Does Question Really Matter? The Attribution of Answer Bias in LLM Evaluation

Boxi Cao, Ruotong Pan, Hongyu Lin, Xianpei Han*, Le Sun*

Chinese Information Processing Laboratory, Institute of Software
Chinese Academy of Sciences, Beijing, China
{caoboxi,hongyu,xianpei,sunle}@iscas.ac.cn

Abstract

Multiple-choices question answering (MCQA) has emerged as one of the most popular task formats for large language models (LLMs) evaluation. Unfortunately, there exist substantial evidence that the evaluation of current MCQA benchmarks suffers from significant *answer bias*, which severely undermines the reliability of the evaluation conclusions. Specifically, many LLMs achieve performance significantly higher than random selection even when the questions are omitted from input information. To this end, we conduct a systematic investigation of the attribution of answer bias, and demonstrate a strong correlation between the degree of data contamination and the severity of answer bias, while the position of options and the popularity of answers have relatively minor effects. Building on these insights, we further propose OPD, a straightforward yet effective tool for contamination detection and dataset debiasing without requiring access to the model's internal training data. Our findings and algorithms provide valuable insights for the design of future trustworthy LLM evaluation protocols.

1 Introduction

The evaluation of large language models (LLMs) has become a cornerstone of their development (Ouyang et al. 2022; Touvron et al. 2023; OpenAI et al. 2023). Therefore, enormous benchmarks have been proposed to comprehensively assess the capabilities of LLMs (Chang et al. 2023). Among them, multiple-choices question answering (MCQA) has emerged as the most commonly employed task format, as it is easy to collect, relatively objective and capable of automatic evaluation. Representative benchmarks such as MMLU (Hendrycks et al. 2021), GPQA (Rein et al. 2023), ARC (Clark et al. 2018) have been extensively utilized to gauge the development of LLMs. Ideally, given a specific question, LLM should first comprehend the information within the context, extract relevant internal knowledge and ultimately select the correct answer from several candidate options. The design, difficulty and scope of questions are crucial determinants of a benchmark's overall quality, and the evaluation performance is expected to robustly and accurately reflect the capabilities of LLMs.

* Corresponding Authors

However, previous studies have identified the presence of answer bias in the evaluation of language models, a phenomenon wherein the models are capable of producing correct answers even in the absence of the given questions (Poliak et al. 2018; Geirhos et al. 2020; Balepur, Ravichander, and Rudinger 2024). Buildings upon these observations, as illustrated in Figure 1, we first conduct a comprehensive option-only evaluation regarding answer bias spanning 6 widely utilized MCQA benchmarks and 17 open-sourced LLMs. The results demonstrate that there exist a widespread issue of answer bias on current MCQA benchmarks for LLMs. Specifically, even when the question is removed from the input prompt, retaining only the candidate options, many LLMs are still able to achieve considerably high performance without even looking at the actual question. For instance, on MMLU (Hendrycks et al. 2021), almost all models manage to achieve over half of their original performance without accessing the question; on HellaSwag (Zellers et al. 2019), there are even LLMs achieve an accuracy rate exceeding 75% when presented with only 4 options. Such phenomenon raises serious concerns about whether the original superior performance achieved by LLMs are due to genuine knowledge or largely affected by other confounders. Furthermore, our extensive experiments across different model families, model scales, and benchmarks reveal more fine-grained findings that were not addressed in previous studies. For example, we find that while MiniCPM3-4b's original performance on CMMLU is significantly lower than that of LLaMA-3-70b, its option-only performance surpasses models that are more than ten times larger. Additionally, the LLaMA-3 series exhibit markedly different behaviors between English and Chinese benchmarks. The aforementioned observations prompt us to delve into the underlying causes of such answer bias, and strive for more reliable and impartial evaluation results.

To this end, we conduct a systematic investigation of the attribution of answer bias. Through interventions on instances and correlation analyses with training data, we investigate the 3 most likely contributing factors: answer popularity, selection bias, and data contamination. The experimental results demonstrate that there exists a high correlation between the degree of data contamination and the severity of answer bias, whereas the position of options and the popularity of answers have relatively minor effects. To further vali-

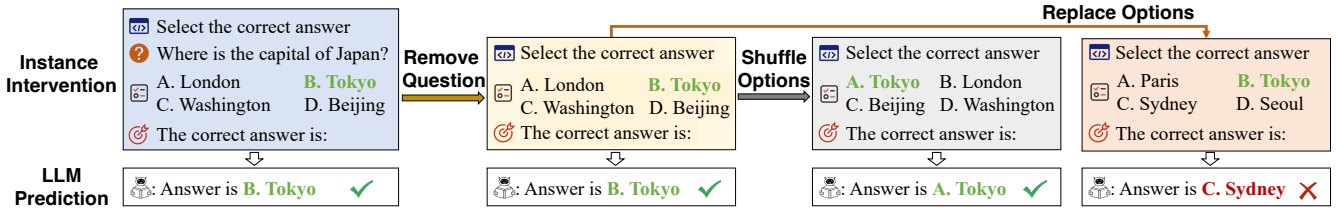

Figure 1: Illustration of the option-only evaluation, which first removes the question from input, and then assesses whether LLMs can still select the correct answer from candidate options.

date the above findings, we propose OPD, a straightforward algorithm based on option-only evaluation for data contamination detection without requiring access to the model's internal training data. Experimental results demonstrate that OPD serves as an effective tool for detecting which test cases have been contaminated. Building on these insights, to mitigate the effects of answer bias and thereby enhance the reliability of existing benchmarks without additional human efforts, we further propose OPD-based benchmark pruning, which filters out the instances that can be easily answered without question. Evaluation results indicate that the performance of current LLMs significantly declines on pruned datasets, highlighting the necessity of revising these benchmarks for more accurate assessments. The major contributions of this paper are summarized as follows:

- We conduct a systematic investigation regarding the existence and causes of answer bias.
- Based on the analysis, we introduce effective tools for contamination detection and dataset debiasing.

2 Widespread Answer Bias

To further verify the existence of answer bias across different LLMs and benchmarks, as illustrated in Figure 1, we conduct option-only evaluation on a wide range LLMs and benchmarks. In the following, we will first describe our experimental settings in detail, and then demonstrate that there exists severe answer bias which affects the reliability of current evaluation performance.

2.1 Experimental Setup

Benchmarks. To ensure the broad applicability of our experimental conclusions, we select 6 representative benchmarks in the form of multiple-choice questions. 1) MMLU (Hendrycks et al. 2021) is a large-scale benchmark designed to comprehensively measure the knowledge within LMs, which covers 57 subjects and ranges in difficulty from elementary to professional level. 2) CMMLU (Li et al. 2023) is a large-scale benchmark aiming to assess the knowledge and reasoning abilities of LLMs in Chinese, which comprises 67 subjects. 3) CEVAL (Huang et al. 2023) is a Chinese evaluation benchmark which consists of 52 diverse disciplines and 4 difficulty levels. We use the dev set since test set is not openly released. 4) ARC (Clark et al. 2018) consists of a set of questions of science exams from grade 3 to grade 9, which is also often used to assess the knowledge capabilities of LMs. We select the challenge subset for experiments which contains more difficult questions. 5) OpenbookQA (Mihaylov et al. 2018) contains elementary-level science questions, which requires both scientific and commonsense knowledge for answering. 6) HellaSwag (Zellers et al. 2019) is designed to evaluate the commonsense natural language inference abilities of LMs.

Models. We conduct experiments on 17 open-source LLMs ranging in size from 4B to 72B parameters: LLaMA-2-7B (Touvron et al. 2023), LLaMA-3-8B & 70B, LLaMA-3.1-8B (Dubey et al. 2024), Mistral-7B & 8*7B (Jiang et al. 2023), Qwen-1.5-7B (Bai et al. 2023), Qwen-2-7B & 72B (Yang et al. 2024), Yi-6B, Yi-1.5-9B & 34B (Young et al. 2024), Baichuan2-7B (Yang et al. 2023), Deepseek-V2-16B (Liu et al. 2024), InternLM-2-7B, InternLM-2.5-7B (Team 2023), MiniCPM-3-4B (Hu et al. 2024).

Option-only Evaluation. For each benchmark, we perform two evaluations. The first is a standard evaluation, where we provide the model with an instruction for question answering, the question itself, and the corresponding candidate options, requiring the model to select the correct answer from these options. The second is an option-only evaluation, where the only difference is that we remove the question from the input information, directly prompting the model to select a "correct answer". To avoid the impact of input prompts on the performance across different benchmarks, we employ a standardized task prompt format for all benchmarks. To obtain more stable evaluation results and to avoid out-of-answer issues, we employ a zero-shot perplexity-based assessment method. All the evaluation is implemented based on the standard evaluation from OpenCompass (Contributors 2023). Please kindly note that our experiments currently do not include chat models since we find that they may learn to reject inputs that do not contain a question during the alignment process. Therefore, we believe that the evaluation results for the pre-trained LMs provide a more accurate reflection of answer bias.

2.2 Overall Results

The comparison of performance under the standard evaluation and option-only evaluation of 17 LLMs across each benchmark is demonstrated in Figure 2. We can clearly find that **there exist widespread answer biases in current LLM evaluation, which significantly undermines the accuracy and reliability of evaluation conclusions**:

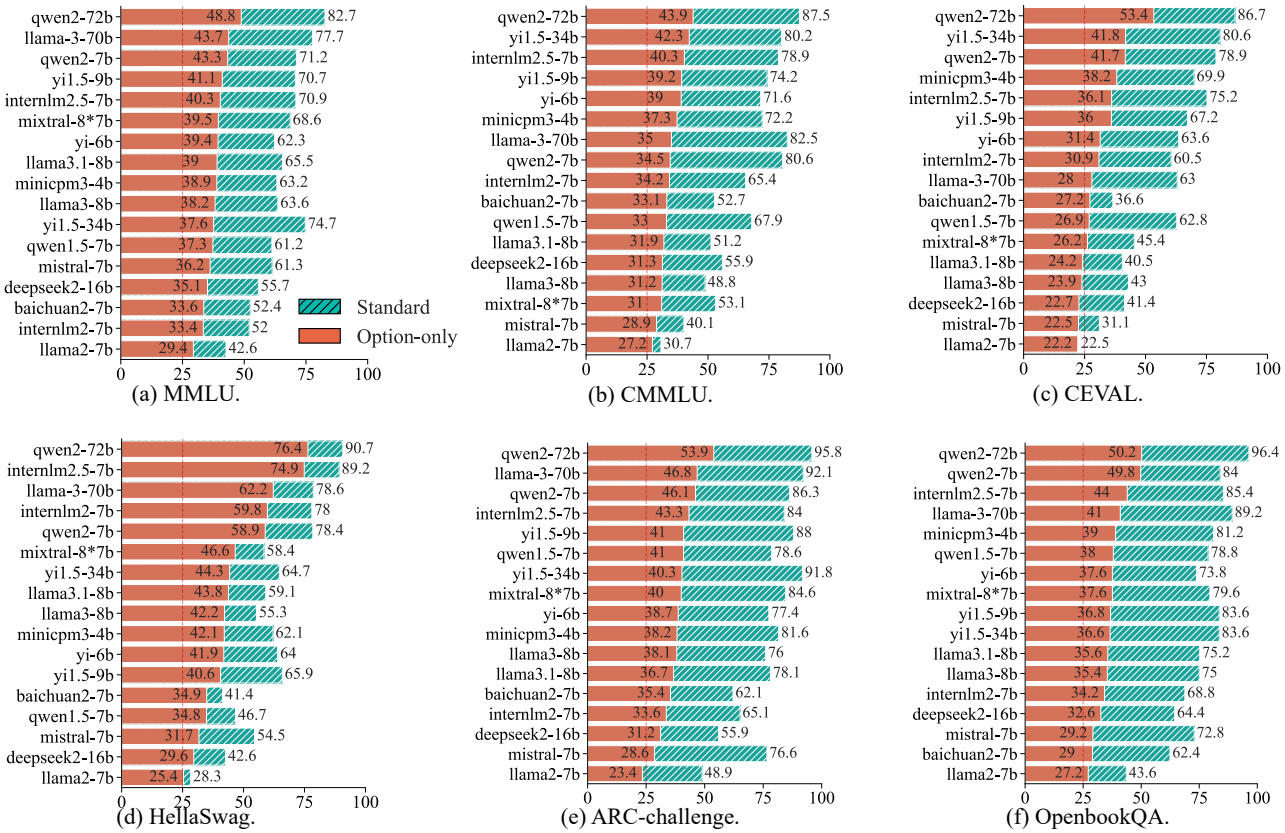

Figure 2: Comparison of performance under the original evaluation and option-only evaluation. The results are sorted in decreasing option-only performance. The red dashed line represents the performance level achieved by random selection. Many LLMs achieve performance significantly higher than random selection when question is removed from the input.

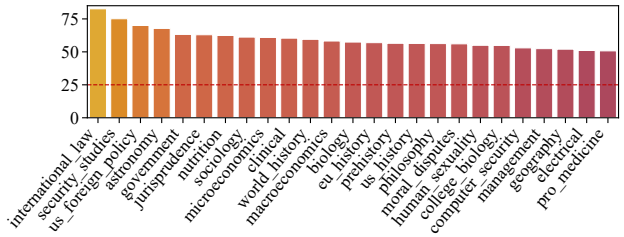

Figure 3: Option-only performance of Qwen2-72B on various subjects in MMLU benchmark.

1) Many LLMs can still achieve considerably high performance even when removing the question from input and only maintain the candidate options, which present concrete evidence for the prevalence of answer bias. Across all benchmarks, the vast majority of models significantly outperform the performance expected from random selection in option-only evaluations. Moreover, the phenomenon of answer bias is even more pronounced among certain LLMs. For instance, Qwen2-72B can achieve over 50% accuracy on 4 benchmarks with only options provided; InternLM-2.5-7B can achieve an option-only performance up to 74.9% on HellaSwag, compared to its standard performance 89.2%; While MiniCPM3-4b's original performance on CMMLU is significantly lower than that of LLaMA-3-70b, its option-only performance surpasses models that are more than ten times larger. 2) Upon further investigation of the models' performance across different subjects, we find that the impact of answer bias becomes more pronounced. Figure 3 demonstrate the option-only performance of Qwen2-72B across various subjects on MMLU. Qwen2-72B is capable of achieving performance exceeding 50% on more than half of the subjects. Meanwhile, in certain subjects such as algebra and logic questions, where all options are highly similar (e.g., true/false, numbers), all models tend to achieve performance that is close to random selection. 3) The correlation between test data and LLMs' pre-training data can affect the models' option-only performance. For instance, there exists a notable performance disparity for LLaMA-3 between English benchmarks (e.g., MMLU, ARC) and Chinese benchmarks(e.g., CMMLU, CEVAL). This phenomenon suggests that the presence of answer bias may be correlated to the distribution of the model's training data, which we will further analyze in subsequent sections.

3 Attribution of Answer Bias

The experiments in Section 2 clearly demonstrate the widespread prevalence of answer bias and its significant impact on current LLM evaluations. In these cases, it is

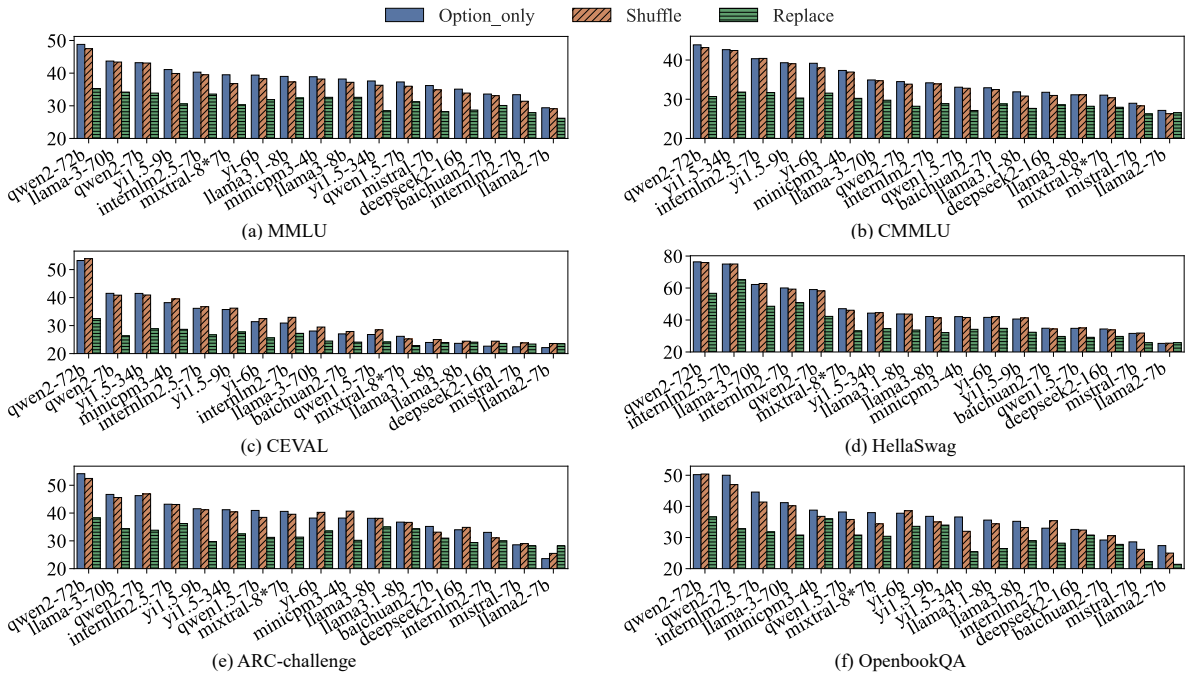

Figure 4: The performance of LLMs is compared under three different testing conditions: 1) Option-only: in this setting, the question is removed from the input while retaining the original options and their sequence. 2) Shuffle: based on the option-only setting, the options are shuffled to analyze the impact of selection bias. 3) Replace: building on the option-only setting, the "correct" option is preserved, while the three "incorrect" options are randomly replaced with incorrect options from other questions within the same dataset.

essential to dive into the underlying causes of this phenomenon. Understanding the key attributions of answer bias will enable the development of targeted strategies to mitigate the impact of answer bias on the reliability of evaluations. Therefore, we conduct a systematic investigation of the attribution of answer bias. Specifically, we investigate 3 most likely contributing factors including selection bias, answer popularity, and data contamination. By analyzing the correlation between answer bias and each factor, we identify that data contamination as the primary factor contributing to answer bias. To further substantiate such findings, we propose OPD, an algorithm for data contamination detection based on option-only evaluation, and experimental results demonstrate that OPD can identify contaminated test instances with a high F1 score compared to previous methods.

3.1 Does Selection Bias Affect?

The initial possible influencing factor is selection bias, referring to the preference of LLMs for options in specific positions, which has been observed in the evaluation of LLMs on MCQA benchmarks (Zheng et al. 2023; Wei et al. 2024). Consequently, we examine whether answer bias arises from the alignment of a model's selection bias with the placement of correct answers. Specifically, as illustrated in Figure 1, we perturb the original evaluation instances in a straightforward manner. For each test instance, we randomly shuffle the order of all options and then assess whether the model could still select the "correct" answer without being provided with the question. By comparing the performance of the option-only evaluation before and after shuffling, we can assess the impact of selection bias.

The results are demonstrated in Figure 4, which show that the model's performance under option-only evaluation does not exhibit significant changes after the options are shuffled. For instance, on ARC-challenge, the average performance across all models shifts minimally from 38.37% to 38.19% after shuffling the options, and similar trends are observed in other benchmarks as well. For the vast majority of models and datasets, the effect of shuffling options on performance did not exceed 0.1%. To further quantify the performance divergence, we employ Mann–Whitney U test (Mann and Whitney 1947) to examine the significance of the performance before and after shuffling options. Following standard conventions, we set the threshold for statistical significance at 0.05. The p-values of all benchmarks are presented in Table 1. In each of these benchmarks, the p-values are notably greater than 0.05, indicating insufficient statistical evidence to assert a noteworthy difference in model performance before and after option shuffling. The above experiment results suggest that selection bias is not the primary cause of answer bias.

3.2 Does Answer Population Affect?

The second potential influencing factor is answer popularity, which refers to the extent to which the correct answer itself is favored among all available options for a specific model.

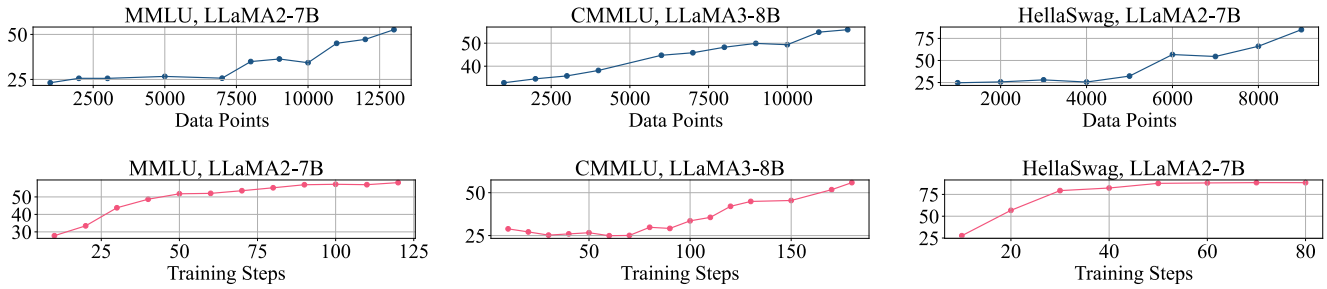

Figure 5: The performance under option-only evaluation improves as the degree of data contamination increases. The base models are selected since their original option-only performance is closely to random selection.

MMLU	CMMLU	CEVAL	HellaSwag	ARC	OBQA
0.21	0.31	0.26	0.47	0.45	0.20

Table 1: P value of U-test of the option-only evaluation performance between and after shuffling options.

Factor	MMLU	CMMLU	HellaSwag
Data Points	0.935	0.991	0.925
Training Steps	0.887	0.915	0.822

Table 2: The correlation coefficient between degree of contamination and performance of option-only evaluation.

Therefore, we aim to investigate whether answer bias arises because the correct answer is often the more popular choice, and as a result, the one that the model tends to select without question. However, due to such answer popularity is correlated with the training data of specific model and the black-box nature of current LLMs' training data, it is impractical to accurately quantify the popularity of each option. To determine the impact of answer popularity on model predictions, we substitute the original incorrect options in each instance with other options of the same type, then assess whether the model can still select the original answer from the set of new candidate options. The underlying hypothesis is intuitive: if the model's choice relies answer popularity, it should continue to select the originally more "popular" answer even after the options have been replaced. As shown in Figure 1, to keep the option distribution consistent, for each subject within a benchmark, we will first divide all options from every instance into two groups: the first group consists of all the correct answers, while the second group comprises all the incorrect options. Then, each "original correct" answer will be paired with three randomly selected incorrect options to form new test samples. By comparing the performance of option-only evaluation before and after replacing the options, we can assess whether LLMs tend to select answers based on popularity.

The results are presented in Figure 4, revealing a significant decline in performance after the replacement of options. For instance, the performance of Qwen2-72B under option-only evaluation decreases from 48.8% to 35.3% on MMLU, from 53.2% to 32.5% on CEVAL and from 43.9% to 30.7% on CMMLU. Nonetheless, we can also observe that the performance of most models maintains higher than random selection after substitution of options. Therefore, based on the above analysis, we posit that answer popularity is one of the factors contributing to answer bias. However, given the significant decline performance after the substitution of options, it is evident that there are other, more predominant influencing factors.

3.3 Does Data Contamination Affect?

Combined with the experimental results in section 3.1 and 3.2, we find that, in most cases, the model can only choose the correct answer corresponding to the original instance when provided with the initial 4 options. This phenomenon leads us to speculate that data contamination could be one of the primary factors attributing answer bias. In other words, the model may have memorized specific combinations of options present in the training data and selected the corresponding answers from the original instances. However, due to the vast scale and opacity of current LLMs' pre-training data, it's presently impractical to precisely identify which test cases are contaminated (Oren et al. 2024; Zhou et al. 2023; Palavalli, Bertsch, and Gormley 2024). Therefore, to investigate the impact of data contamination, we refer to the experimental setups in previous studies (Dong et al. 2024; Shi et al. 2024; Zhou et al. 2023; Oren et al. 2024; Ying et al. 2024; Xu et al. 2024). We first train a base model with test data, and then calculate the correlation coefficient between the degree of contamination and the severity of answer bias. To avoid the influence of other spurious correlations on the experimental results, we further conduct comparative experiments on data with the same distribution. Furthermore, we propose a data contamination detection algorithm based on option-only evaluation to further verify the interrelationship between data contamination and answer bias.

Correlation with Data Contamination To conduct a comprehensive analysis of the impact of data contamination on answer bias, we calculate the correlation coefficients between the degree of contamination and the severity of answer bias from two dimensions: the number of contaminated samples and the number of training steps. For con-

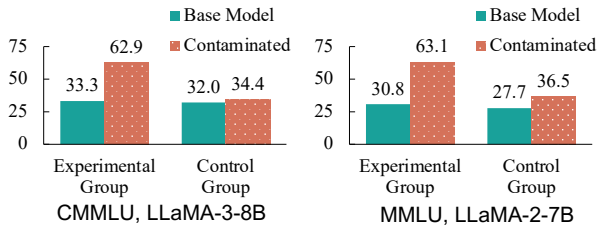

Figure 6: Option-only performance improvement within 2 groups of data: the experimental group comprises of data from 15 contaminated subjects, and the control group consists of data from uncontaminated 15 subjects within the same benchmark.

taminated samples: for a dataset comprising N samples, we train $K = \lfloor \frac{N}{1000} \rfloor$ models, where the $k-th$ model is trained with $1000 * k$ randomly sampled instances. Subsequently, we assess the performance of each of these K models under option-only evaluation. For training steps, we train the model on the entire dataset, save a checkpoint at every 10 steps, and assess the option-only performance for each saved checkpoint. For the selection of data and models, we conduct experiments on datasets with a larger amount of instances: MMLU, CMMLU, and HellaSwag. Additionally, we select base models with option-only performance close to random selection. Each model is trained with batch size of 256 sequences for 4 epoch with learning rate $2e-5$.

Figure 5 demonstrates how model performance varies with the degree of data contamination, and Table 2 demonstrates the Pearson correlation coefficient between these two factors. It's obvious that **there exists a strong correlation between the degree of contamination and the performance under option-only evaluation**: On all 3 benchmarks, the model's option-only performance exhibits a correlation coefficient exceeding 0.9 with amount of contaminated instances and a correlation coefficient exceeding 0.8 with number of training steps. Additionally, the final performance on these 3 benchmarks achieved through data contamination closely approximated the highest LLM performance demonstrated in Figure 2.

However, aside from data contamination, there could be other spurious correlations contributing to such performance improvement. To better disentangle the effects of data contamination, we select 30 subjects from MMLU and CMMLU respectively, randomly sampling 15 subjects as the experimental group and assigning the remaining 15 as the control group. Subsequently, we train a base LLM using data from the experimental group, observing changes in option-only performance in both groups. In this experiment, if there are other major influencing factors beyond data contamination, such as the model learning specific shortcuts, due to the similar data distribution between the two groups, we should be able to observe a significant performance improvement in both groups. The comparison is demonstrated in Figure 6, we can observe that the improvement in option-only performance of the model is significantly greater in the experimental group (contaminated) compared to the control group

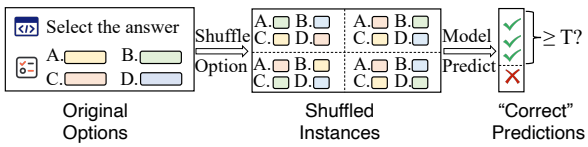

Figure 7: Illustration of our contamination detection algorithm OPD.

Method	R	P	F
BLD (Ni et al. 2024)			
Scenario A	0.040	0.250	0.069
Scenario B ($\sigma = 0.15$)	0.600	0.522	0.558
Scenario B ($\sigma = 0.17$)	0.440	0.524	0.478
Scenario B ($\sigma = 0.20$)	0.220	0.537	0.312
Our OPD			
Option-only	0.785	0.730	0.757
$T = 2$	**0.847**	0.764	**0.802**
$T = 3$	0.766	0.830	0.797
$T = 4$	0.583	**0.921**	0.714

Table 3: Recall, precision and F1 of data contamination detection methods. For baseline, we report results under different thresholds as referenced in the original paper.

(clean), which demonstrate **it is primarily data contamination instead of other spurious correlations contributing to the option-only performance improvement.**

Data Contamination Detection Considering the strong correlation between data contamination and answer bias, we propose an intuitive algorithm named *OPD* for data contamination detection based on option-only evaluation. As illustrated in Figure 7, to mitigate the influence of random factors, for each instance in original benchmark, we first shuffle the options 4 times to create 4 new test samples. We then evaluate the model's prediction results under option-only evaluation for each sample. If the model correctly predicts the original answer at least T times, we label the instance as contaminated. To verify the effectiveness of OPD, we reference the experimental setup of previous studies by employing llama-2-7B as the base model and sampling 5,000 data points from HellaSwag as contaminated set for model training. The model is trained with batch size of 256 sequences for 3 epochs, using Adam with learning rate $2e-5$. Then, we sample 5000 instances from MMLU as a clean set. Subsequently, we employ OPD to detect which 5,000 instances among the 10,000 total instances have been contaminated.

Table 3 demonstrates the performance comparison between our OPD algorithm and the detection method proposed by Ni et al. (2024) . We can see that, compared to the baseline, the OPD method achieves substantial performance improvements in both recall and precision, effectively detecting contaminated test data. OPD results in an increase of 24.4% in F1 compared to the optimal configuration of baseline. Moreover, the precision of OPD is further enhanced as the value of T increases. **This experiment not only demonstrates the effectiveness of our proposed data contamination detection method but also further substantiates the**

Model	MMLU-S	CMMLU-S	HellaSwag-S	ARC-S	OBQA-S
LLaMA-2-7B	$25.02_{17.60\downarrow}$	$23.94_{06.80\downarrow}$	$25.51_{02.77\downarrow}$	$28.04_{20.89\downarrow}$	$27.60_{16.00\downarrow}$
LLaMA-3-8B	$47.53_{16.11\downarrow}$	$36.43_{12.39\downarrow}$	$31.45_{23.81\downarrow}$	$65.87_{10.18\downarrow}$	$59.45_{15.55\downarrow}$
LLaMA-3.1-8B	$50.54_{14.94\downarrow}$	$40.00_{11.16\downarrow}$	$34.84_{24.29\downarrow}$	$68.48_{09.63\downarrow}$	$64.71_{10.49\downarrow}$
LLaMA-3-70B	$65.50_{12.16\downarrow}$	$53.44_{29.06\downarrow}$	$49.18_{29.43\downarrow}$	$87.61_{04.49\downarrow}$	$80.77_{08.43\downarrow}$
Mistral-7B	$42.86_{18.47\downarrow}$	$31.05_{09.03\downarrow}$	$30.64_{23.89\downarrow}$	$60.22_{16.43\downarrow}$	$57.35_{15.45\downarrow}$
Mixtral-8*7B	$54.00_{14.64\downarrow}$	$38.56_{14.56\downarrow}$	$40.88_{17.52\downarrow}$	$74.78_{09.77\downarrow}$	$66.52_{13.08\downarrow}$
Qwen-1.5-7B	$42.69_{18.51\downarrow}$	$56.68_{11.26\downarrow}$	$29.00_{17.71\downarrow}$	$67.83_{10.80\downarrow}$	$61.88_{16.92\downarrow}$
Qwen-2-7B	$53.24_{17.91\downarrow}$	$70.70_{09.86\downarrow}$	$33.09_{45.34\downarrow}$	$75.43_{10.84\downarrow}$	$76.24_{07.76\downarrow}$
Qwen-2-72B	$70.17_{12.53\downarrow}$	$82.57_{04.93\downarrow}$	$59.53_{31.20\downarrow}$	$92.61_{03.18\downarrow}$	$91.12_{05.28\downarrow}$
Yi-6B	$45.65_{16.64\downarrow}$	$60.50_{11.11\downarrow}$	$34.94_{29.10\downarrow}$	$68.26_{09.16\downarrow}$	$58.37_{15.43\downarrow}$
Yi-1.5-9B	$55.71_{15.03\downarrow}$	$60.40_{13.83\downarrow}$	$38.01_{27.87\downarrow}$	$79.13_{08.85\downarrow}$	$71.72_{11.88\downarrow}$
Yi-1.5-34B	$61.91_{12.74\downarrow}$	$71.30_{08.92\downarrow}$	$41.91_{22.77\downarrow}$	$88.04_{03.72\downarrow}$	$78.96_{04.64\downarrow}$
Deepseek-V2-16B	$38.12_{17.57\downarrow}$	$45.15_{10.72\downarrow}$	$29.20_{13.41\downarrow}$	$47.83_{08.11\downarrow}$	$47.85_{16.55\downarrow}$
InternLM-2-7B	$39.72_{12.25\downarrow}$	$51.65_{13.75\downarrow}$	$55.84_{22.12\downarrow}$	$58.04_{07.02\downarrow}$	$72.40_{03.60\uparrow}$
InterLM-2.5-7B	$53.65_{17.25\downarrow}$	$66.04_{12.83\downarrow}$	$62.50_{26.71\downarrow}$	$76.96_{06.99\downarrow}$	$83.26_{02.14\downarrow}$
MiniCPM-3-4B	$45.63_{17.57\downarrow}$	$57.18_{15.04\downarrow}$	$32.79_{29.35\downarrow}$	$72.83_{08.80\downarrow}$	$67.76_{13.44\downarrow}$
Baichuan-2-7B	$33.36_{19.00\downarrow}$	$38.48_{14.24\downarrow}$	$24.90_{16.45\downarrow}$	$41.09_{20.97\downarrow}$	$40.55_{21.85\downarrow}$

Table 4: The evaluation performance on pruned benchmarks based on OPD. Subscript numbers show performance changes relative to the original benchmark. CEVAL isn't included due to an insufficient amount of remaining data.

significant impact of data contamination on answer bias. These results align with our previous findings. OPD can be used to test data contamination, as we have shown that data contamination—not selection bias—is the main cause of answer bias. Even after shuffling the options multiple times, a contaminated instance may still cause the model to choose the correct answer.

4 OPD-based Benchmark Pruning

The above experiments demonstrate that it is critical to optimize the existing evaluation datasets to mitigate the impact of answer bias, and achieve more equitable assessment outcomes. To this end, we prune the existing benchmarks based on our proposed OPD algorithm, which can mitigate the influence of answer bias without the need for additional human annotation. The idea behind this is intuitive, if a test case can be consistently answered correctly even when the question is removed, its value for evaluation is significantly diminished. Specifically, we place all the models to be tested into a model pool. For each test case in a benchmark, we first use the OPD algorithm to test each model in the pool, and then remove all test cases labeled as low-value by the OPD. Based on the experimental results in Table 3, we set T to 2 to yield the highest F1 score.

Table 4 demonstrates the performance of each model on our pruned benchmarks, as well as the performance divergence compared with the performance on original benchmarks. We can see that all the models exhibit a significant decrease in performance when evaluated on the pruned benchmark. For instance, on the HellaSwag dataset, which previously exhibited the most severe answer bias, the average performance of all models decreased by 22.77% after data pruning. Such results further highlight the impact of answer bias on LLM evaluation and the necessity of revising current benchmarks for more accurate assessment. Meanwhile, we observe that larger models such as LLaMa-3-70B and Qwen-2-72B, continue to achieve superior performance on pruned benchmarks. This indicates that despite the influence of answer bias on previous evaluation results, these large models do indeed possess greater capabilities. Consequently, there is a necessity for updated, high-quality, and more challenging benchmarks to provide a more reliable evaluation of their capabilities.

5 Related Works

The investigation and mitigation of bias in the evaluation of models are crucial for ensuring their fairness, reliability and applicability, which has consistently garnered extensive attention with the advancement of machine learning (Geirhos et al. 2020; Liang et al. 2022; Gallegos et al. 2024; Alzahrani et al. 2024). Many early studies focus on analyzing the annotation artifacts (Gururangan et al. 2018; Kaushik and Lipton 2018) and syntactic heuristics (Poliak et al. 2018; McCoy 2019) in datasets related to tasks such as NLI. When it comes to LLM evaluation, researchers have identified various types of biases present in different evaluation frameworks. For instance, prompt bias in cloze-style evaluation (Zhao et al. 2021; Cao et al. 2022), selection bias in MCQA benchmarks (Zheng et al. 2023; Pezeshkpour and Hruschka 2023), position bias and knowledge bias in LLM-as-judge evaluation (Koo et al. 2023; Wang et al. 2023). These studies expose serious deficiencies in current LLM evaluation, thereby advancing the development of more reliable evaluation paradigms.

6 Conclusions

We show that answer bias is widespread in MCQA-based LLM evaluation and that data contamination is its primary cause. To address this issue, we propose OPD, an option-only based method for contamination detection and benchmark pruning without accessing training data. Our results underscore the necessity of revising current benchmarks toward more reliable LLM evaluation.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by Beijing Natural Science Foundation (L243006), the Natural Science Foundation of China (No. 62476265), the Basic Research Program of ISCAS (Grant No. ISCAS-JCZD-202303).

References

Alzahrani, N.; Alyahya, H.; Alnumay, Y.; Alrashed, S.; Alsubaie, S.; Almushayqih, Y.; Mirza, F.; Alotaibi, N.; Al-Twairesh, N.; Alowisheq, A.; et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13787–13805.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Balepur, N.; Ravichander, A.; and Rudinger, R. 2024. Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? *arXiv preprint arXiv:2402.12483*.

Cao, B.; Lin, H.; Han, X.; Liu, F.; and Sun, L. 2022. Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5796–5808. Dublin, Ireland: Association for Computational Linguistics.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2023. A Survey on Evaluation of Large Language Models. arXiv:2307.03109.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

Dong, Y.; Jiang, X.; Liu, H.; Jin, Z.; Gu, B.; Yang, M.; and Li, G. 2024. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 12039–12050. Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.

Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. arXiv:2305.08322.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kaushik, D.; and Lipton, Z. C. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Koo, R.; Lee, M.; Raheja, V.; Park, J. I.; Kim, Z. M.; and Kang, D. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

McCoy, R. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv preprint arXiv:1902.01007*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, 2381–2391. Brussels, Belgium: Association for Computational Linguistics.

Ni, S.; Kong, X.; Li, C.; Hu, X.; Xu, R.; Zhu, J.; and Yang, M. 2024. Training on the Benchmark Is Not All You Need. *arXiv preprint arXiv:2409.01790*.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt; et al. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Oren, Y.; Meister, N.; Chatterji, N. S.; Ladhak, F.; and Hashimoto, T. 2024. Proving Test Set Contamination in Black-Box Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155.

Palavalli, M.; Bertsch, A.; and Gormley, M. R. 2024. A Taxonomy for Data Contamination in Large Language Models. *CoRR*, abs/2407.08716.

Pezeshkpour, P.; and Hruschka, E. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Shi, W.; Ajith, A.; Xia, M.; Huang, Y.; Liu, D.; Blevins, T.; Chen, D.; and Zettlemoyer, L. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wei, S.-L.; Wu, C.-K.; Huang, H.-H.; and Chen, H.-H. 2024. Unveiling Selection Biases: Exploring Order and Token Sensitivity in Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 5598–5621.

Xu, R.; Wang, Z.; Fan, R.; and Liu, P. 2024. Benchmarking Benchmark Leakage in Large Language Models. *CoRR*, abs/2404.18824.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ying, J.; Cao, Y.; Bai, Y.; Sun, Q.; Wang, B.; Tang, W.; Ding, Z.; Yang, Y.; Huang, X.; and Yan, S. 2024. Automating Dataset Updates Towards Reliable and Timely Evaluation of Large Language Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, 12697–12706. PMLR.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Zhou, K.; Zhu, Y.; Chen, Z.; Chen, W.; Zhao, W. X.; Chen, X.; Lin, Y.; Wen, J.; and Han, J. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. *CoRR*, abs/2311.01964.