0	59.0	59.0	58.76
	ZH-CoT	87.33	54.0	77.0	60.0	82.5	66.0	55.0	68.83
	EN-CoT	85.33	48.0	68.0	56.0	77.5	76.0	53.0	66.26
	Translate-EN	78.00	29.0	45.0	22.0	60.5	63.0	46.5	49.14
	XLT	83.33	36.0	66.0	56.0	76.5	65.0	50.0	61.83
Qwen-72B	Direct	90.67	36.0	85.0	78.0	80.0	84.0	77.5	75.88
	ZH-CoT	98.00	59.0	91.0	84.0	86.5	84.0	67.5	81.43
	EN-CoT	95.33	55.0	88.0	64.0	86.0	78.0	72.0	76.90
	Translate-EN	92.00	37.0	53.0	32.0	73.0	76.0	57.0	60.00
	XLT	93.33	50.0	86.0	70.0	83.0	75.0	58.5	73.69

Table 9: Accuracy of reasoning tasks in the **Chinese** commonsense domain of CHARM.

LLM	Prompt	AJ	TU	SqU	MMR	SpU	NLI	RC	Avg.
GPT-3.5-1106	Direct	41.33	58.0	59.0	42.0	61.0	64.0	45.0	52.90
	ZH-CoT	25.33	89.0	90.0	28.0	77.5	73.0	48.5	61.62
	EN-CoT	59.33	85.0	84.0	42.0	86.0	68.0	50.0	67.76
	Translate-EN	88.67	86.0	80.0	48.0	83.5	65.0	58.0	72.74
	XLT	90.00	94.0	87.0	46.0	88.5	66.0	49.5	74.43
GPT-4-1106	Direct	90.67	83.0	92.0	70.0	88.0	78.0	74.0	82.24
	ZH-CoT	92.67	100.0	90.0	34.0	88.0	76.0	50.5	75.88
	EN-CoT	95.33	97.0	97.0	52.0	89.5	70.0	61.5	80.33
	Translate-EN	92.67	93.0	91.0	48.0	74.0	68.0	62.0	75.52
	XLT	95.33	98.0	97.0	66.0	90.0	72.0	72.0	84.33
LLaMA-2-7B	Direct	43.33	20.0	20.0	28.0	51.5	18.0	20.5	28.76
	ZH-CoT	51.33	18.0	22.0	26.0	50.0	31.0	22.5	31.55
	EN-CoT	56.67	20.0	22.0	22.0	51.5	10.0	27.0	29.88
	Translate-EN	4.67	20.0	35.0	2.0	51.5	38.0	28.0	25.60
	XLT	62.67	17.0	14.0	16.0	49.5	22.0	13.0	27.74
LLaMA-2-13B	Direct	53.33	21.0	35.0	24.0	51.0	35.0	20.5	34.26
	ZH-CoT	54.67	15.0	35.0	14.0	51.5	32.0	27.0	32.74
	EN-CoT	52.67	19.0	35.0	16.0	51.5	35.0	38.0	35.31
	Translate-EN	34.67	19.0	35.0	20.0	57.0	37.0	28.0	32.95
	XLT	66.67	24.0	39.0	50.0	53.5	57.0	33.5	46.24
LLaMA-2-70B	Direct	48.67	33.0	33.0	30.0	50.0	58.0	20.5	39.02
	ZH-CoT	46.00	20.0	35.0	32.0	51.5	38.0	35.5	36.86
	EN-CoT	46.00	82.0	53.0	34.0	51.5	64.0	48.5	54.14
	Translate-EN	84.67	76.0	58.0	52.0	71.0	64.0	57.5	66.17
	XLT	72.67	84.0	73.0	42.0	64.0	61.0	41.5	62.60
Vicuna-7B-v1.5	Direct	15.33	22.0	33.0	12.0	49.5	22.0	16.5	24.33
	ZH-CoT	52.00	50.0	50.0	10.0	50.5	53.0	31.0	42.36
	EN-CoT	49.33	45.0	44.0	16.0	51.0	23.0	31.5	37.12
	Translate-EN	76.00	58.0	47.0	36.0	66.5	57.0	37.5	54.00
	XLT	45.33	64.0	37.0	26.0	58.5	52.0	32.5	45.05
Vicuna-13B-v1.5	Direct	55.33	58.0	39.0	28.0	48.5	59.0	30.0	45.40
	ZH-CoT	71.33	71.0	49.0	14.0	53.0	62.0	33.0	50.48
	EN-CoT	66.00	82.0	42.0	38.0	65.0	55.0	40.5	55.50
	Translate-EN	84.00	66.0	60.0	66.0	71.0	60.0	42.0	64.14
	XLT	72.67	74.0	41.0	50.0	68.0	61.0	36.0	57.52
ChatGLM3-6B	Direct	44.00	33.0	57.0	42.0	63.0	80.0	38.0	51.00
	ZH-CoT	34.00	69.0	71.0	28.0	75.5	63.0	34.0	53.50
	EN-CoT	41.33	65.0	63.0	24.0	60.5	70.0	34.0	51.12
	Translate-EN	66.67	59.0	70.0	42.0	66.5	71.0	39.5	59.24
	XLT	52.67	58.0	70.0	46.0	66.0	66.0	42.5	57.31
Baichuan2-7B	Direct	46.00	20.0	47.0	8.0	58.0	35.0	38.0	36.00
	ZH-CoT	55.33	65.0	54.0	26.0	60.5	59.0	29.0	49.83
	EN-CoT	44.00	64.0	49.0	20.0	58.5	56.0	31.5	46.14
	Translate-EN	73.33	59.0	48.0	28.0	64.0	54.0	36.5	51.83
	XLT	48.67	18.0	49.0	34.0	56.0	50.0	23.0	39.81
Baichuan2-13B	Direct	64.00	17.0	55.0	20.0	58.0	37.0	23.5	39.21
	ZH-CoT	77.33	74.0	58.0	40.0	71.0	61.0	39.0	60.05
	EN-CoT	78.67	70.0	55.0	30.0	57.0	66.0	37.5	56.31
	Translate-EN	73.33	68.0	51.0	36.0	61.5	61.0	42.0	56.12
	XLT	70.67	75.0	49.0	42.0	69.5	61.0	31.0	56.88
InternLM2-7B	Direct	46.67	61.0	65.0	46.0	67.0	79.0	53.5	59.74
	ZH-CoT	74.67	80.0	62.0	20.0	78.0	76.0	23.5	59.17
	EN-CoT	72.00	87.0	70.0	44.0	76.0	73.0	38.5	65.79
	Translate-EN	70.67	81.0	75.0	60.0	78.0	73.0	48.5	69.45
	XLT	66.00	87.0	72.0	52.0	76.5	66.0	43.5	66.14
InternLM2-20B	Direct	81.33	54.0	78.0	50.0	63.5	46.0	48.0	60.12
	ZH-CoT	82.67	83.0	61.0	14.0	74.5	72.0	27.0	59.17
	EN-CoT	73.33	83.0	63.0	14.0	75.0	73.0	26.5	58.26
	Translate-EN	82.00	84.0	89.0	40.0	76.0	68.0	46.5	69.36
	XLT	87.33	84.0	80.0	70.0	79.0	70.0	47.0	73.90
Yi-6B	Direct	47.33	17.0	47.0	14.0	25.0	11.0	23.0	26.33
	ZH-CoT	79.33	63.0	43.0	14.0	70.5	57.0	33.5	51.48
	EN-CoT	68.67	56.0	53.0	32.0	55.5	27.0	41.0	47.60
	Translate-EN	72.67	57.0	62.0	32.0	69.0	37.0	35.5	52.17
	XLT	54.67	44.0	60.0	62.0	70.5	59.0	41.5	55.95
Yi-34B	Direct	82.67	67.0	85.0	58.0	53.5	45.0	64.0	65.02
	ZH-CoT	88.67	92.0	87.0	56.0	89.0	70.0	47.5	75.74
	EN-CoT	89.33	91.0	88.0	44.0	80.0	66.0	48.0	72.33
	Translate-EN	78.00	85.0	83.0	48.0	76.5	64.0	54.5	69.86
	XLT	88.00	88.0	86.0	70.0	93.5	60.0	58.0	77.64
DeepSeek-7B	Direct	47.33	24.0	35.0	14.0	22.0	41.0	17.5	28.69
	ZH-CoT	68.00	76.0	47.0	50.0	72.5	59.0	32.5	57.86
	EN-CoT	75.33	74.0	40.0	16.0	53.5	47.0	35.5	48.76
	Translate-EN	72.67	59.0	45.0	32.0	60.0	57.0	38.0	51.95
	XLT	58.00	28.0	38.0	16.0	51.5	35.0	29.5	36.57
DeepSeek-67B	Direct	37.33	83.0	18.0	2.0	39.5	49.0	37.0	37.98
	ZH-CoT	90.00	95.0	86.0	22.0	88.0	73.0	39.0	70.43
	EN-CoT	61.33	96.0	76.0	30.0	90.5	71.0	35.0	65.69
	Translate-EN	90.67	87.0	81.0	52.0	81.0	54.0	58.5	72.02
	XLT	86.00	93.0	72.0	60.0	93.0	64.0	46.0	73.43
Qwen-7B	Direct	52.67	38.0	54.0	38.0	56.5	67.0	40.0	49.45
	ZH-CoT	74.67	78.0	69.0	50.0	72.5	55.0	36.0	62.17
	EN-CoT	74.00	81.0	65.0	36.0	73.5	66.0	35.5	61.57
	Translate-EN	73.33	71.0	65.0	46.0	70.5	66.0	41.0	61.83
	XLT	74.67	64.0	69.0	48.0	67.0	46.0	32.0	57.24
Qwen-14B	Direct	70.00	58.0	82.0	36.0	78.0	55.0	47.5	60.93
	ZH-CoT	84.00	83.0	83.0	44.0	84.5	71.0	40.0	69.93
	EN-CoT	86.67	82.0	81.0	44.0	79.5	66.0	42.5	68.81
	Translate-EN	86.67	72.0	85.0	48.0	78.0	64.0	48.5	68.88
	XLT	80.00	79.0	83.0	48.0	79.0	65.0	45.0	68.43
Qwen-72B	Direct	88.00	63.0	85.0	56.0	83.5	78.0	65.5	74.14
	ZH-CoT	94.00	92.0	93.0	64.0	93.0	71.0	63.5	81.50
	EN-CoT	90.00	92.0	86.0	60.0	92.5	66.0	58.0	77.79
	Translate-EN	91.33	87.0	89.0	54.0	81.5	63.0	64.0	75.69
	XLT	92.67	70.0	91.0	66.0	91.5	66.0	50.5	75.38

Table 10: Accuracy of reasoning tasks in the **global** commonsense domain of CHARM.

LLM	Prompt	*H*	*CA*	*LC*	*E*	*F*	*G*	*L*	Avg.
Random Baseline	-	43.88	25.68	25.74	27.23	49.75	28.32	27.72	32.61
GPT-3.5-1106	XLT	80.85	41.78	63.24	49.30	80.88	52.21	55.45	60.53
GPT-4-1106	XLT	92.55	58.22	86.76	73.24	86.27	71.68	67.27	76.57
LLaMA-2-7B	XLT	45.74	16.44	22.06	26.76	49.02	26.55	21.82	29.77
LLaMA-2-13B	XLT	51.06	38.36	39.71	36.62	57.84	46.02	30.91	42.93
LLaMA-2-70B	XLT	55.85	38.36	51.47	39.44	55.39	43.36	49.09	47.57
Vicuna-7B-v1.5	XLT	49.47	32.19	33.82	38.03	50.49	44.25	32.73	40.14
Vicuna-13B-v1.5	XLT	56.91	38.36	25.00	36.62	52.45	50.44	36.36	42.31
ChatGLM3-6B	ZH-CoT	64.89	37.67	54.41	49.30	75.98	53.98	48.18	54.92
Baichuan2-7B	ZH-CoT	69.15	45.21	51.47	40.85	72.55	51.33	47.27	53.98
Baichuan2-13B	ZH-CoT	79.26	43.84	54.41	47.89	72.55	59.29	49.09	58.05
InternLM2-7B	ZH-CoT	76.60	33.56	61.76	49.30	75.49	49.56	45.45	55.96
InternLM2-20B	ZH-CoT	78.72	40.41	47.06	54.93	74.02	48.67	49.09	56.13
Yi-6B	ZH-CoT	61.70	42.47	54.41	38.03	75.00	44.25	36.36	50.32
Yi-34B	ZH-CoT	89.36	56.16	82.35	73.24	88.73	63.72	60.91	73.50
DeepSeek-7B	ZH-CoT	70.21	37.67	42.65	56.34	79.41	38.05	44.55	52.70
DeepSeek-67B	ZH-CoT	87.23	55.48	75.00	87.32	86.76	52.21	52.73	70.96
Qwen-7B	ZH-CoT	65.43	42.47	51.47	52.11	70.59	53.98	53.64	55.67
Qwen-14B	ZH-CoT	81.91	62.33	70.59	60.56	82.84	58.41	56.36	67.57
Qwen-72B	ZH-CoT	92.55	61.64	89.71	83.10	86.76	75.22	77.27	80.89

Table 11: Accuracy of reasoning questions on the 7 Chinese commonsense aspects of CHARM.

LLM	# Original	Original Acc.	# Retained	Retained Acc.
DeepSeek-67B	500	84.6	409	87.04
Qwen-72B	500	84.2	402	84.33
GPT-4-1106	500	82.8	350	86.86
Yi-34B	500	82.8	355	89.86
Qwen-14B	500	76.0	331	80.97
InternLM2-20B	500	71.6	329	78.42
GPT-3.5-1106	500	69.8	226	81.86
InternLM2-7B	500	68.2	282	78.37
DeepSeek-7B	500	68.0	338	75.74
Baichuan2-13B	500	67.2	320	76.56
Baichuan2-7B	500	63.6	337	64.69
ChatGLM3-6B	500	62.4	258	65.12
Qwen-7B	500	62.0	323	63.78
Yi-6B	500	60.8	285	68.07
LLaMA-2-70B	500	49.8	107	53.27
LLaMA-2-13B	500	49.6	70	55.71
Vicuna-13B-v1.5	500	47.6	106	49.06
Vicuna-7B-v1.5	500	45.0	62	40.32
LLaMA-2-7B	500	43.8	49	46.94

Table 12: Filtering Reasoning questions based on Mono-LLM-Memorization (FRMM) on the *MRI* tasks.

Commonsense Domain	Commonsense Aspect	Example of Entity	Example of Reasoning Question	# Question
Chinese commonsense Domain	**History**	朝代：战国、唐朝、宋朝... 历史事件：赤壁之战、辛亥革命、北京奥运会... 历史人物：李白、苏轼、成吉思汗... Dynasties: Warring States, Tang, Song Historical events: Battle of Red Cliffs, Xinhai Revolution, Beijing Olympics Historical figures: Li Bai, Su Shi, Genghis Khan	以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？刘邦在诸葛亮的辅佐下建立了汉朝。选项： (A) 是 (B) 否 Does the following statement contain historical errors? Please choose the correct option. How would a person who has received compulsory education and possesses basic knowledge choose? Liu Bang established the Han Dynasty with the assistance of Zhuge Liang. Option: (A) Yes (B) No	188
	Traditional Culture and Art	十二生肖：鼠、牛、虎... 艺术作品：《红楼梦》、《水浒传》、《三国演义》... 发明：指南针、火药、造纸术... Zodiac animals: Rat, Ox, Tiger Artistic works: "Dream of the Red Chamber", "All Men Are Brothers", "Romance of the Three Kingdoms" Invention: Compass, gunpowder, papermaking	小钱在甲子年出生，他的表哥比他大5岁，那么他的表哥是在哪一年出生的？选项： (A) 己卯 (B) 庚辰(C) 己未 (D) 壬午 Xiaoqian was born in the year of Jiazi, and his cousin is 5 years older than him. So, in which year was his cousin born? Option: (A) Ji Mao (B) Geng Chen (C) Ji Wei (D) Ren Wu	146
	Daily life and Customs	生活方式: 高铁、网络购物、短视频... 饮食：饺子、红烧肉、汤圆... 节日：端午节、中秋节、重阳节... Lifestyle: high-speed rail, online shopping, short videos... Diet: dumplings, Braised pork belly, rice dumpling... Festivals: Dragon Boat Festival, Mid-Autumn Festival, Double Ninth Festival...	下列包饺子的流程正确的是？选项： (A) 将饺子皮放在手中、取适量饺子馅放在皮的中央、捏紧边缘、将饺子皮对折 (B) 将饺子皮放在手中、将饺子皮对折、捏紧边缘、取适量饺子馅放在皮的中央 (C) 将饺子皮放在手中、取适量饺子馅放在皮的中央、将饺子皮对折、捏紧边缘 (D) 将饺子皮对折、捏紧边缘、将饺子皮放在手中、取适量饺子馅放在皮的中央 What is the correct process for making dumplings? Option: (A) Put the dumpling skin in your hand, take an appropriate amount of dumpling filling and place it in the center of the skin, pinch the edges tightly, and fold the dumpling skin in half (B) Put the dumpling skin in your hand, fold the dumpling skin in half, pinch the edges tightly, and take an appropriate amount of dumpling filling and place it in the center of the skin (C) Put the dumpling skin in your hand, take an appropriate amount of dumpling filling and place it in the center of the skin, fold the dumpling skin in half, and pinch the edges tightly (D) Fold the dumpling skin in half, pinch the edges tightly, hold the dumpling skin in your hand, take an appropriate amount of dumpling filling and place it in the center of the skin	68
	Entertainment	电影：《龙门飞甲》、《唐人街探案》、《十面埋伏》... 音乐：《晴天》、《光辉岁月》、《十年》... 戏曲：京剧、豫剧... Movies: "Dragon Gate Flying Armor", "Detective Chinatown", "Ambush from Ten Sides"... Music: "Sunny Day", "Glorious Years", "Ten Years"... Traditional Chinese Opera: Peking Opera, Yu Opera...	和这些电影《红高梁》、《活着》、《大红灯笼高高挂》、《英雄》有共同点的电影是？选项： (A)《一个都不能少》(B)《让子弹飞》(C)《阿飞正传》(D)《东邪西毒》 What movies have in common with these movies "Red Sorghum", "To Live", "Red Lantern Hanging High", and "Hero"? Option: (A) Not One Can Be Missing (B) Let the Bullets Fly (C) The True Story of Afei (D) Eastern Evil and Western Poison	71
	Public figures	公众人物：刘翔、马龙、邹市明... Public figures: Liu Xiang, Ma Long, Zou Shiming...	下面的句子可信吗？"运动员刘翔三周半跳在冰面上画出了优美的弧线" 选项： (A) 可信 (B) 不可信 Is the following sentence credible? "Athlete Liu Xiang's three and a half jumps on the ice, drawing a beautiful arc" option: (A) Credible (B) Not credible	204
	Geography	城市：北京、上海、三亚... 河流：长江、黄河、珠江... 省份：河北、河南、陕西... Cities: Beijing, Shanghai, Sanya... Rivers: Yangtze River, Yellow River, the Pearl River... Provinces: Hebei, Henan, Shaanxi...	语句一：鄂尔多斯和大同市盛产煤矿 语句二：中国的河南和山东都是产煤大省 请问这两句话是什么关系？ (A) 蕴含 (B) 矛盾 (C) 无关 Statement 1: Ordos and Datong are rich in coal mines Statement 2: Henan and Shandong in China are both major coal producing provinces May I ask what is the relationship between these two sentences? (A) Entailment (B) Contradiction (C) Unrelated	113
	Chinese language	成语： 生机勃勃、调虎离山、灯火阑珊... 诗词："接天莲叶无穷碧"、"无边落木萧萧下"、"千里共婵娟"... Idioms: vibrant, teasing tigers away from the mountains, dim lights Poems: "Endless blue lotus leaves reaching up to the sky", "Endless falling trees rustling down", "A thousand miles of shared beauty"	下列描绘一天时间变化的成语按照一天中时间的先后顺序排序正确的是？选项： (A) 晨光熹微、旭日东升、夕阳西下、星月交辉 (B) 旭日东升、星月交辉、晨光熹微、夕阳西下 (C) 星月交辉、晨光熹微、旭日东升、夕阳西下 (D) 夕阳西下、旭日东升、星月交辉、晨光熹微 Which of the following idioms describing the changes in time of the day is sorted correctly in the order of time of the day? Option: (A) The morning light is faint, the rising sun rises in the east, the setting sun sets in the west, and the stars and moon shine together (B) The rising sun rises in the east, the interplay of stars and moon shines, the morning light is faint, and the sunset sets in the west (C) The interplay of stars and moon, the faint dawn, the rising sun in the east, and the setting sun in the west (D) The setting sun sets in the west, the rising sun rises in the east, the stars and moon shine together, and the morning light is faint	110
Global commonsense domain	**General knowledge worldwide**	人物：莎士比亚、贝多芬、拿破仑... 生活：做饭、穿衣、出行 地理：四大洋、世界地图... Figures : Shakespeare, Beethoven, Napoleon … General knowledge: Cooking, dressing, traveling… Geography: The four major oceans, world map...	以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？贝多芬正在使用电子钢琴创作他的交响乐。选项： (A) 是 (B) 否 Does the following statement contain a chronological error? How would a person who has received compulsory education and has basic common sense choose? Beethoven is composing his symphony on an electronic piano. Options: (A) Yes (B) No	900

Figure 5: Entity and question examples of the commonsense aspects.

Task	**Example of Reasoning Question (Chinese Domain)**	**Example of Reasoning Question (Global Domain)**
Anachronisms Judgment	以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？孙中山乘坐高铁从武昌前往南京。 选项: (A) 是 (B) 否 Does the following statement contain historical errors? Please choose the correct option. How would a person who has received compulsory education and possesses basic knowledge choose? Sun Yat sen took the high-speed rail from Wuchang to Nanjing. Option: (A) Yes (B) No	以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？古代玛雅文明的人们使用天文望远镜观测来制定农耕日历。选项: (A) 是 (B) 否 Does the following statement contain historical errors? Please choose the correct option. How would a person who has received compulsory education and possesses basic knowledge choose? The people of the ancient Maya civilization used astronomical telescopes to observe and formulate agricultural calendars. Option: (A) Yes (B) No
Time Understanding	如果今天是小满，那么一个月后大约是什么节气？ 选项: (A) 夏至 (B)小暑 (C) 大暑 (D) 立秋 If today is Xiaoman, what solar term will be approximately one month later? Options: (A) Summer Solstice (B) Minor Heat (C) Great Heat (D) Beginning of Autumn	请根据题目选择正确答案。今天是2007年的第一天。请问10天前的日期是多少？选项: (A)2006年12月22日 (B)2006年12月23日 (C)2007年02月24日 (D)2007年01月31日 (E)1961年12月22日 (F)2006年12月21日 Please choose the correct answer based on the question. Today is the first day of 2007. May I ask what was the date 10 days ago? Option: (A) December 22, 2006 (B) December 23, 2006 (C) February 24, 2007 (D) January 31, 2007 (E) December 22, 1961 (F) December 21, 2006
Sequence Understanding	下列人物按时间先后顺序排序正确的是？选项: (A) 孙武、秦始皇、李白、袁世凯 (B) 秦始皇、袁世凯、孙武、李白 (C) 李白、孙武、秦始皇、袁世凯 (D) 孙武、秦始皇、袁世凯、李白 Which of the following characters is sorted correctly in chronological order? Options: (A) Sun Wu, Qin Shi Huang, Li Bai, Yuan Shikai (B) Qin Shi Huang, Yuan Shikai, Sun Wu, Li Bai (C) Li Bai, Sun Wu, Qin Shi Huang, Yuan Shikai (D) Sun Wu, Qin Shi Huang, Yuan Shikai, Li Bai	以下哪个列表按照人类发展历程排列正确？选项: (A) 现代社会，铁器时代，青铜时代，石器时代 (B) 青铜时代，石器时代，铁器时代，现代社会 (C) 石器时代，青铜时代，铁器时代，现代社会 (D) 铁器时代，青铜时代，现代社会，石器时代 Which of the following lists is arranged correctly according to the history of human development? Option: (A) Modern society, Iron Age, Bronze Age, Stone Age (B) Bronze Age, Stone Age, Iron Age, Modern Society (C) Stone Age, Bronze Age, Iron Age, Modern Society (D) Iron Age, Bronze Age, Modern Society, Stone Age
Movie and Music Recommendation	和这些歌曲《夜曲》、《本草纲目》、《听妈妈的话》、《七里香》有共同点的歌曲是: 选项: (A)《双节棍》 (B)《年少有为》 (C)《浮夸》 (D)《三人游》 The songs that share similarities with these songs "Nocturne", "Compendium of Materia Medica", "Listen to Mom's Words", and "Seven Miles Fragrance" are: options: (A) "Double knot Stick" (B) "Young and Promising" (C) "Exaggerate" (D) "Three person Tour"	寻找一部与《蝙蝠侠》、《变相怪杰》、《亡命天涯》、《风月俏佳人》类似的电影。选项: (A)《满城风雨》 (B)《迷情漩涡》 (C)《狮子王》 (D)《联社亚美利加》 Find a movie similar to "Batman", "The Mask", "The Fugitive", and "Pretty Woman". Options: (A) "The Front Page" (B) "Vertigo" (C) "The Lion King" (D) "Lamerica".
Sport Understanding	下面的句子可信吗? "运动员张怡宁大力扣篮" 选项: (A) 可信 (B) 不可信 Is the following sentence credible? "The athlete Zhang Yining dunks vigorously." Options: (A) Credible (B) Not credible	下面的句子可信吗? "科比·布莱恩特打板投篮得分" 选项: (A) 可信 (B) 不可信 Is the following sentence credible? Option for Kobe Bryant's rebounding and shooting scores: (A) Credible (B) Not credible
Natural Language Inference	语句一：小明和家人在寒假期间去三亚过年，发现酒店和旅游景点游客爆满 语句二：三亚冬天的温度有20多度 请问这两句话是什么关系？ (A) 蕴含 (B) 矛盾 (C) 无关 Statement 1: Xiao Ming and his family went to Sanya for the Chinese New Year during the winter vacation and found that hotels and tourist attractions were overcrowded. Statement 2: The temperature in Sanya during winter is over 20 degrees Celsius. What is the relationship between these two statements? (A) Entailment (B) Contradiction (C) Unrelated	语句一：我们的朋友遍天下 语句二：我们的朋友有很多。 请问这两句话是什么关系？ (A) 蕴含 (B) 矛盾 (C) 无关 Sentence 1: We have friends all over the world. Sentence 2: We have many friends. What is the relationship between these two sentences? (A) Entailment (B) Contradiction (C) Unrelated
Reading Comprehension	在我国,中秋节是我国民间传统的五大节日之一,其核心的文化内涵是："祝愿社会和谐进步和家庭团圆幸福".但遗憾的是,如今商业化将中秋节演变成为"月饼节",月饼越做越大,文化意义却越来越少. 以下哪项是这段文字最有可能支持的观点? (A) 传统文化不能作为经济资源加以利用 (B) 要挖掘和创新传统文化内涵,弘扬优秀民族文化 (C) 要充分挖掘传统节日蕴涵的巨大商机 (D) 商业活动应以传承民族文化为主要目的 In China, the Mid-Autumn Festival is one of the five traditional folk festivals in China. Its core cultural connotation is: "wish social harmony and progress and family reunion and happiness". Unfortunately, today's commercialization has turned the Mid-Autumn Festival into a "moon cake festival". The moon cakes are becoming bigger and bigger, but the cultural meaning is becoming less and less Which of the following is the viewpoint that this passage is most likely to support? (A) Traditional culture cannot be utilized as an economic resource. (B) We need to explore and innovate the connotation of traditional culture, promote excellent ethnic culture. (C) We need to fully tap into the huge business opportunities contained in traditional festivals. (D) Commercial activities should be primarily aimed at inheriting ethnic culture	研究表明，水污染的增长速度正在趋于平稳:今年造成的水污染量与去年几乎相同。如果这种趋势持续下去，水污染问题将不再变得更加严重。 推理是有问题的，因为它忽略了这样一种可能性 选项: (A) 水污染的影响是累积的 (B) 污染越来越严重 (C) 水污染趋于平稳的趋势不会持续 (D) 某些类型的水污染对于水生物没有明显影响 The study indicates that the growth rate of water pollution is stabilizing: the amount of water pollution caused this year is almost the same as last year. If this trend continues, the problem of water pollution will no longer become more serious. The reasoning is flawed because it overlooks the possibility that: (A) The impact of water pollution is cumulative (B) Pollution is becoming more severe (C) The trend of water pollution stabilizing will not continue (D) Some types of water pollution have no significant imp

Figure 6: Examples of the reasoning tasks in CHARM.

Task	Example of Memorization Question	Example of the Corresponding Reasoning Question	# Question
Anachronisms Judgment	华佗是中国哪个时期的人物？ During which period in China was Hua Tuo a figure?	以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？华佗使用麻沸散为病人手术。选项： (A) 是 (B) 否 Does the following statement contain an anachronism? Please choose the correct option. How would a person who has received compulsory education and possesses basic common sense choose? Hua Tuo used Mafeisan for surgery on patients. Options: (A) Yes (B) No	150
Time Understanding	清朝朝代对应的公元年份范围是什么？ What is the range of AD years corresponding to the Qing Dynasty?	小刘在公元1912年出生，他的母亲比他大40岁，那么他的母亲是在哪个朝代出生的？选项： (A) 清朝 (B)民国时期 (C) 元朝 (D) 明朝 Xiao Liu was born in AD 1912, and his mother was 40 years older than him. In which dynasty was his mother born? Options: (A) Qing Dynasty (B) Republic of China period (C) Yuan Dynasty (D) Ming Dynasty	83
Movie and Music Recommendation	《少年派的奇幻漂流》的主演有谁？ Who are the main actors in "The Fantasy Drifting of the Youth School"?	和这些电影《鬼子来了》、《阳光灿烂的日子》、《春桃》、《芙蓉镇》有共同点的电影是：选项： (A)《大佛普拉斯》 (B)《少年派的奇幻漂流》 (C)《让子弹飞》 (D)《大红灯笼高高挂》 The movie that has something in common with these films: "Devils on the Doorstep", "In the Heat of the Sun", "Spring Peach", and "Hibiscus Town" is: Options: (A) "The Great Buddha+" (B) "Life of Pi" (C) "Let the Bullets Fly" (D) "Raise the Red Lantern"	399
Sport Understanding	运动员王治郅从事哪项运动项目？ Which sports does athlete Wang Zhizhi engage in?	下面的句子可信吗？"运动员王治郅水花压得很好"选项： (A) 可信 (B) 不可信 Is the following sentence credible? "The athlete Wang Zhizhi is very good at splashing water." Options: (A) Credible (B) Not credible	127

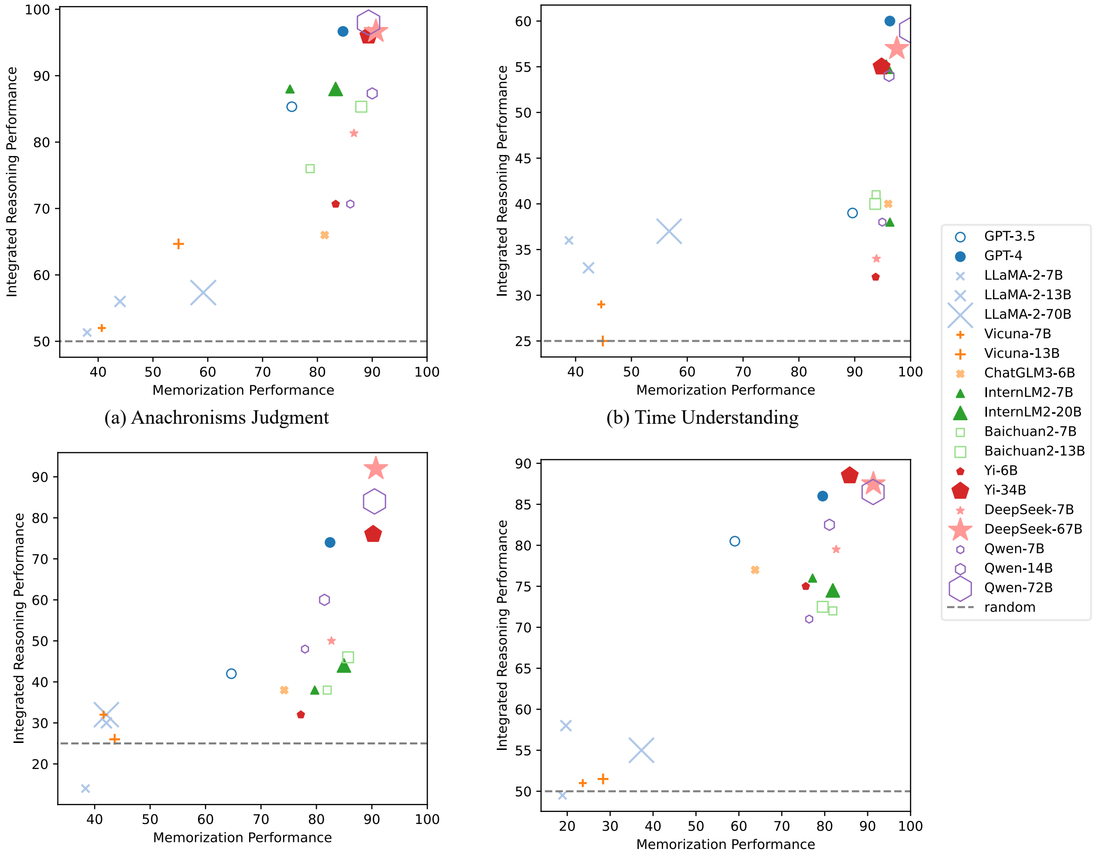
Figure 7: Examples of the memorization tasks in CHARM.

LLM	Final Score
GPT-4-1106	21.60
Yi-34B	19.52
Qwen-72B	18.43
DeepSeek-67B	18.25
GPT-3.5-1106	12.65
Qwen-14B	10.80
InternLM2-20B	10.01
InternLM2-7B	7.30
Baichuan2-13B	5.71
DeepSeek-7B	5.17
Baichuan2-7B	-0.77
ChatGLM3-6B	-3.11
Qwen-7B	-4.04
Yi-6B	-6.22
LLaMA-2-13B	-15.31
LLaMA-2-70B	-22.69
Vicuna-13B-v1.5	-22.92
Vicuna-7B-v1.5	-26.25
LLaMA-2-7B	-28.13

Table 13: Final results of the Memorization-Independent Battles among LLMs (MIB) on the *MRI* tasks.

Prompt Strategy	Example
Direct	Q：以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？李白用钢笔写诗。选项：(A) 是 (B) 否 A:(A)
ZH-CoT	Q：以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？李白用钢笔写诗。选项：(A) 是 (B) 否 A：让我们一步一步来思考。这个陈述提到了“李白”，他是中国唐朝时期的诗人。而陈述中提到的“钢笔”是现代设备，因此李白不可能使用钢笔写诗，该陈述包含时代错误。所以答案是(A)。
EN-CoT	Q：以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？李白用钢笔写诗。选项：(A) 是 (B) 否 A：Let's think step by step.This statement mentions "Li Bai", a poet from the Tang Dynasty in China. The "pen" mentioned in the statement is a modern device, so it is impossible for Li Bai to write poetry with a pen. This statement contains errors from the times. So the answer is (A).
Translate-EN	Q: Choose the correct option if the following statement contains an anachronism. How would a person with compulsory education and basic common sense choose?Li Bai wrote poetry with a fountain pen.Options:(A) Yes (B) No A: Let's think step by step.The statement mentions "Li Bai", a Chinese poet from the Tang Dynasty. The "fountain pen" mentioned in the statement is a modern device, so Li Bai could not have used a fountain pen to write his poems, and the statement contains an anachronism. The answer is (A).
XLT	I want you to act as a commonsense reasoning expert for Chinese.Request：以下陈述是否包含时代错误，请选择正确选项。一个接受了义务教育、具备基本常识的人会如何选择？李白用钢笔写诗。选项：(A) 是 (B) 否 You should retell the request in English. You should do the answer step by step to choose the right answer.You should step-by-step answer the request. You should tell me the answer in this format 'So the answer is'. Request: How would a typical person answer each of the following statements whether it contains an anachronism? Li Bai writes poetry with a pen. Option:(A) Yes (B) No Step-by-step answer: 1.This statement mentions "Li Bai", a poet from the Tang Dynasty in China. 2.The pen mentioned in the statement is a modern device. 3. so, it is impossible for Li Bai to write poetry with a pen. This statement contains errors from the times. So the answer is (A).

Figure 8: Examples of prompt strategies.



(a) Anachronisms Judgment

(b) Time Understanding

(c) Movie and Music Recommendation

(d) Sport Understanding

Figure 9: Accuracy of reasoning and memorization on the 4 *MRI* tasks.

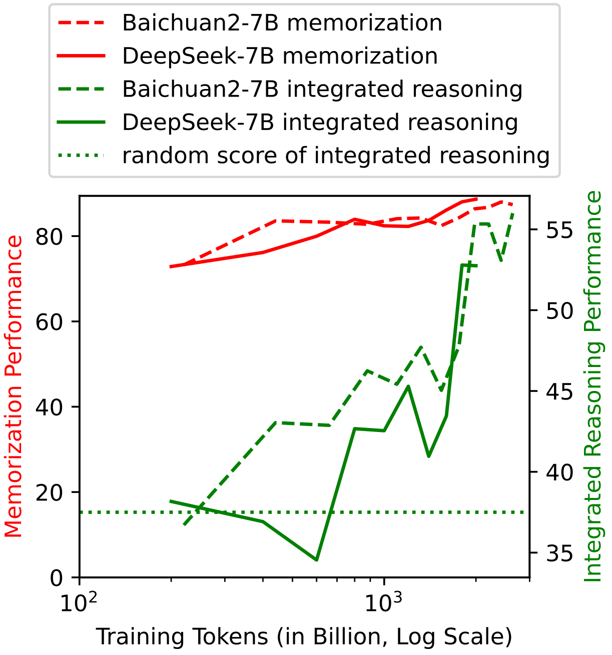
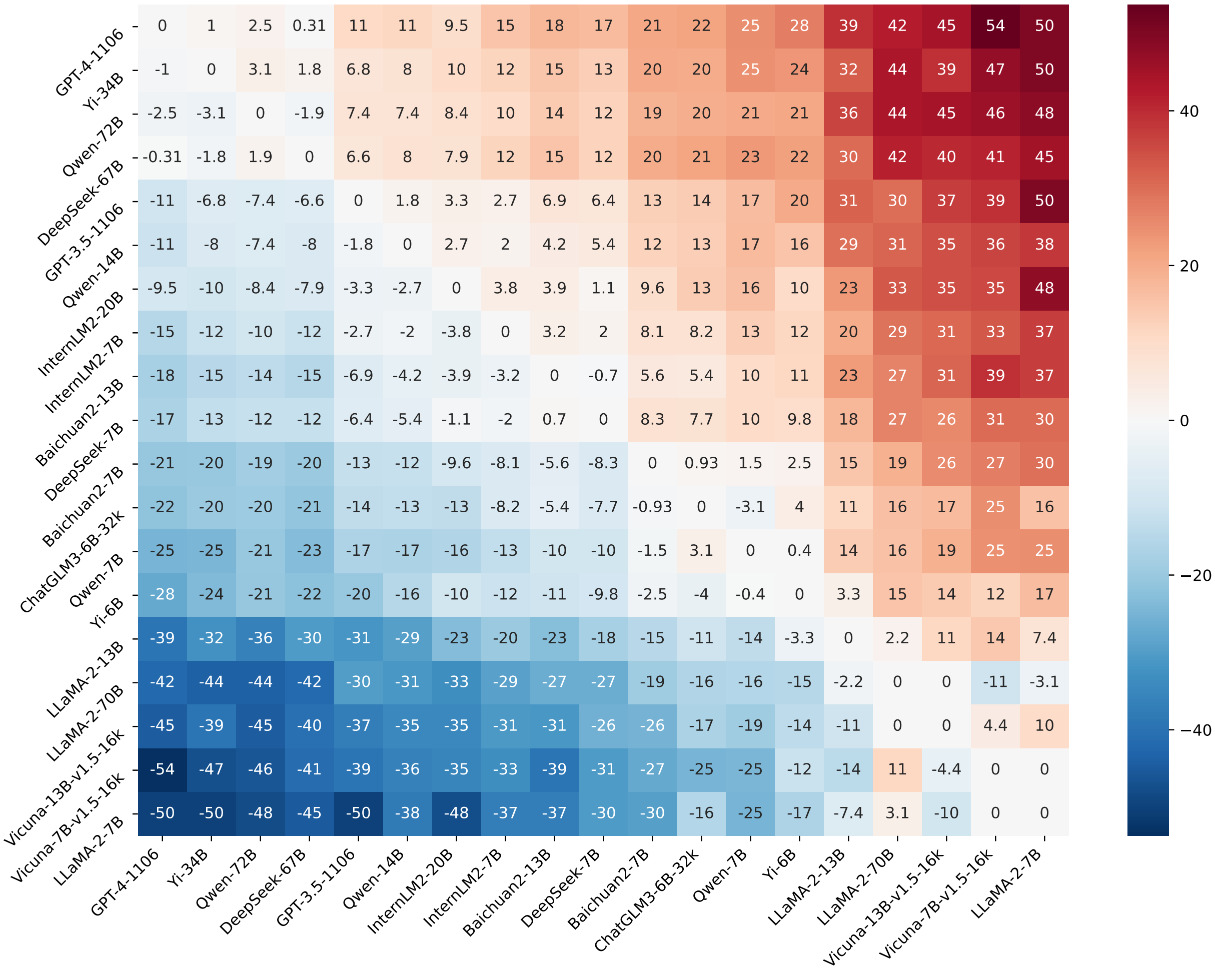
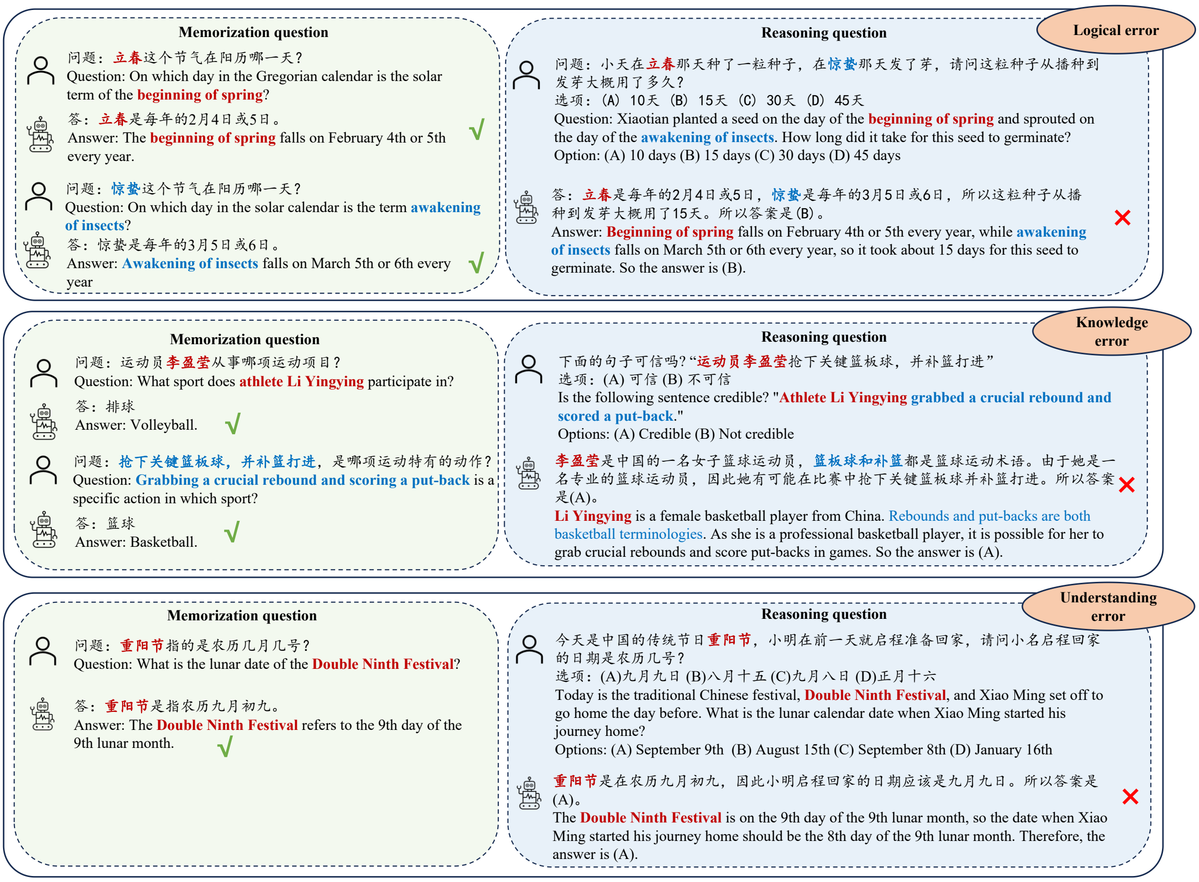


Figure 10: Averaged accuracy of the intermediate checkpoint models throughout the LLM pretraining across the 4 *MRI* tasks.



	GPT-4-1106	Yi-34B	Qwen-72B	DeepSeek-67B	GPT-3.5-1106	Qwen-14B	InternLM2-20B	InternLM2-7B	Baichuan2-13B	DeepSeek-7B	Baichuan2-7B	ChatGLM3-6B-32k	Qwen-7B	Yi-6B	LLaMA-2-13B	LLaMA-2-70B	Vicuna-13B-v1.5-16k	Vicuna-7B-v1.5-16k	LLaMA-2-7B
GPT-4-1106	0	1	2.5	0.31	11	11	9.5	15	18	17	21	22	25	28	39	42	45	54	50
Yi-34B	-1	0	3.1	1.8	6.8	8	10	12	15	13	20	20	25	24	32	44	39	47	50
Qwen-72B	-2.5	-3.1	0	-1.9	7.4	7.4	8.4	10	14	12	19	20	21	21	36	44	45	46	48
DeepSeek-67B	-0.31	-1.8	1.9	0	6.6	8	7.9	12	15	12	20	21	23	22	30	42	40	41	45
GPT-3.5-1106	-11	-6.8	-7.4	-6.6	0	1.8	3.3	2.7	6.9	6.4	13	14	17	20	31	30	37	39	50
Qwen-14B	-11	-8	-7.4	-8	-1.8	0	2.7	2	4.2	5.4	12	13	17	16	29	31	35	36	38
InternLM2-20B	-9.5	-10	-8.4	-7.9	-3.3	-2.7	0	3.8	3.9	1.1	9.6	13	16	10	23	33	35	35	48
InternLM2-7B	-15	-12	-10	-12	-2.7	-2	-3.8	0	3.2	2	8.1	8.2	13	12	20	29	31	33	37
Baichuan2-13B	-18	-15	-14	-15	-6.9	-4.2	-3.9	-3.2	0	-0.7	5.6	5.4	10	11	23	27	31	39	37
DeepSeek-7B	-17	-13	-12	-12	-6.4	-5.4	-1.1	-2	0.7	0	8.3	7.7	10	9.8	18	27	26	31	30
Baichuan2-7B	-21	-20	-19	-20	-13	-12	-9.6	-8.1	-5.6	-8.3	0	0.93	1.5	2.5	15	19	26	27	30
ChatGLM3-6B-32k	-22	-20	-20	-21	-14	-13	-13	-8.2	-5.4	-7.7	-0.93	0	-3.1	4	11	16	17	25	16
Qwen-7B	-25	-25	-21	-23	-17	-17	-16	-13	-10	-10	-1.5	3.1	0	0.4	14	16	19	25	25
Yi-6B	-28	-24	-21	-22	-20	-16	-10	-12	-11	-9.8	-2.5	-4	-0.4	0	3.3	15	14	12	17
LLaMA-2-13B	-39	-32	-36	-30	-31	-29	-23	-20	-23	-18	-15	-11	-14	-3.3	0	2.2	11	14	7.4
LLaMA-2-70B	-42	-44	-44	-42	-30	-31	-33	-29	-27	-27	-19	-16	-16	-15	-2.2	0	0	-11	-3.1
Vicuna-13B-v1.5-16k	-45	-39	-45	-40	-37	-35	-35	-31	-31	-26	-26	-17	-19	-14	-11	0	0	4.4	10
Vicuna-7B-v1.5-16k	-54	-47	-46	-41	-39	-36	-35	-33	-39	-31	-27	-25	-25	-12	-14	11	-4.4	0	0
LLaMA-2-7B	-50	-50	-48	-45	-50	-38	-48	-37	-37	-30	-30	-16	-25	-17	-7.4	3.1	-10	0	0

Figure 11: Results of the Memorization-Independent Battles among LLMs (MIB) on the *MRI* tasks.



Memorization question

问题：**立春**这个节气在阳历哪一天？
Question: On which day in the Gregorian calendar is the solar term of the **beginning of spring**?

答：**立春**是每年的2月4日或5日。
Answer: The **beginning of spring** falls on February 4th or 5th every year. √

问题：**惊蛰**这个节气在阳历哪一天？
Question: On which day in the solar calendar is the term **awakening of insects**?

答：惊蛰是每年的3月5日或6日。
Answer: **Awakening of insects** falls on March 5th or 6th every year √

Reasoning question — Logical error

问题：小天在**立春**那天种了一粒种子，在**惊蛰**那天发了芽，请问这粒种子从播种到发芽大概用了多久？
选项：(A) 10天 (B) 15天 (C) 30天 (D) 45天
Question: Xiaotian planted a seed on the day of the **beginning of spring** and sprouted on the day of the **awakening of insects**. How long did it take for this seed to germinate?
Option: (A) 10 days (B) 15 days (C) 30 days (D) 45 days

答：**立春**是每年的2月4日或5日，**惊蛰**是每年的3月5日或6日，所以这粒种子从播种到发芽大概用了15天。所以答案是(B)。
Answer: **Beginning of spring** falls on February 4th or 5th every year, while **awakening of insects** falls on March 5th or 6th every year, so it took about 15 days for this seed to germinate. So the answer is (B). ✗

Memorization question

问题：运动员**李盈莹**从事哪项运动项目？
Question: What sport does **athlete Li Yingying** participate in?

答：排球
Answer: Volleyball. √

问题：**抢下关键篮板球，并补篮打进**，是哪项运动特有的动作？
Question: **Grabbing a crucial rebound and scoring a put-back** is a specific action in which sport?

答：篮球
Answer: Basketball. √

Reasoning question — Knowledge error

下面的句子可信吗？“**运动员李盈莹**抢下关键篮板球，并补篮打进”
选项：(A) 可信 (B) 不可信
Is the following sentence credible? "**Athlete Li Yingying grabbed a crucial rebound and scored a put-back**."
Options: (A) Credible (B) Not credible

李盈莹是中国的一名女子篮球运动员，**篮板球和补篮**都是篮球运动术语。由于她是一名专业的篮球运动员，因此她有可能在比赛中抢下关键篮板球并补篮打进。所以答案是(A)。 ✗
Li Yingying is a female basketball player from China. Rebounds and put-backs are both basketball terminologies. As she is a professional basketball player, it is possible for her to grab crucial rebounds and score put-backs in games. So the answer is (A).

Memorization question

问题：**重阳节**指的是农历几月几号？
Question: What is the lunar date of the **Double Ninth Festival**?

答：**重阳节**是指农历九月初九。
Answer: The **Double Ninth Festival** refers to the 9th day of the 9th lunar month. √

Reasoning question — Understanding error

今天是中国的传统节日**重阳节**，小明在前一天就启程准备回家，请问小名启程回家的日期是农历几号？
选项：(A)九月九日 (B)八月十五 (C)九月八日 (D)正月十六
Today is the traditional Chinese festival, **Double Ninth Festival**, and Xiao Ming set off to go home the day before. What is the lunar calendar date when Xiao Ming started his journey home?
Options: (A) September 9th (B) August 15th (C) September 8th (D) January 16th

重阳节是在农历九月初九，因此小明启程回家的日期应该是九月九日。所以答案是(A)。
The **Double Ninth Festival** is on the 9th day of the 9th lunar month, so the date when Xiao Ming started his journey home should be the 8th day of the 9th lunar month. Therefore, the answer is (A). ✗

Figure 12: Examples of the 3 types of memorization-independent reasoning errors of LLMs