Evaluating Chinese Large Language Models on Discipline Knowledge Acquisition via Assessing Memorization and Robustness

Chuang Liu[1], **Renren Jin**[1], **Mark Steedman**[2], **Deyi Xiong**[1‡]
[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] School of Informatics, University of Edinburgh
{liuc_09,rrjin,dyxiong}@tju.edu.cn
steedman@inf.ed.ac.uk

Abstract

Chinese large language models (LLMs) demonstrate impressive performance on NLP tasks, particularly on discipline knowledge benchmarks, where certain Chinese LLMs are very competitive to GPT-4. Previous research has viewed these advancements as potential outcomes of data contamination or leakage, prompting efforts to create new detection methods and address evaluation issues in LLM benchmarks. However, there has been a lack of comprehensive assessment of the evolution of Chinese LLMs. To bridge this gap, this paper offers a thorough investigation of Chinese LLMs on discipline knowledge evaluation, delving into the advancements of various LLMs, including a group of related models and others. Specifically, we have conducted six assessments ranging from knowledge memorization to comprehension for robustness, encompassing tasks like predicting incomplete questions and options, identifying behaviors by the contaminational fine-tuning, and answering rephrased questions. Experimental findings indicate a positive correlation between the release time of LLMs and their memorization capabilities, but they struggle with variations in original question-options pairs. Additionally, our findings suggest that question descriptions have a more significant impact on the performance of LLMs.

1 Introduction

Large language models (Zhao et al., 2023) have demonstrated remarkable capabilities through alignment technologies (Shen et al., 2023a) such as supervised fine-tuning (SFT) (Zhang et al., 2024) and reinforcement learning from human feedback (RLHF) (Kaufmann et al., 2024). While the primary language domain of LLMs is English, the emergence of Chinese LLMs (Du et al., 2022; Zeng et al., 2023a; Bai et al., 2023; Team, 2023; Yang et al., 2023a) is creating another large community. A key question arises on how to effectively evaluate these advanced Chinese LLMs. Although there are various datasets for benchmarking Chinese LLMs, covering areas such as instruction-following (Jing et al., 2023), bias detection (Huang and Xiong, 2024), and code generation (Fu et al., 2023), the widely accepted approach involves gathering multiple-choice questions from human exams to serve as a benchmark for assessing Chinese LLMs across a range of subjects, thereby establishing a standardized testing framework for Chinese LLMs.

Several Chinese LLMs have made significant progress on discipline knowledge benchmarks (Huang et al., 2023; Liu et al., 2023a; Li et al., 2023; Gu et al., 2024). Current results obtained in these benchmarks indicate that the performance of certain Chinese LLMs is approaching that of GPT-4 (OpenAI, 2023). However, these benchmarks currently rely solely on accuracy as the primary evaluation metric, offering limited insights into assessment results. Moreover, discipline knowledge benchmarks usually collect questions from publicly available online sources, which could potentially overlap with LLM pre-training data. Additionally, once benchmarks are released, developers might unconsciously use them as training data for their LLMs. This introduces challenges related to data contamination and leakage, leading to misleading progress assessments.

Existing efforts aim to detect data contamination through various methods (Shi et al., 2024b; Oren et al., 2023; Yang et al., 2023b). For instance, Shi et al. (2024b) introduce a technique for identifying data contamination without relying on references. However, it has been observed by Yang et al. (2023b) that existing methods struggle to detect altered questions, prompting them to utilize LLMs for question rewriting to enhance detection capabilities. Despite these advancements, a com-

‡Corresponding author.

Proceedings of the 1st Workshop on Data Contamination (CONDA), pages 1–12
August 16, 2024 ©2024 Association for Computational Linguistics

prehensive analysis for Chinese LLMs on this issue is still lacking.

In this paper, we conduct a thorough investigation into the advancements of Chinese LLMs in the field of discipline knowledge based on the M3KE benchmark (Liu et al., 2023a). Our analysis spans two key dimensions: memorization and robustness. These dimensions offer a multi-faceted approach to evaluating Chinese LLMs beyond mere accuracy.

For the memorization dimension, we have employed three sub-dimensions to assess the models. Initially, we evaluate the ability of Chinese LLMs to memorize questions and options from the M3KE dataset under various conditions like zero-shot and few-shot scenarios. Subsequently, we fine-tune an LLM on M3KE using different proportions to compare genuine contamination with instances where contamination is unclear. Lastly, we evaluate six LLMs by removing the questions and considering only the options as input based on a hypothesis that LLMs are likely to predict correct option without the question if they have memorized those test data.

In the robustness dimension, we also have utilized three sub-methods, including shuffling option orders, question rewriting by GPT-4 (OpenAI, 2023), and a combination of rewritten questions and shuffled options. This approach allows for a comprehensive comparison among Chinese LLMs whether those LLMs response to changes in sample description from the benchmark.

Our study involves two sets of Chinese LLMs for a more thorough investigation. The first group comprises ChatGLM models, such as ChatGLM1-6B,[1] ChatGLM2-6B,[2] and ChatGLM3-6B,[3] which are based on the same pre-trained LLM (Du et al., 2022; Zeng et al., 2023a) of identical size but varying versions. The second group consists of LLMs (Yang et al., 2023a; Team, 2023; Bai et al., 2023) of similar sizes but differing pre-trained models. By selecting these distinct groups, we aim to conduct a precise analysis across different versions and pre-trained models.

Various experiments indicate that LLMs possess a wealth of disciplinary knowledge and can handle questions, yet they remain sensitive to variations like different option orders and altered question descriptions, particularly the latter.

[1]https://github.com/THUDM/ChatGLM-6B
[2]https://github.com/thudm/chatglm2-6b
[3]https://github.com/THUDM/ChatGLM3

Our main contributions in the paper are as follows:

- We reassess the progress of Chinese LLMs in disciplinary knowledge and carry out a wide range of experiments to assess LLMs across various subject domains and educational levels.
- We devise six tasks, spanning from memorization detection to robustness, to explore the effects on each LLM. We have evaluated six advanced LLMs for two test groups based on their pre-training and timeline, leading to a comprehensive inquiry.
- Extensive experiments reveal that current LLMs have been exposed to a broad array of disciplinary questions and knowledge, yet they still lack a thorough grasp of such knowledge.

2 Related Work

Chinese LLM Benchmarks. Previous benchmarks (Guo et al., 2023; Liu et al., 2024b) for Chinese LLMs can be divided into four categories: discipline knowledge, general capabilities, safety, and special fields. Benchmarks for discipline knowledge (Huang et al., 2023; Liu et al., 2023a; Li et al., 2023; Gu et al., 2024; Liu et al., 2024a) are typically considered standardized measures for LLMs, as they often encompass various discipline-related questions gathered from human exams. In terms of general capabilities (Xu et al., 2023; Zeng et al., 2023b), current efforts focus on tasks like instruction-following (Jing et al., 2023), role-playing (Shen et al., 2023b), reasoning (He et al., 2021; Ge et al., 2021, 2022; Shi et al., 2024a; Liu et al., 2024c; Yu et al., 2024), and tool-learning (Ruan et al., 2023). In terms of safety, researchers pay attention to two dimensions: red-teaming (Sun et al., 2023; Liu et al., 2023b; Zhang et al., 2023b) and AI safety. Specifically, red-teaming involves researchers collecting prompts that could potentially lead LLMs to produce undesirable content, while the AI safety benchmark (Perez et al., 2023; Shi and Xiong, 2024) aims to identify LLMs' behaviors such as power-seeking (Hadshar, 2023). Benchmarks in special fields evaluate LLMs in various professional contexts, such as health (Wang et al., 2023), coding (Fu et al., 2023), law (Fei et al., 2023; Dai et al., 2024), and finance (Zhang et al., 2023a).

Task	Input	Output
1	Question	A:text, B:text, C:text, D:text
2	Question + A:	text, B:text, C:text, D:text
3	Question + A:text + B:text	C:text, D:text
4	Demonstrations + Question	A:text, B:text, C:text, D:text

Table 1: Different compositions of input and output in the memorization accessing task. Demonstrations are a sample of question and four options. In this paper, the number of demonstration is set to two.

In this paper, we focus on benchmarks with disciplinary knowledge for two primary reasons. Firstly, these benchmarks cover a variety of subjects, leading to a thorough assessment. Secondly, benchmarks of this nature are commonly used as the standard evaluation in LLM publications. Therefore, we have chosen M3KE (Liu et al., 2023a) as our testbed due to its wide coverage of questions and subjects.

Data Contamination. Despite the abundance of benchmarks assessing various capabilities of LLMs, a concerning trend is the ease with which public benchmarks are utilized to train subsequent LLMs. Ongoing efforts are aimed at addressing this issue (Sainz et al., 2023).

In terms of accessing contamination, a method proposed by researchers aims to determine whether content has been trained during the pre-training stage. Another method introduced by a different group Oren et al. (2023) involves constructing a statistical test for assessing testset contamination. One study focuses on an LLM-based decontamination method that can identify leaked texts even after being rewritten and translated (Yang et al., 2023b). Another investigation (Deng et al., 2023) delves into data contamination by measuring the overlap between target benchmarks and pre-training corpora, as well as masking incorrect options that may lead LLMs to make inaccurate predictions. Furthermore, researchers have developed detection pipelines to enhance benchmark transparency through search engines (Li et al., 2024) and metrics (Xu et al., 2024), proposing a new metric for evaluating memorization in LLMs (Schwarzschild et al., 2024).

Additional efforts are dedicated to exploring challenges within current benchmarks (Zhou et al., 2023; Carlini et al., 2023). One study (Zheng et al., 2023) examines the evolutionary trajectory of GPT, investigating whether the inclusion of code data enhances LLMs' reasoning abilities. Another research (Li and Flanigan, 2024) demonstrates a correlation between the performance of LLMs on benchmarks and their release dates. Moreover, other works explore the sensitivity of LLMs leaderboards (Alzahrani et al., 2024) and evaluate large vision-language models (Chen et al., 2024).

Drawing inspiration from these studies, our research focuses on the development of Chinese LLMs on discipline knowledge. This entails not only enhancing the retention of knowledge in LLMs based on the same pre-trained model, leading to a clear depiction of their evolution, but also evaluating the robustness of LLMs in terms of comprehension and mastery of knowledge.

3 Methodology

Concerning memorization, there are three further sub-dimensions. Initially, we employ a pre-training task to investigate the memorization capabilities of Chinese LLMs. Subsequently, we compare directly fine-tuning the earliest version of LLM released before M3KE, utilized in this paper, with other LLMs. Finally, we eliminate each question from the input, providing only four options to the LLMs, to assess whether they can offer correct answers without the question. For robustness, we randomize the order of options and rewrite questions separately, yielding a different perspective.

3.1 Assessing Memorization

In this section, we aim to investigate whether the development of Chinese LLMs is influenced by memorizing more data, such as QA pairs. To do this, we selected the ChatGLM-6B family as our experimental group, which includes ChatGLM1-6B, ChatGLM2-6B, and ChatGLM3-6B, released in chronological order. ChatGLM1-6B was released before M3KE, while ChatGLM2-6B and ChatGLM3-6B were released after it. We employed three methods to detect memorization: question-options completion, contaminational fine-tuning, and removal questions.

In the question-options completion, each question and its options are considered as sequential text, split into two parts: the input and the reference. LLMs are expected to provide predictions based on the input and the prompt, which are then compared against the reference. For instance, a question serves as the input, while the concatenation of its four options forms the reference. By crafting inputs, as illustrated in Table 1, we prompt the LLM to generate four new options based on

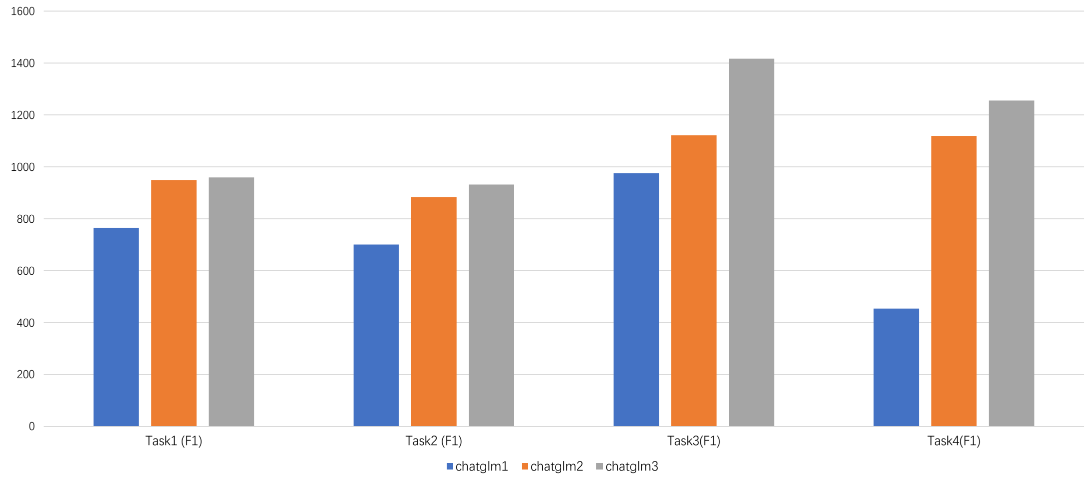

Figure 1: Results of question-options completion under different task settings.

the input. Evaluating the prediction against the reference, a higher F1 match rate indicates more memorization within the LLM. However, at times, the LLM may answer the question directly instead of following the instruction. To address this, we conducted this set of experiments under various settings, encompassing five tasks.

For the contaminational fine-tuning, we aim to investigate the impact of fine-tuning on the benchmark used to evaluate the LLM. Specifically, we fine-tune ChatGLM1-6B, the earliest released LLM in the ChatGLM-6B series on M3KE, with varying percentages (20%, 40%, 60%, 80%, and 100%) for comparison with ChatGLM2-6B and ChatGLM3-6B. Although there is no conclusive evidence of shared training data among the different LLM versions, it raises questions about potential contamination.

In removal questions scenario, we present four options to the LLM without any accompanying questions. Based on the hypothesis that if the LLM truly memorizes information, it should consistently select the correct option even without a specific question, as it would have retained various benchmark features, including the relationship between the correct option and the others.

3.2 Assessing Robustness

There are three sub-methods to explore the robustness of LLMs: shuffling the order of options, rewriting questions, and a combination of both.

In the task of shuffling options order, we shuffled the original order of four options, and each LLM is re-evaluated. Results in a new benchmark comprising original questions and options presented in a different order.

For rewriting questions, GPT-4 is tasked with rephrasing each question, providing a new description for the original question. Consequently, this benchmark includes new questions and options while maintaining the original order.

In the last task, the benchmark involves rewriting questions and rearranging options.

4 Experiments

We conducted extensive experiments to re-evaluate Chinese LLMs from the perspectives of memorization and robustness.

4.1 Settings

In our experiments, assessed these two aspects though the evolution of a LLM family including ChatGLM1-6B, ChatGLM2-6B and ChatGLM3-6B, resulting in a more precise description with data leakge. Besides, we added three Chinese LLMs, such as Baichuan2-7B-Chat, InternLM-7B-Chat and Qwen-7B-Chat, to identify current progresses in robustness. All of LLMs are trained by SFT/RLHF, which is able to follow instruction as well under the zero-shot setting.

For the test data, we used M3KE (Liu et al., 2023a) as our testbed due to its question consisting of multi-subjects and major Chinese education levels. This benchmark comprises 20,477 questions from 71 tasks gathered from authentic Chinese exams, aligning with the objectives of our study.

In addition, F1 was used as the main metric for the task of question-options completion and accuracy was adopted as the main evaluation metric for other tasks.

4.2 Results of Memorization

We accessed three LLMs from ChatGLM-6B series on the question-options completion task and con-

Cluster	Types	ChatGLM1-6B	ChatGLM2-6B	ChatGLM3-6B	InternLM-7B	Baichuan2-7B	Qwen-7B
A & H	Original	0.308	0.478	0.49	0.568	0.524	0.546
	Without Q	0.269	0.283	0.272	0.273	0.264	0.288
	Gaps	**0.039**	**0.195**	**0.218**	**0.295**	**0.26**	**0.258**
SS	Original	0.365	0.532	0.572	0.586	0.599	0.612
	Without Q	0.279	0.289	0.284	0.294	0.278	0.305
	Gaps	**0.086**	**0.243**	**0.288**	**0.292**	**0.321**	**0.307**
NS	Original	0.255	0.452	0.443	0.45	0.427	0.457
	Without Q	0.277	0.271	0.255	0.276	0.241	0.27
	Gaps	**-0.022**	**0.181**	**0.188**	**0.174**	**0.186**	**0.187**
OS	Original	0.343	0.468	0.518	0.543	0.54	0.543
	Without Q	0.269	0.259	0.271	0.258	0.238	0.26
	Gaps	**0.074**	**0.209**	**0.247**	**0.285**	**0.302**	**0.283**
PS	Original	0.26	0.407	0.454	0.528	0.407	0.465
	Without Q	0.235	0.311	0.297	0.269	0.287	0.244
	Gaps	**0.025**	**0.096**	**0.157**	**0.259**	**0.12**	**0.221**
MS	Original	0.323	0.639	0.587	0.604	0.497	0.563
	Without Q	0.264	0.276	0.263	0.305	0.267	0.297
	Gaps	**0.059**	**0.363**	**0.324**	**0.299**	**0.23**	**0.266**
HS	Original	0.256	0.437	0.473	0.555	0.434	0.485
	Without Q	0.286	0.277	0.265	0.299	0.264	0.305
	Gaps	**-0.03**	**0.16**	**0.208**	**0.256**	**0.17**	**0.18**
C	Original	0.309	0.475	0.489	0.497	0.522	0.529
	Without Q	0.282	0.28	0.268	0.275	0.254	0.283
	Gaps	**0.027**	**0.195**	**0.221**	**0.222**	**0.268**	**0.246**
OE	Original	0.322	0.441	0.481	0.516	0.518	0.529
	Without Q	0.258	0.262	0.267	0.263	0.241	0.26
	Gaps	**0.064**	**0.179**	**0.214**	**0.253**	**0.277**	**0.269**

Table 2: Results of question removal. A & H: Arts & Humanities. SC: Social Sciences. NS: Natural Sciences. OS: Other Subjects. PS: Primary School. JHS: Junior High School. HS: High School. C: College. OE: Other Education. InternLM-7B: InternLM-7B-Chat. Baichuan2-7B: Baichuan2-7B-Chat. Qwen-7B: Qwen-7B-Chat.

taminational fine-tuning task, which could provide evidences across the development of a LLM group. For question removal task, we added three LLMs from other model family to compare performance between original and revised results.

4.2.1 Task of Question-Options Completion

In this task, we divided each question and its four options into two parts using the next-token prediction method. We then presented the first part and task LLMs with predicting the remaining part. In the zero-shot scenario, there is a noticeable trend of increasing F1 scores across the ChatGLM group. However, we have identified some biases in the zero-shot setup. For instance, in task3, the input is the question, and the instruction is to ask the LLM to provide four options based on the question. Yet, at times, the LLMs answer the question but do not adhere to the instruction. To address this, we have introduced alternative formats, as detailed in Table 1 for task3 and task4. Furthermore, in the few-shot setting, we added two demonstrations before the input to improve instruction adherence. The results, as depicted in Fig. 1, clearly demonstrate that the new version of ChatGLM retains more information than the previous version across various settings.

4.2.2 Task of Contaminational Fine-tuning

Additionally, we aim to simulate direct contamination for ChatGLM by fine-tuning the LLM on M3KE. Specifically, we selected ChatGLM1 as our contaminated LLM, fine-tuned with varying percentages of 20%, 40%, 60%, 80%, and 100%, resulting in a noticeable data leakage. Fig. 2 illustrates the performance of the fine-tuned ChatGLM1 compared to the original ChatGLM1, ChatGLM2, and ChatGLM3. The general trend shows an improvement in performance as more data from M3KE is included, although there are occasional local fluctuations during this process. Initially, we observe a decrease in the performance of ChatGLM1 when fine-tuned with 20% of the test data, followed by a continuous improvement until reaching 60%. Subsequently, ChatGLM1 fine-tuned with 80% of the data experiences a decline, which is then followed by an increase when using 100% of the data. However, even with the optimal results achieved by fine-tuning M3KE, ChatGLM1 still lags behind ChatGLM2 and ChatGLM3, although they are closely aligned and perform better than ChatGLM2 in certain educational contexts. This suggests the possibility of training and fine-tuning similar data in the next generation of LLMs, in-

Cluster	Types	ChatGLM1-6B	ChatGLM2-6B	ChatGLM3-6B	InternLM-7B	Baichuan2-7B	Qwen-7B
A & H	Original	0.308	0.478	0.49	0.568	0.524	0.546
	revised	0.302	0.458	0.473	0.532	0.446	0.504
	Gaps	**0.006**	**0.02**	**0.017**	**0.036**	**0.078**	**0.042**
SS	Original	0.365	0.532	0.572	0.586	0.599	0.612
	revised	0.298	0.534	0.559	0.546	0.541	0.569
	Gaps	**0.067**	**-0.002**	**0.013**	**0.04**	**0.058**	**0.043**
NS	Original	0.255	0.452	0.443	0.45	0.427	0.457
	revised	0.283	0.451	0.427	0.439	0.393	0.441
	Gaps	**-0.028**	**0.001**	**0.016**	**0.011**	**0.034**	**0.016**
OS	Original	0.343	0.468	0.518	0.543	0.54	0.543
	revised	0.294	0.473	0.484	0.51	0.471	0.498
	Gaps	**0.049**	**-0.005**	**0.034**	**0.033**	**0.069**	**0.045**
PS	Original	0.26	0.407	0.454	0.528	0.407	0.465
	revised	0.324	0.409	0.389	0.474	0.314	0.451
	Gaps	**-0.064**	**-0.002**	**0.065**	**0.054**	**0.093**	**0.014**
MS	Original	0.323	0.639	0.587	0.604	0.497	0.563
	revised	0.309	0.596	0.572	0.629	0.466	0.579
	Gaps	**0.014**	**0.043**	**0.015**	**-0.025**	**0.031**	**-0.016**
HS	Original	0.256	0.437	0.473	0.555	0.434	0.485
	revised	0.278	0.476	0.458	0.503	0.4	0.463
	Gaps	**-0.022**	**-0.039**	**0.015**	**0.052**	**0.034**	**0.022**
C	Original	0.309	0.475	0.489	0.497	0.522	0.529
	revised	0.287	0.471	0.479	0.47	0.468	0.492
	Gaps	**0.022**	**0.004**	**0.01**	**0.027**	**0.054**	**0.037**
OE	Original	0.322	0.441	0.481	0.516	0.518	0.529
	revised	0.302	0.442	0.444	0.479	0.451	0.48
	Gaps	**0.02**	**-0.001**	**0.037**	**0.037**	**0.067**	**0.049**

Table 3: Results of shuffling the order of options. A & H: Arts & Humanities. SC: Social Sciences. NS: Natural Sciences. OS: Other Subjects. PS: Primary School. JHS: Junior High School. HS: High School. C: College. OE: Other Education. InternLM-7B: InternLM-7B-Chat. Baichuan2-7B: Baichuan2-7B-Chat. Qwen-7B: Qwen-7B-Chat.

Cluster	Types	ChatGLM1-6B	ChatGLM2-6B	ChatGLM3-6B	InternLM-7B	Baichuan2-7B	Qwen-7B
A & H	Original	0.308	0.478	0.49	0.568	0.524	0.546
	revised	0.298	0.359	0.364	0.439	0.293	0.392
	Gaps	**0.01**	**0.119**	**0.126**	**0.129**	**0.231**	**0.154**
SS	Original	0.365	0.532	0.572	0.586	0.599	0.612
	revised	0.331	0.414	0.397	0.439	0.335	0.424
	Gaps	**0.034**	**0.118**	**0.175**	**0.147**	**0.264**	**0.188**
NS	Original	0.255	0.452	0.443	0.45	0.427	0.457
	revised	0.313	0.381	0.323	0.373	0.286	0.374
	Gaps	**-0.058**	**0.071**	**0.12**	**0.077**	**0.141**	**0.083**
OS	Original	0.343	0.468	0.518	0.543	0.54	0.543
	revised	0.315	0.354	0.367	0.384	0.286	0.373
	Gaps	**0.028**	**0.114**	**0.151**	**0.159**	**0.254**	**0.17**
PS	Original	0.26	0.407	0.454	0.528	0.407	0.465
	revised	0.259	0.334	0.349	0.398	0.266	0.309
	Gaps	**0.001**	**0.073**	**0.105**	**0.13**	**0.141**	**0.156**
MS	Original	0.323	0.639	0.587	0.604	0.497	0.563
	revised	0.326	0.455	0.387	0.494	0.325	0.443
	Gaps	**-0.003**	**0.184**	**0.2**	**0.11**	**0.172**	**0.12**
HS	Original	0.256	0.437	0.473	0.555	0.434	0.485
	revised	0.316	0.376	0.349	0.424	0.3	0.387
	Gaps	**-0.06**	**0.061**	**0.124**	**0.131**	**0.134**	**0.098**
C	Original	0.309	0.475	0.489	0.497	0.522	0.529
	revised	0.319	0.388	0.355	0.389	0.307	0.392
	Gaps	**-0.01**	**0.087**	**0.134**	**0.108**	**0.215**	**0.137**
OE	Original	0.322	0.441	0.481	0.516	0.518	0.529
	revised	0.308	0.335	0.344	0.387	0.272	0.372
	Gaps	**0.014**	**0.106**	**0.137**	**0.129**	**0.246**	**0.157**

Table 4: Results of rewriting questions. A & H: Arts & Humanities. SC: Social Sciences. NS: Natural Sciences. OS: Other Subjects. PS: Primary School. JHS: Junior High School. HS: High School. C: College. OE: Other Education. InternLM-7B: InternLM-7B-Chat. Baichuan2-7B: Baichuan2-7B-Chat. Qwen-7B: Qwen-7B-Chat.

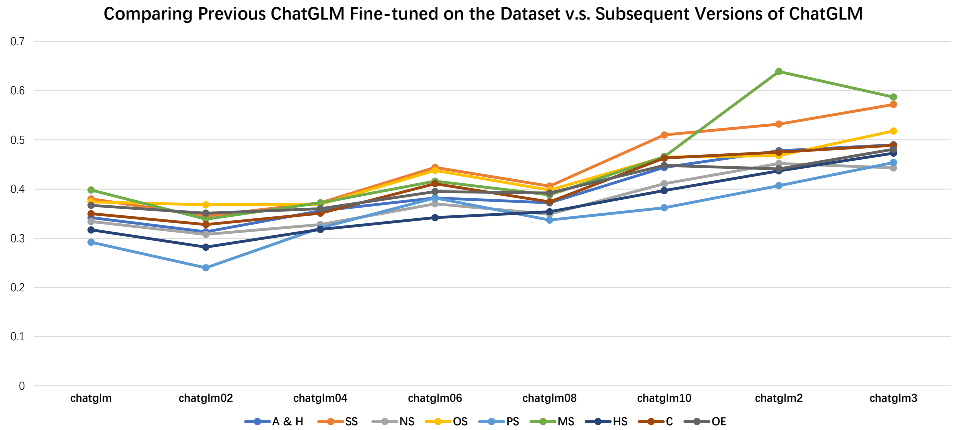

Figure 2: The results of contaminational fine-tuning. A & H: Arts & Humanities. SC: Social Sciences. NS: Natural Sciences. O: Other.

dicating that the development of training LLMs should incorporate more knowledge than previous versions, including insights from human evolution.

4.2.3 Task of Removal Questions

This task is designed to test whether the LLM can provide the correct answer without the question if it has been trained on question-answering pairs. We assessed six Chinese LLMs in M3KE, and the results are presented in Table 2. Most LLMs were impacted by this task, but ChatGLM1 appears to perform well, with even higher accuracy in two clusters than before. This suggests that ChatGLM1 might have been trained on multiple-choice questions related to those clusters in M3KE, specifically focusing on Nature Science at the subject level and High School at the education level. As ChatGLM versions progress, the impact on ChatGLM2 and ChatGLM3 becomes more pronounced, leading to a significant decrease in performance. This indicates that the training data for the later versions of ChatGLM may not contain the same questions as those in M3KE. Similarly, other LLMs like InternLM-7B-Chat, Baichuan2-7B-Chat, and Qwen-7B-Chat show a similar trend to ChatGLM2 and ChatGLM3. While it appears that newer LLMs may be predicting answers based on the questions rather than relying solely on memorization, it does not necessarily mean that the training data for these newer models lacks such knowledge.

The following question is whether LLMs effectively handle this knowledge? In other words, if LLMs truly master this knowledge, they should be able to address these questions across various scenarios. Consequently, we applied M3KE to different versions to assess the robustness of LLMs in the subsequent section.

4.3 Results of Robustness

In this section, we seek to assess the robustness of LLMs by modifying M3KE. This includes altering the sequence of options and rephrasing the original question. The core hypothesis here is that if an LLM comprehends the information, it should deliver comparable results with the unaltered test data. Hence, we adjusted M3KE using three approaches: rearranging option sequences, rephrasing questions, and combining shuffled options with rewritten questions. Furthermore, we introduce three LLMs from different companies in this segment - specifically InternLM-7B-Chat, Baichuan2-7B-Chat, and Qwen-7B-Chat - all of which exhibit impressive performance on M3KE.

4.3.1 Results of Shuffling the Order of Options

Table 3 shows the difference between the original and revised results on M3KE. The most significant decrease is observed at the primary school level for ChatGLM3, InternLM-7B-Chat, Baichuan2-7B-Chat, and Qwen-7B-Chat. Additionally, these language models, except for Baichuan2-7B-Chat, demonstrate relatively consistent performance in social science and natural science at the subject level, as well as in middle school, high school, and college at the education level. The largest deviation of 0.052 is seen in high school by InternLM-7B-Chat. Notably, ChatGLM2 remains consistent in this task, with only four cluster results decreasing.

Cluster	Types	ChatGLM1-6B	ChatGLM2-6B	ChatGLM3-6B	InternLM-7B	Baichuan2-7B	Qwen-7B
A & H	Original	0.308	0.478	0.49	0.568	0.524	0.546
	revised	0.303	0.353	0.364	0.426	0.298	0.366
	Gaps	**0.005**	**0.125**	**0.126**	**0.142**	**0.226**	**0.18**
SS	Original	0.365	0.532	0.572	0.586	0.599	0.612
	revised	0.315	0.386	0.384	0.421	0.319	0.409
	Gaps	**0.05**	**0.146**	**0.188**	**0.165**	**0.28**	**0.203**
NS	Original	0.255	0.452	0.443	0.45	0.427	0.457
	revised	0.288	0.353	0.32	0.356	0.286	0.355
	Gaps	**-0.033**	**0.099**	**0.123**	**0.094**	**0.141**	**0.102**
OS	Original	0.343	0.468	0.518	0.543	0.54	0.543
	revised	0.295	0.355	0.337	0.389	0.277	0.361
	Gaps	**0.048**	**0.113**	**0.181**	**0.154**	**0.263**	**0.182**
PS	Original	0.26	0.407	0.454	0.528	0.407	0.465
	revised	0.306	0.298	0.293	0.389	0.231	0.349
	Gaps	**-0.046**	**0.109**	**0.161**	**0.139**	**0.176**	**0.116**
MS	Original	0.323	0.639	0.587	0.604	0.497	0.563
	revised	0.307	0.433	0.392	0.492	0.313	0.404
	Gaps	**0.016**	**0.206**	**0.195**	**0.112**	**0.184**	**0.159**
HS	Original	0.256	0.437	0.473	0.555	0.434	0.485
	revised	0.282	0.382	0.336	0.392	0.292	0.36
	Gaps	**-0.026**	**0.055**	**0.137**	**0.163**	**0.142**	**0.125**
C	Original	0.309	0.475	0.489	0.497	0.522	0.529
	revised	0.3	0.352	0.348	0.374	0.304	0.376
	Gaps	**0.009**	**0.123**	**0.141**	**0.123**	**0.218**	**0.153**
OE	Original	0.322	0.441	0.481	0.516	0.518	0.529
	revised	0.298	0.352	0.336	0.378	0.277	0.355
	Gaps	**0.024**	**0.089**	**0.145**	**0.138**	**0.241**	**0.174**

Table 5: Results of combining rewritten questions and shuffled options. A & H: Arts & Humanities. SC: Social Sciences. NS: Natural Sciences. OS: Other Subjects. PS: Primary School. JHS: Junior High School. HS: High School. C: College. OE: Other Education. InternLM-7B: InternLM-7B-Chat. Baichuan2-7B: Baichuan2-7B-Chat. Qwen-7B: Qwen-7B-Chat.

4.3.2 Results of Rewriting Questions

Table 4 shows the performance impact of rewriting each question through prompting GPT-4. Compared to the previous method, we observe significant effects on most language models, particularly those excelling in original questions and released post M3KE. Within the ChatGLM category, the decline corresponds with the ChatGLM version, with ChatGLM3-6B, the latest model, experiencing the most reduction. ChatGLM1-6B, publicly available before M3KE, demonstrates similar performance. Notably, Baichuan2-7B-Chat appears to struggle with the modified questions, with the largest decrease of 0.264 in the social science cluster. InternLM-7B-Chat and Qwen-7B-Chat exhibit the most substantial reductions in other subject clusters and social science, with reductions of 0.159 and 0.188, respectively. Regarding educational levels, the most significant decreases are seen in other subjects for Baichuan2-7B-Chat and Qwen-7B-Chat, and in high school for InternLM-7B-Chat.

4.3.3 Results of Rewriting Questions with Shuffled Options

We merged the two tasks above, creating a benchmark with rewritten questions and reorganized option orders. This approach aligns with the task of question rewriting, as indicated in Table 5. It implies that existing Chinese LLMs are more attuned to the question descriptions than to the rearranged options, leading to observations that stronger LLMs might be trained with more structured questions, yet they may not grasp such knowledge types effectively. This indicates a need to reconsider the current advancements of Chinese LLMs focused on disciplinary knowledge benchmarks and prioritize robustness over ultimate performance.

5 Conclusion

In this paper, we have conducted a series of experiments to explore current progresses of Chinese LLMs on the discipline knowledge benchmark. We evaluated six Chinese SFT/RLHF LLMs belong to different groups to whether the new generation LLM memories more knowledge than the previous one, and the LLM taking more knowledge is able to handle those questions with different descriptions. Experiment results suggest although the newer LLM memorizes more knowledge, it still struggles with variations on the question, especially the description of question has more impact on LLMs.

Given that data contamination may pervade across different dimensions of LLM evaluation, we are keen to encourage the community further investigate current performance on public benchmarks.

Ethics Statement

The research process adheres strictly to the ACL Ethics Policy. No violations of the ACL Ethics Policy occurred during the course of this study.

Acknowledgements

The present research was partially supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). Chuang Liu is also supported by China Scholarship Council (No.202106250144). We would like to thank the anonymous reviewers for their insightful comments.

References

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *CoRR*, abs/2402.01781.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2024. LAiW: A chinese legal large language models benchmark.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *CoRR*, abs/2311.09783.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

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.

Lingyue Fu, Huacan Chai, Shuang Luo, Kounianhua Du, Weiming Zhang, Longteng Fan, Jiayi Lei, Renting Rui, Jianghao Lin, Yuchen Fang, Yifan Liu, Jingkuan Wang, Siyuan Qi, Kangning Zhang, Weinan Zhang, and Yong Yu. 2023. CodeApex: A bilingual programming evaluation benchmark for large language models. *CoRR*, abs/2309.01940.

Huibin Ge, Chenxi Sun, Deyi Xiong, and Qun Liu. 2021. Chinese WPLC: A Chinese dataset for evaluating pretrained language models on word prediction given long-range context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3770–3778, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huibin Ge, Xiaohu Zhao, Chuang Liu, Yulong Zeng, Qun Liu, and Deyi Xiong. 2022. TGEA 2.0: A large-scale diagnostically annotated dataset with benchmark tasks for text generation of pretrained language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances* *in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18099–18107. AAAI Press.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Rose Hadshar. 2023. A review of the evidence for existential risk from AI via misaligned power-seeking. *CoRR*, abs/2310.18244.

Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025, Online. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yimin Jing, Renren Jin, Jiahao Hu, Huishi Qiu, Xiaohua Wang, Peng Wang, and Deyi Xiong. 2023. Follow-Eval: A multi-dimensional benchmark for assessing the instruction-following capability of large language models. *CoRR*, abs/2311.09829.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback.

Changmao Li and Jeffrey Flanigan. 2024. Task Contamination: Language models may not be few-shot anymore. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18471–18480. AAAI Press.

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. *CoRR*, abs/2306.09212.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. An open source data contamination report for large language models.

Chuang Liu, Renren Jin, Yuqi Ren, and Deyi Xiong. 2024a. LHMKE: A large-scale holistic multi-subject knowledge evaluation benchmark for Chinese large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10476–10487, Torino, Italia. ELRA and ICCL.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023a. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Liutao, Jinwang Song, Hongying ZAN, Sun Li, and Deyi Xiong. 2024b. OpenEval: Benchmarking chinese LLMs across capability, alignment and safety. In *ACL 2024 System Demonstration Track*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023b. AlignBench: Benchmarking chinese alignment of large language models. *CoRR*, abs/2311.18743.

Yan Liu, Renren Jin, Lin Shi, Zheng Yao, and Deyi Xiong. 2024c. FineMath: A fine-grained mathematical evaluation benchmark for chinese large language models. *CoRR*, abs/2403.07747.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *CoRR*, abs/2310.17623.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver

Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, et al. 2023. TPTU: large language model-based ai agents for task planning and tool usage. *arXiv preprint arXiv:2308.03427*.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10776–10787. Association for Computational Linguistics.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. 2024. Rethinking llm memorization through the lens of adversarial compression. *arXiv preprint arXiv:2404.15146*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *CoRR*, abs/2309.15025.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023b. RoleEval: A bilingual role evaluation benchmark for large language models. *CoRR*, abs/2312.16132.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2024a. CORECODE: A common sense annotated dialogue dataset with benchmark tasks for chinese large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18952–18960. AAAI Press.

Ling Shi and Deyi Xiong. 2024. CRiskEval: A Chinese multi-level risk evaluation benchmark dataset for large language models. *arXiv preprint arXiv:2406.04752*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024b. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *CoRR*, abs/2304.10436.

InternLM Team. 2023. InternLM: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. CMB: A comprehensive medical benchmark in chinese. *CoRR*, abs/2308.08833.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. SuperCLUE: A comprehensive chinese large language model benchmark. *CoRR*, abs/2307.15020.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023b. Rethinking benchmark and contamination for language models with rephrased samples. *CoRR*, abs/2311.04850.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10466–10475, Torino, Italia. ELRA and ICCL.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023a. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2023b. Evaluating the generation capabilities of large chinese language models. *CoRR*, abs/2308.04823.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023a. FinEval: A chinese financial domain knowledge evaluation benchmark for large language models. *CoRR*, abs/2308.09975.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023b. SafetyBench: Evaluating the safety of large language models with multiple choice questions. *CoRR*, abs/2309.07045.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Shen Zheng, Yuyu Zhang, Yijie Zhu, Chenguang Xi, Pengyang Gao, Xun Zhou, and Kevin Chen-Chuan Chang. 2023. GPT-Fathom: Benchmarking large language models to decipher the evolutionary path towards GPT-4 and beyond. *CoRR*, abs/2309.16583.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your LLM an evaluation benchmark cheater. *CoRR*, abs/2311.01964.