Empirical Study on Data Attributes Insufficiency of Evaluation Benchmarks for LLMs

Chuang Liu[1], **Renren Jin**[1], **Zheng Yao**[2], **Tianyi Li**[3], **Liang Cheng**[3], **Mark Steedman**[3], **Deyi Xiong**[1]*
[1] College of Intelligence and Computing, Tianjin University
[2] School of Electrical Engineering and Computer Science, University of Queensland
[3] School of Informatics, University of Edinburgh
{liuc_09,rrjin,dyxiong}@tju.edu.cn

Abstract

Previous benchmarks for evaluating large language models (LLMs) have primarily emphasized quantitative metrics, such as data volume. However, this focus may neglect key qualitative data attributes that can significantly impact the final rankings of LLMs, resulting in unreliable leaderboards. In this paper, we investigate whether current LLM benchmarks adequately consider these data attributes. We specifically examine three attributes: diversity, redundancy, and difficulty. To explore these attributes, we propose a framework with three separate modules, each designed to assess one of the attributes. Using a method that progressively incorporates these attributes, we analyze their influence on the benchmark. Our experimental results reveal a meaningful correlation between LLM rankings on the revised benchmark and the original benchmark when these attributes are accounted for. These findings indicate that existing benchmarks often fail to meet all three criteria, highlighting a lack of consideration for multifaceted data attributes in current evaluation datasets.

1 Introduction

Large Language Models (LLMs) such as Llama (Touvron et al., 2023b; Dubey et al., 2024) and GPT-4 (OpenAI, 2023) have redefined the boundaries of natural language processing, delivering remarkable capabilities in understanding and applying knowledge ranging from social science to nature science. The evaluation of these models primarily relies on broadly constructed evaluation benchmarks (Guo et al., 2023), which are developed through methods including manual (Hendrycks et al., 2021), automated (Li et al., 2024), and semi-automated (Huang and Xiong, 2024) methods that involve either human-authored, machine-generated, or both inputs. The tasks covered in these benchmarks span discipline knowledge (Hendrycks et al., 2021; Liu et al., 2023; Huang et al., 2023), instruction following (Zhou et al., 2023) and alignment (Liu et al., 2024b; Sun et al., 2023; Zhang et al., 2023d), where each benchmark is used to rank the strength of various LLMs.

However, these publicly available benchmarks (Gu et al., 2024) often highlight their data scale but seldom mention other data attributes such as diversity, redundancy, and difficulty. Diversity indicates whether the dataset comprehensively covers varied knowledge points, redundancy refers to the presence of similar or duplicate questions that may lead to a waste of computational resource, and difficulty signifies whether the collected questions possess sufficient discriminative power. Thus these attributes are equally reflective of the benchmark's quality as data volume, if not more so.

To mitigate this gap, in this paper, we aim to investigate whether the current evaluation benchmarks for LLMs possess these data attributes. This also means we must answer two critical questions. First, how do we extract these three attributes from the benchmarks? Second, how do we measure whether these attributes are sufficiently considered within the benchmarks?

For the first question, we address it by proposing a novel framework to individually explore these data attributes within the benchmarks. This framework comprises three distinct modules: (1) a diversity module that employs DBSCAN (Ester et al., 1996) for analyzing topic variety and distribution, (2) a difficulty module that categorizes questions into various difficulty levels based on performance across 20 LLMs, drawing inspiration from standards used in human examinations, and (3) a redundancy module that utilizes text entailment tasks (Bowman et al., 2015) to assess conceptual overlaps in each question. Then, we assign a label to each question in the benchmark based on these attributes.

* Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 6024–6038
January 19–24, 2025.

For the second question, based on the data attributes we have already labeled, we adopt a step-by-step pipeline to measure the presence of these attributes in the original datasets. Specifically, we start by listing the performance rankings of 40 LLMs on the original datasets as a reference. Then, we examine the changes in the correlation coefficients (Spearman, 1904; Kendall, 1938) between model rankings after progressively considering the attributes of diversity, redundancy, and difficulty, as labeled in our framework. We posit that if a dataset satisfies all three attributes, the inclusion of these considerations should correlate positively with the original rankings. Conversely, a decrease in correlation upon considering a specific attribute suggests a lack of consideration for that attribute in the original dataset.

Since our study is language-agnostic, we select three LLM benchmarks of the same type for our experiments, including two Chinese benchmarks, M3KE (Liu et al., 2023) and CMMLU (Li et al., 2023a) and one English benchmark, MMLU (Hendrycks et al., 2021). We further categorize these benchmarks into three categories: Humanities, Social, and STEM, because all three benchmarks are oriented towards subject knowledge, making these categories comparable.[1]

Overall, the majority of benchmarks lack consideration for redundancy, as evidenced by a notable decrease in the correlation of rankings when this attribute is considered, with the exception of the STEM subject in M3KE (Liu et al., 2023). Conversely, the level of difficulty is well-balanced in more than half of the benchmarks. However, because redundancy only has two labels—redundant and non-redundant—while the granularity of difficulty classification is more varied, we further divide difficulty into three levels: easy, normal, and hard, and conduct a deeper analysis. Experiment results reveal that M3KE (Liu et al., 2023) and CMMLU (Li et al., 2023a) predominantly feature normal and easy levels of difficulty, whereas MMLU (Hendrycks et al., 2021) maintains a better balance across all three categories. However, the actual difference in difficulty levels is much smaller in the STEM tasks than in the humanities and social categories. This suggests that current benchmarks do not adequately consider these attributes, and it is recommended that future benchmark construction should fully consider these factors rather than merely increasing scale.

Our main contributions are summarized as follows.

- We propose an automated analysis framework aimed at detecting the attributes of LLM benchmarks, including diversity, redundancy, and difficulty.
- we also design a pipeline to measure the quality of LLM evaluation benchmarks by incrementally adding attributes and observing changes in the correlation coefficients with the original dataset rankings.
- Extensive experiments show that current benchmarks do not adequately balance consideration of each attribute.

2 Related Work

Currently, evaluation benchmarks for LLMs have expanded across multiple dimensions (Shevlane et al., 2023; Guo et al., 2023), including capability, value alignment, and safety. Among these, capability-oriented benchmarks are the most diverse in terms of quantity and type. They cover a wide range of abilities such as general knowledge (Zhang et al., 2023c; Yu et al., 2024a; Zhang et al., 2023b; Yu et al., 2024c; Liu et al., 2024a), instruction following (Jing et al., 2023), commonsense reasoning (He et al., 2021; Shi et al., 2024), mathematical reasoning (Wei et al., 2023; Liu et al., 2024d), tool usage (Zhuang et al., 2023), agent evaluation (Li et al., 2023b; Guo et al., 2024; Zhou et al., 2024; Liu et al., 2024c) and machine programming (Fu et al., 2023; Peng et al., 2024). Value alignment evaluation focuses on testing LLMs' performance in areas like bias (Zhang et al., 2023a; Zhou et al., 2022; Huang and Xiong, 2024), offensiveness (Yang and Lin, 2020; Jiang et al., 2022; Deng et al., 2022), and social morality (Yu et al., 2024b). Lastly, safety evaluations (Perez et al., 2022; Shi and Xiong, 2024) are conducted to monitor whether LLMs may cause catastrophic behavioral risks (Hendrycks et al., 2023).

In terms of language, both English and Chinese have become the primary languages for current LLM evaluation benchmarks, with corresponding datasets in each language available for assessment across various dimensions. However, apart from benchmarks based on language-specific cultural

[1]Other categories which often involve culturally specific questions are not included.

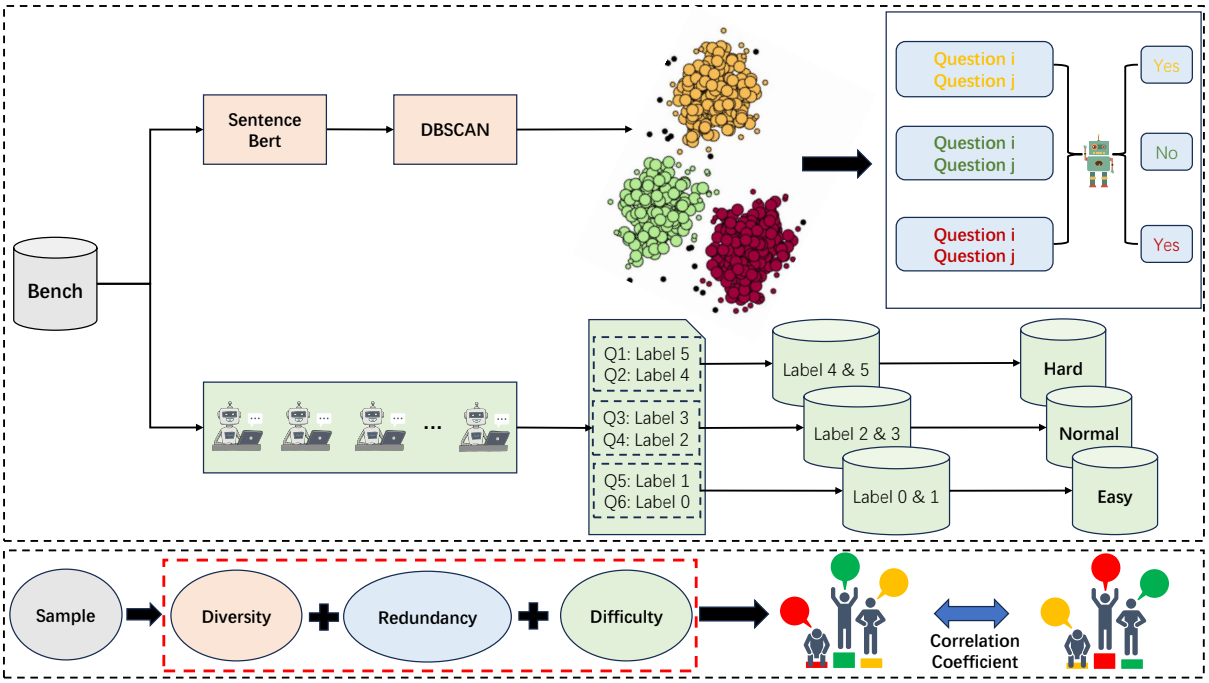

Figure 1: Diagrams of the proposed framework (upper part) and pipeline (lower part). Orange, blue, and green represent diversity, redundancy, and difficulty, respectively.

contexts, the motivation and content of most benchmarks are generally universal. Unfortunately, when these datasets are released, most authors only emphasize the differences in language and the scale of the dataset, with little introduction to other data attributes in the benchmarks, such as diversity, redundancy, and difficulty. As other data-centric studies (Jha et al., 2023; Xia et al., 2024; Xie et al., 2023) have already revealed the importance of data perspective for training LLMs, the core purpose of this paper is to explore whether current LLM benchmarks overlook the consideration of these data attributes and to demonstrate the potential impact different data attributes can have on evaluation leaderboards.

At the same time, researchers have begun to reflect on issues present in previous benchmarks (Singh et al., 2024). Gema et al. (2024) revisit MMLU (Hendrycks et al., 2021) and identify defects in data quality through manual comparison. Perlitz et al. (2024) investigate ways to reduce the computational costs of evaluating language models without compromising the reliability of the results. Mazumder et al. (2023) emphasizes fostering innovation in data-centric AI by enhancing competition, comparability, and reproducibility. AutoBencher (Li et al., 2024) addresses the challenge of balancing three key criteria in dataset creation: salience, novelty, and difficulty.

This paper differs from these works in three significant ways. First, we focus on three attributes, data diversity, redundancy, and difficulty, to explore current benchmarks. Second, we design an automated framework to uncover these attributes within benchmarks. Finally, we employ a method that incrementally introduces these attributes to observe fluctuations in correlation with the original benchmark rankings, thereby reflecting whether these attributes were adequately considered in the original benchmarks.

3 Methodology

Figure 1 displays the proposed attribute mining framework analysis pipeline.

3.1 Framework

As demonstrated in Figure 1 (upper part), the framework is composed of three modules designed to evaluate diversity, difficulty, and redundancy. Specifically, the diversity module employs a density-based clustering algorithm to detect benchmarks containing a greater variety of topics and consistent question distribution. The difficulty module measures question complexity through accuracy distribution across different models. Lastly, the redundancy module focuses on identifying and filtering out similar or duplicate questions within the benchmarks.

3.1.1 Module for Diversity

The diversity module is based on DBSCAN (Ester et al., 1996), a density-based clustering algorithm tailored to detect benchmarks with a broad and consistent distribution of topics. Initially, benchmarks within the same dimension, such as MMLU, CMMLU, and M3KE in the discipline dimension, are grouped for assessment. It is essential that benchmarks in the same group contain comparable categories to ensure a consistent evaluation.

We employed Sentence-BERT (Reimers and Gurevych, 2019) as our text encoding model to convert textual content into embeddings suitable for clustering. After clustering, we compared benchmarks across each dimension by analyzing the number of questions and topics, as well as their variance and standard deviation.

3.1.2 Module for Difficulty

Inspired by standards used in human examinations, this module assesses the difficulty of each benchmark, with the premise that a high-quality benchmark should balance the distribution of question difficulty. We used twenty LLMs, both open-source and proprietary, exhibiting varied performance across benchmarks. These LLMs evaluate the selected benchmarks, and questions are classified into six difficulty levels based on the number of correct responses: a question is tagged with a difficulty level of 0 if more than 15 LLMs answer it correctly. For every three fewer models that answer correctly, the difficulty level decreases by one.

A benchmark should ideally contain questions of varying difficulty levels instead of focusing solely on one. We will compare the impact of each difficulty attribute of the benchmarks across these levels in Section 5.

3.1.3 Module for Redundancy

In the redundancy module, we define the detection of redundant questions as a text entailment task. This method leverages the natural language understanding capabilities of models to determine whether two questions within the same cluster are conceptually similar enough to be considered redundant.

To implement this, each question pair from the same cluster is formatted into a structured query that resembles a natural language understanding task:

Prompt: I have two multiple-choice questions and I need a simple answer to determine if they test the same concepts. Here are the questions: *question 1* and *question 2*. Do these two questions test the same concepts? Please answer with "Yes" or "No".

Using this format, we employ GPT-4 (OpenAI, 2023) to assess whether the two questions are conceptually similar because the advanced LLM ensures that we are accurately identifying redundancies based on conceptual similarity rather than superficial textual characteristics. The model's responses ("Yes" or "No") indicate the presence of redundancy. Questions that are determined to be testing the same concepts by receiving a "Yes" response are flagged as redundant.

The effectiveness of this module is evaluated by measuring the impact of removing identified redundancies on the overall performance metrics of LLMs on the benchmark. A lower variance in performance before and after redundancy removal indicates a more distinct and essential set of questions, thereby validating the quality of the benchmark.

3.2 Pipeline

Once the data is assessed through our framework, questions within the same topic are clustered into groups, and questions within each cluster are labeled by the redundancy module. In addition, we determine the varying difficulty levels of these questions through the application of our dedicated difficulty module, which evaluates and categorizes them based on their complexity. Furthermore, we also establish a baseline ranking that encompasses a total of 40 LLMs, which is grounded in the performance metrics derived from the analysis of the original dataset.

Then, we begin to gradually explore the significance of each attribute in the dataset, as illustrated in the Figure 1 (lower part). Specifically, we first consider diversity for our initial correlation estimation, selecting data from each cluster to derive new model rankings and calculate the correlation coefficient with the original rankings as the initial coefficient. Next, we filter out questions marked as redundant, i.e., those with a "Yes" label, and derive the model rankings after eliminating those questions. If the updated ranking's correlation coefficient is higher than the initial coefficient, it can be inferred that the original dataset may include this attribute, and vice versa. Finally, considering the difficulty attribute means we eliminate questions with duplicate difficulty labels to balance the distribution of difficulty levels in the data, with trends

Benchmark	Subject	Attribute	Spearman ↑	Kendall ↑	Rank Change ↓	Standard Deviation ↓
M3KE	Humanity	Diversity	**0.992**	0.941	**50**	**0.786**
		DR	0.989	0.933	60	0.943
		DDR	0.986	0.915	72	1.131
	Social	Diversity	**0.996**	**0.964**	**32**	**0.503**
		DR	**0.996**	**0.964**	34	0.534
		DDR	0.995	0.958	36	0.566
	STEM	Diversity	0.991	0.939	54	0.849
		DR	**0.992**	**0.943**	**50**	**0.786**
		DDR	0.990	0.937	54	0.849
CMMLU	Humanity	Diversity	**0.996**	0.956	38	0.597
		DR	**0.996**	**0.958**	**36**	**0.566**
		DDR	0.990	0.927	64	1.006
	Social	Diversity	**0.994**	**0.952**	**44**	**0.691**
		DR	0.993	0.947	**44**	**0.691**
		DDR	0.992	0.943	48	0.754
	STEM	Diversity	**0.987**	**0.923**	**66**	**1.037**
		DR	0.985	0.917	72	1.131
		DDR	**0.987**	0.915	70	1.100
MMLU	Humanity	Diversity	**0.997**	**0.972**	**26**	**0.409**
		DR	0.995	0.959	38	0.597
		DDR	0.992	0.941	50	0.786
	Social	Diversity	**0.991**	**0.935**	**58**	**0.911**
		DR	0.989	0.921	66	1.037
		DDR	0.976	0.883	98	1.540
	STEM	Diversity	**0.996**	**0.958**	**36**	**0.566**
		DR	0.992	0.943	46	0.723
		DDR	0.991	0.933	52	0.817

Table 1: Overall results. ↑ represents that the higher is better, while ↓ denotes that the lower is better. Diversity & Redundancy (DR): assessing the inclusion of the redundancy attribute. Diversity & Difficulty & Redundancy (DDR): assessing the inclusion of the difficulty attribute. Correlation.C: Correlation Coefficients.

in the correlation coefficient changes following the same logic as above.

Ultimately, by incrementally controlling attributes through this pipeline, we can investigate the impact of different data attributes on the original rankings to explore possible data attributes insufficiency in the benchmark.

4 Experiment

In this section, we first introduce the datasets, models, and correlation metrics used in our experiments. We then present the main experimental results, detailing the performance of our analytical framework and how each attribute—diversity, redundancy, and difficulty—affects the rankings of LLMs in comparison to the original rankings.

4.1 Assessed Datasets

We selected three disciplinary knowledge benchmarks for our experiments, including an English benchmark, MMLU, and two Chinese benchmarks, M3KE and CMMLU. Firstly, disciplinary knowledge is often gathered manually from publicly accessible exam questions on the Internet, and since disciplinary knowledge is equivalent regardless of whether the educational context is in English or Chinese, these benchmarks are comparable. Secondly, although the knowledge points and examination points within these disciplines are fixed, the sources of the questions are diverse, including practice problems, mock exams, regional unified examinations, and national examinations. Therefore, the data attributes within these benchmarks can help us understand how each attribute was considered during the data collection process.

We conducted our investigations into the three benchmarks across three disciplinary categories: Humanities, Social Sciences, and STEM. It is important to note that after clustering each disciplinary subject, we selected a subset of data from each cluster for subsequent experiments. This selection was necessary because our experiments on

Benchmark	Subject	Topics	Avg. Count of Qs
M3KE	Humanity	119	30.353
	Social	224	27.777
	STEM	271	30.118
CMMLU	Humanity	80	31.113
	Social	117	31.214
	STEM	97	26.093
MMLU	Humanity	200	23.355
	Social	78	26.526
	STEM	228	23.969

Table 2: Diversity statistics for the three evaluated benchmarks. Avg. Count of Qs: Average counts of Questions.

redundancy required using GPT-4 (OpenAI, 2023) to evaluate paired questions. Given that we have N questions, which would result in $\frac{N\times(N-1)}{2}$ pairs, the cost of processing all the data would be prohibitively high. Therefore, a selective approach was necessary to manage the scale and feasibility of the experiments efficiently.

4.2 Model

To accurately depict the difficulty attribute and create objective leaderboards, we first labeled all questions in each benchmark using 60 LLMs, with 0 representing an incorrect answer and 1 a correct one. We then selected 20 LLMs, varying from 0.5B to 72B parameters, to assess model difficulty. The remaining 40 models, with parameters ranging from 0.5B to 110B, were used to generate a baseline leaderboard for each disciplinary subject within each benchmark. A detailed list of all models used in the experiments is provided in Appendix 4.

4.3 Evaluation Metrics

We used Spearman and Kendall metrics to assess correlations pre- and post-attribute control, while Rank Change and Standard Deviation show shifts in rankings.

Spearman Spearman (Spearman, 1904) correlation coefficient is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. .

Kendall Kendall (Kendall, 1938) is another non-parametric statistic used to measure the ordinal association between two measured quantities. Unlike the Spearman coefficient, Kendall's tau is more sensitive to outliers in rankings.

Rank Change This metric directly measures the discrepancy between two rankings over the original and revised benchmark.

Standard Deviation The metric helps understand the volatility of ranking changes; the larger the standard deviation, the more unstable the ranking changes.

4.4 Main Results

We conducted a statistical analysis based on the clustering results to determine the number of topics and the average number of questions per benchmark in order to measure diversity, as shown in Table 2. Overall, we find an imbalance in the subject categories within each benchmark. For instance, the number of topics in the humanities in M3KE and in social sciences in MMLU is only about half that of the other two categories. Although the topic numbers across categories in CMMLU are closer to each other, the total number of topics is significantly less diverse compared to the other benchmarks.

Since samples corresponding to each topic were selected, we used the LLM rankings associated with these samples as the leaderboard under controlled diversity. Table 1 describes the correlation and changes in rankings under controlled diversity, redundancy, and difficulty compared to the original rankings. Generally, while there is a high overall correlation between previous and current rankings, metrics like rank change and standard deviation indicate local fluctuations. In most disciplines, the best outcomes—stronger correlations and minimal rank changes—are observed when only diversity is controlled. However, as redundancy and difficulty are progressively included, over half of the disciplines show a trend of negative correlation with the original rankings, indicating a general oversight of these attributes in current benchmarks.

For M3KE, redundancy is only considered in STEM subjects, where controlling for this attribute improves ranking correlation and reduces both ranking changes and standard deviation. In contrast, in the humanities and social sciences, accounting for redundancy and difficulty led to greater deviations from the original rankings.

In CMMLU, redundancy is observed in Humanities through four indicators, but difficulty was not prioritized. This is evident from the notable change in rankings—an increase from 36 to 64—when we adjusted for difficulty by equalizing the distribu-

Benchmark	Subject	Difficulty	Spearman ↑	Kendall ↑	Rank Change ↓	Standard Deviation ↓
M3KE	Humanity	Easy	0.946	0.818	146	2.294
		Normal	**0.982**	**0.897**	**86**	**1.351**
		Hard	0.712	0.560	282	4.431
	Social	Easy	0.935	0.792	162	2.546
		Normal	**0.977**	**0.879**	**100**	**1.571**
		Hard	0.721	0.545	292	4.588
	STEM	Easy	0.904	0.731	208	3.268
		Normal	**0.946**	**0.802**	**154**	**2.420**
		Hard	0.669	0.477	330	5.185
CMMLU	Humanity	Easy	0.946	0.808	152	2.388
		Normal	**0.970**	**0.863**	**106**	**1.666**
		Hard	0.698	0.521	314	4.934
	Social	Easy	0.941	0.800	162	2.546
		Normal	**0.969**	**0.863**	**112**	**1.760**
		Hard	0.682	0.509	318	4.997
	STEM	Easy	0.911	0.752	198	3.111
		Normal	**0.946**	**0.810**	**144**	**2.623**
		Hard	0.658	0.481	344	5.405
MMLU	Humanity	Easy	0.685	0.505	**358**	**5.625**
		Normal	**0.696**	**0.511**	370	5.814
		Hard	0.409	0.283	502	7.888
	Social	Easy	0.722	0.533	**334**	**5.248**
		Normal	**0.727**	**0.535**	338	5.311
		Hard	0.447	0.295	470	7.385
	STEM	Easy	0.689	0.507	**358**	**5.625**
		Normal	**0.699**	**0.521**	366	5.751
		Hard	0.427	0.289	490	7.699

Table 3: Difficulty level results. ↑ denotes that the higher is better, while ↓ signifies that the lower is better.

tion of difficulty labels among the questions. This suggests a concentrated distribution of difficulty labels in the Humanities. Other disciplines exhibited smaller ranking changes, indicating a similar neglect of redundancy and difficulty considerations.

In the context of the MMLU evaluations, it becomes apparent that all disciplines tend to overlook important factors such as redundancy and difficulty. This oversight is evidenced by the increasingly significant discrepancies that arise between the new rankings and the original rankings. Adding more controlled attributes increases differences, highlighting the impact of neglected aspects on evaluation.

The experiments have confirmed that the current LLM evaluation benchmarks indeed do not sufficiently consider these data attributes. This is first reflected in the imbalance in diversity, with significant disparities in the number of topics. Additionally, the lack of consideration for redundancy and difficulty could lead to issues in the benchmarks, such as questions involving similar knowledge points and comparable levels of difficulty.

5 Fine-grained Analysis of the Difficulty Attribute

With a difficulty attribute ranging from 0 to 5, we can more precisely control difficulty labels to analyze fluctuations in the original rankings. We categorize these labels into Easy (less than 2), Normal (2 to 4), and Hard (greater than 4).

An observation of the rank changes within a particular difficulty type can provide meaningful insights. If we notice that a difficulty category exhibits only a minimal number of rank changes and aligns closely with the original rankings, this can be interpreted as a strong indication that the original data effectively represented this type of difficulty. On the other hand, when we encounter significant fluctuations within a specific difficulty category, this can serve as a clear signal that the original data may not have adequately captured the nuances of that particular category.

Table 3 compares original rankings to different difficulty levels across various subjects under benchmarks in this paper.

5.1 Analysis of Various Difficulty Attributes in the Humanity Subject

In the M3KE benchmark, there is a significant divergence between the original and adjusted rankings in the hard category. This indicates that the hard level in the M3KE contribute minimally to the rankings, suggesting that the difficulty of questions under the humanities subject of M3KE primarily comprises easy and normal questions. Furthermore, considering only the rankings at normal difficulty, which show the least fluctuation, suggests that questions at this level contribute slightly more to the rankings than easy questions.

Moving to the CMMLU benchmark, the shifts in rankings are generally more pronounced than in M3KE, especially when considering only the normal difficulty level. The changes between the original and adjusted rankings in the CMMLU benchmark are closer in the easy and normal levels than hard level, suggesting that this benchmark is also primarily composed of easy and normal questions.

Finally, MMLU exhibits a completely different trend compared to M3KE and CMMLU, where considering any single difficulty level alone leads to significant fluctuations in rankings. This indicates that the questions in MMLU have distinct boundaries in terms of difficulty, meaning that the differences between each difficulty level are very pronounced.

5.2 Analysis of Various Difficulty Attributes in the Social Subject

The social subject reflects a broader trend observed in the field of the humanities, particularly regarding the types of questions being presented across various datasets. In both M3KE and CMMLU, there is a notable predominance of Easy and Normal level questions, while the occurrence of Hard questions remains quite limited. This trend indicates a preference or perhaps a tendency to focus on questions that are more accessible and straightforward. On the other hand, when we examine MMLU, we can observe a distinct and pronounced differentiation among the various levels of question difficulty.

5.3 Analysis of Various Difficulty Attributes in the STEM Subject

Interestingly, the STEM subject exhibits fluctuations that are completely different from those in other categories.

For M3KE, even though its overall trend is relatively similar to the other two categories, notable variations can still be observed. Although considering only the hard questions results in the most significant fluctuations, the variations in the other two difficulty levels are also more pronounced compared to other subject categories, especially at the normal level. This suggests a decline in the contribution of normal level questions to the rankings, suggesting a reduced proportion of these questions, even though they still hold the highest correlation with the rankings.

For CMMLU, the correlation of normal and easy questions to the rankings has significantly decreased, particularly at the normal difficulty level. Compared to the three categories, questions of normal category have emerged as the difficulty level with the strongest correlation to rankings, implying that the STEM questions in CMMLU are predominantly moderate difficulty.

However, in MMLU, considering any single difficulty type alone shows less consistency with the original rankings, especially when compared to their performance with M3KE and CMMLU. This also indicates that, overall, MMLU possesses a more balanced difficulty distribution in STEM subjects.

6 Conclusion

In this paper, we have explored the consideration of data attributes in current LLM benchmarks, including diversity, redundancy, and difficulty. We initially propose an automated attribute mining framework to detect these attributes and then design a step-by-step process to assess each attribute's impact on the benchmark's original leaderboard to determine whether the original datasets lack these attributes. The experimental results indicate that redundancy is the most commonly overlooked attribute in LLM benchmarks, followed by difficulty. Further analysis reveals that changes in difficulty significantly affect the final model rankings. This trend highlights a systemic issue within the benchmark, suggesting that further adjustments are necessary to adequately account for these aspects. Therefore, we recommend that, in the construction of LLM evaluation benchmarks, a balance of various data attributes should be maintained, rather than relying solely on expanding dataset size, which can lead to computational resource waste and unreliable evaluation rankings. This balanced approach is crucial for achieving a reliable evaluation.

Limitations

Although we have conducted extensive experiments to investigate potential issues of inadequate consideration of data attributes in current LLM benchmarks and proposed a framework and pipeline solution to explore the impact of different data attributes on the original rankings of benchmarks, our work has two significant limitations. First, the definition of the quality of LLM benchmarks remains unclear, and there is no quantifiable metric to accurately represent the quality of a benchmark, so our comparison of the three attributes only covers a portion of this aspect. Second, due to cost considerations, we cannot conduct experiments on the entire dataset. Lastly, LLM benchmarks involve multiple dimensions, and in this study, we only select discipline knowledge-oriented LLM benchmarks. It is important to emphasize that we carefully select the most representative LLM benchmarks for our experiments, aiming to explore whether the construction process of LLM benchmarks adequately considers different attribute dimensions and to demonstrate their impact on benchmark rankings by controlling these attributes.

Ethics Statement

This work presents an Empirical Study on Data Attributes Insufficiency of Evaluation Benchmarks for LLMs. All data and models in this study are open sourced.

Acknowledgments

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, and et al. 2024. InternLM2 technical report. *CoRR*, abs/2403.17297.

Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li, Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng Zhang, Zhipeng Zhang, and Kun Han. 2024. Orion-14B: Open-source multilingual large language models. *CoRR*, abs/2401.12246.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong,

Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231. AAAI Press.

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.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2024. Are we done with mmlu? *CoRR*, abs/2406.04127.

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, Yufei Huang, and Deyi Xiong. 2024. CToolEval: A Chinese benchmark for LLM-powered agent evaluation in real-world API interactions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15711–15724, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395.

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*.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. 2023. LIMIT: less is more for instruction tuning across evaluation paradigms. *CoRR*, abs/2311.13133.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

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

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023b. API-Bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3102–3116. Association for Computational Linguistics.

Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. 2024. AutoBencher: Creating salient, novel, difficult datasets for language models. *CoRR*, abs/2407.08351.

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. 2023. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024b. AlignBench: Benchmarking chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11621–11640. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024c. AgentBench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Mark Mazumder, Colby R. Banbury, Xiaozhe Yao, Bojan Karlas, William Gaviria Rojas, Sudnya Frederick Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Raje, Max Bartolo, Evan Sabri Eyuboglu, Amirata Ghorbani, Emmett D. Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas W. Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen K. Paritosh, Ce Zhang, James Y. Zou, Carole-Jean Wu, Cody Coleman, Andrew Y. Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. DataPerf: Benchmarks for data-centric AI development. 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*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth

Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

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

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8383–8394. ELRA and ICCL.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2024. 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*.

Shweta Singh, Aayan Yadav, Jitesh Jain, Humphrey Shi, Justin Johnson, and Karan Desai. 2024. Benchmarking object detectors with COCO: A new path forward. *CoRR*, abs/2403.18819.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

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

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

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. CMATH: can your language model pass chinese elementary school math test? *arXiv preprint arXiv:2306.16636*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. DoReMi: Optimizing data mixtures speeds up language model pretraining. 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*.

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. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Hsu Yang and Chuan-Jie Lin. 2020. TOCP: A dataset for Chinese profanity processing. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 6–12, Marseille, France. European Language Resources Association (ELRA).

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu Gong, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2024a. KoLA: carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024b. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11817–11837, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024c. 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, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Bo-Wen Zhang, Liangdong Wang, Jijie Li, Shuhao Gu, Xinya Wu, Zhengduo Zhang, Boyan Gao, Yulong Ao, and Guang Liu. 2024. Aquila2 Technical Report. *arXiv preprint arXiv:2408.07410*.

Ge Zhang, Yizhi Li, Yaoyao Wu, Linyuan Zhang, Chenghua Lin, Jiayi Geng, Shi Wang, and Jie Fu. 2023a. CORGI-PM: A chinese corpus for gender bias probing and mitigation. *arXiv preprint arXiv:2301.00395*.

Sarah J Zhang, Samuel Florin, Ariel N Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, et al. 2023b. Exploring the mit mathematics and eecs curriculum using large language models. *arXiv preprint arXiv:2306.08997*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023c. Evaluating the performance of large language models on gaokao benchmark.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. *arXiv preprint arXiv:2202.08011*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue

Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A dataset for LLM question answering with external tools. 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*.

A Model Overview

Table 4 displays all the models used in the experiments conducted for this paper.

Model	Parameters	Pre-trained	Role
Yi-1.5-6B-Chat (Young et al., 2024)	6B	Yi 1.5	Leaderboard
Yi-1.5-34B-Chat (Young et al., 2024)	34B	Yi 1.5	Leaderboard
Yi-1.5-9B-Chat (Young et al., 2024)	9B	Yi 1.5	Leaderboard
Yi-34B-Chat (Young et al., 2024)	34B	Yi	Leaderboard
AquilaChat-7B (Zhang et al., 2024)	7B	Aquila	Leaderboard
aya-23-35B (Aryabumi et al., 2024)	35B	Aya	Leaderboard
c4ai-command-r-plus[2]	104B	C4AI	Leaderboard
c4ai-command-r-v01[3]	35B	C4AI	Leaderboard
Orion-14B-Chat (Chen et al., 2024)	14B	Orion	Leaderboard
Qwen2-1.5B-Instruct (Yang et al., 2024)	1.5B	Qwen 2	Leaderboard
Qwen2-7B-Instruct (Yang et al., 2024)	7B	Qwen 2	Leaderboard
Qwen1.5-72B-Chat (Bai et al., 2023)	72B	Qwen 1.5	Leaderboard
Qwen1.5-7B-Chat (Bai et al., 2023)	7B	Qwen 1.5	Leaderboard
Qwen2-57B-A14B-Instruct (Yang et al., 2024)	57B	Qwen 2	Leaderboard
Qwen1.5-110B-Chat (Bai et al., 2023)	110B	Qwen 1.5	Leaderboard
Qwen1.5-14B-Chat (Bai et al., 2023)	14B	Qwen 1.5	Leaderboard
Qwen1.5-MoE-A2.7B-Chat (Bai et al., 2023)	2.7B	Qwen 1.5	Leaderboard
Qwen1.5-1.8B-Chat (Bai et al., 2023)	1.8B	Qwen 1.5	Leaderboard
Qwen2-0.5B-Instruct (Yang et al., 2024)	0.5B	Qwen 2	Leaderboard
glm-4-9b-chat (Zeng et al., 2024)	9B	GLM 4	Leaderboard
Baichuan2-13B-Chat (Yang et al., 2023)	13B	Baichuan 2	Leaderboard
Baichuan-13B-Chat (Yang et al., 2023)	13B	Baichuan	Leaderboard
gemma-2b-it (Mesnard et al., 2024)	2B	Gemma 2	Leaderboard
gemma-2-2b-it (Mesnard et al., 2024)	2B	Gemma 2	Leaderboard
gemma-2-9b-it (Mesnard et al., 2024)	9B	Gemma 2	Leaderboard
gemma-7b-it (Mesnard et al., 2024)	7B	Gemma	Leaderboard
gemma-1.1-2b-it (Mesnard et al., 2024)	2B	Gemma	Leaderboard
internlm2.5-7b-chat (Cai et al., 2024)	7B	InternLM 2.5	Leaderboard
internlm-chat-20b (Cai et al., 2024)	20B	InternLM	Leaderboard
internlm2-chat-1_8b(*Caiet al.*, 2024)	1.8B	InternLM 2	Leaderboard
internlm2-chat-20b (Cai et al., 2024)	20B	InternLM 2	Leaderboard
internlm2-chat-7b (Cai et al., 2024)	7B	InternLM 2	Leaderboard
internlm2.5-$1_8b - chat$(*Caiet al.*, 2024)	1.8B	InternLM 2.5	Leaderboard
Llama-3-70B-Instruct (Dubey et al., 2024)	70B	Llama 3	Leaderboard
Llama-3.1-70B-Instruct (Dubey et al., 2024)	70B	Llama 3.1	Leaderboard
Llama-3-8B-Instruct (Dubey et al., 2024)	8B	Llama 3.1	Leaderboard
Llama-2-7b-chat-hf (Touvron et al., 2023a)	7B	Llama 2	Leaderboard
Mixtral-8x22B-Instruct-v0.1 (Jiang et al., 2024)	22B	Mixtral	Leaderboard
Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024)	7B	Mixtral	Leaderboard
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	7B	Mistral	Leaderboard
Mistral-Large-Instruct-2407 (Jiang et al., 2023)	2407	Mistral	Leaderboard
MiniCPM-2B-sft-bf16 (Hu et al., 2024)	2B	MiniCPM	Leaderboard
MiniCPM-2B-dpo-bf16 (Hu et al., 2024)	2B	MiniCPM	Leaderboard
XVERSE-13B-Chat[4]	13B	XVERSE	Leaderboard
XVERSE-7B-Chat[5]	7B	XVERSE	Leaderboard
Yi-6B-Chat (Young et al., 2024)	6B	Yi	Difficulty
AquilaChat2-7B (Zhang et al., 2024)	7B	Aquila 2	Difficulty
aya-23-8B (Aryabumi et al., 2024)	8B	Aya	Difficulty
Qwen1.5-0.5B-Chat (Bai et al., 2023)	0.5B	Qwen 1.5	Difficulty
Qwen2-72B-Instruct (Yang et al., 2024)	72B	Qwen 2	Difficulty
Qwen1.5-32B-Chat (Bai et al., 2023)	32B	Qwen 1,5	Difficulty
Qwen1.5-4B-Chat (Bai et al., 2023)	4B	Qwen 1.5	Difficulty
Baichuan2-7B-Chat (Yang et al., 2023)	7B	Baichuan 2	Difficulty
gemma-1.1-7b-it (Mesnard et al., 2024)	7B	Gemma 1.1	Difficulty
gemma-2-27b-it (Mesnard et al., 2024)	27B	Gemma 2	Difficulty
internlm2.5-20b-chat (Cai et al., 2024)	20B	InternLM 2.5	Difficulty
internlm-chat-7b (Cai et al., 2024)	7B	InternLM	Difficulty
Llama-3.1-8B-Instruct (Dubey et al., 2024)	8B	Llama 3.1	Difficulty
Llama-2-13b-chat-hf (Touvron et al., 2023a)	13B	Llama 2	Difficulty
Mistral-Nemo-Instruct-2407 (Jiang et al., 2023)	2407	Mistral	Difficulty
MiniCPM-1B-sft-bf16 (Hu et al., 2024)	1B	MiniCPM	Difficulty
XVERSE-65B-Chat[6]	65B	XVERSE	Difficulty
AquilaChat2-34B (Zhang et al., 2024)	34B	Aquila 2	Difficulty

Table 4: The list of evaluated LLMs.