LLMs May Perform MCQA by Selecting the Least Incorrect Option

Haochun Wang, Sendong Zhao*, Zewen Qiang, Nuwa Xi, Bing Qin and Ting Liu
Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, China
{hcwang,sdzhao}@ir.hit.edu.cn

Abstract

In the field of NLP, Large Language Models (LLMs) have markedly enhanced performance across a variety of tasks. However, the comprehensive evaluation of LLMs remains an inevitable challenge for the community. Recently, the adoption of Multiple Choice Question Answering (MCQA) as a benchmark for assessing LLMs has gained considerable traction. However, concerns regarding the robustness of this evaluative method persist. Building upon previous discussions on the issue of *variability*, we reveal an additional dimension of concern: LLMs may perform MCQA by selecting the least incorrect option rather than distinctly correct. This observation suggests that LLMs might regard multiple options as correct, which could undermine the reliability of MCQA as a metric for evaluating LLMs. To address this challenge, we introduce an enhanced dataset augmentation method for MCQA, termed MCQA+, to provide a more accurate reflection of the model performance, thereby highlighting the necessity for more sophisticated evaluation mechanisms in the assessment of LLM capabilities.

1 Introduction

The emergence of Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023a), and ChatGPT (OpenAI, 2022), represents a paradigm shift in the field of Natural Language Processing (NLP). These models have exhibited exceptional proficiency in mimicking human-like textual outputs, establishing their significance across various applications. However, the challenge of effectively evaluating LLMs persists (Chang et al., 2023). This difficulty arises from the intricate nature of natural language. Conventional evaluation metrics for generative tasks often fall short in accurately assessing the performance of LLMs, since most LLMs can generate text contextually rich and coherent (Thoppilan et al., 2022), complicating the assessment of the outputs through merely quantitative evaluation based on text matching such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).

Multiple-Choice Question Answering (MCQA) is a fundamental format for various tasks in NLP, such as commonsense reasoning (Talmor et al., 2019; Sap et al., 2019; Zellers et al., 2018), reading comprehension (Lai et al., 2017; Huang et al., 2019) and cloze-style tasks (Zellers et al., 2019; Mostafazadeh et al., 2016). Each MCQA instance comprises a question paired with several answer options, requiring models to identify the correct response as depicted in Figure 1. As a non-subjective metric, MCQA serves as a prominent automatic evaluation method with accuracy as an evaluation metric for numerous LLMs to test for the commonsense knowledge or knowledge for specific domain (Gao et al., 2021; Touvron et al., 2023a; OpenAI et al., 2023).

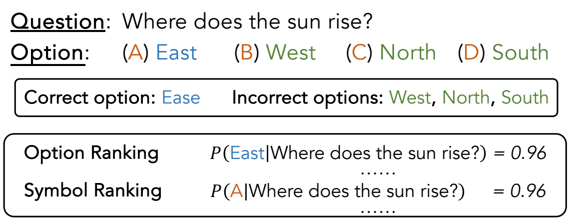

Figure 1: An MCQA example and ranking strategies.

Despite the advanced performance of LLMs on the accuracy of MCQA-format benchmarks like MMLU (Hendrycks et al., 2021), previous studies have discussed a key challenge that persists in evaluating LLMs is maintaining *invariability* in responses when confronted with different orders of answer choices for a same question (Robinson and Wingate, 2022; Wang et al., 2023; Zheng et al., 2023), which underscores an issue that the accuracy of MCQA-format tasks may not reflect the

* Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 5852–5862
January 19–24, 2025.

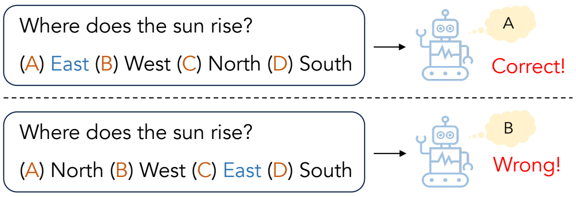

Figure 2: A case for variability issue of LLMs.

authentic capability of LLMs as an example in Figure 2. However, the above phenomenon may not be the *only* issue in the evaluation of LLMs with MCQA-format questions.

To eliminate the potential impact of variability in model responses, we begin by filtering a dataset, denoted as $\mathcal{D}$, to extract a subset $\mathcal{D}^{\blacklozenge}$, which contains instances where the LLMs can consistently predict the correct answer across all permutations of the answer options, thereby demonstrating invariability. Following this, we conduct a comprehensive experimental analysis using various configurations derived from the original MCQs in $\mathcal{D}^{\blacklozenge}$. Our experimental results indicate that while LLMs often select the *most* correct answer, they may also regard other options as correct to some extent. Consequently, evaluating LLM performance solely based on MCQA can produce ambiguous results. This newly identified issue prompts a reconsideration of the suitability of MCQA as a reliable metric for LLM evaluation and offers a possible explanation for the observed differences in LLM performance on generative versus discriminative tasks (West et al., 2024).

To address this issue, which is inherently difficult to resolve, we propose an augmentation of the MCQA dataset, termed MCQA+, which introduces variations of the original MCQs and is designed to more accurately reflect LLM capabilities. Empirical findings demonstrate that LLM performance on the MCQA+ dataset is significantly lower than on the original MCQA dataset, indicating that MCQA+ can serve as a more effective benchmark for developing robust and adaptable NLP models. This augmentation may ultimately contribute to narrowing the gap between machine learning models and human-like understanding and reasoning in NLP tasks.

In summary, our contributions are as follows:

- We identify a novel issue with MCQA-based evaluation of LLMs, beyond the variability in answer options, where LLMs may approach MCQA by selecting the option that is “least incorrect”.
- This issue implies that while LLMs consistently select the correct answer for specific MCQs, they may also incorrectly identify certain other options as correct under different circumstances.
- We introduce a dataset augmentation method, expanding the original MCQA into MCQA+, which more accurately reveals LLM capacities and performance.

2 Related Work

LLMs (Brown et al., 2020; Touvron et al., 2023b; OpenAI, 2022) have led the research of NLP into a new era. Recent advancements, including supervised fine-tuning and alignment with human values (Ouyang et al., 2022; Chung et al., 2022; Bai et al., 2022), have further augmented the capabilities of LLMs, enabling them to adhere more closely to human instructions and ethical considerations. Nonetheless, challenges persist since LLMs may show variability in the model responses, especially under the scenarios of MCQA. Robinson and Wingate (2022) termed the ability to associate the answer options and corresponding symbols as multiple choice symbol binding (MCSB) and proved that the MCSB ability varied significantly by models. Additionally, Wang et al. (2023) revealed vulnerabilities in the ranking of candidate responses, which could be manipulated by altering the presentation order. Zheng et al. (2023) investigated the token selection bias in LLMs. Kadavath et al. (2022) explored the reliability of the LLM performance and calibration, focusing exclusively on a set of private models under the MCQA settings. Recently, West et al. (2024) examined the performance gap between generative and discriminative tasks in LLMs. Pezeshkpour and Hruschka (2024) proposed two calibration techniques to reduce variability in LLM responses. Previous work has focused on mitigating bias in answer options or developing techniques to ensure that LLMs exhibit consistency across different orders of options. A common thread among these studies is the belief that if LLMs can demonstrate robustness to variations in the order of answer options, their predictive reliability can be improved. However, our research identifies another limitation: even if LLMs consistently predict the correct answer across varied option orders, they may still struggle to accurately

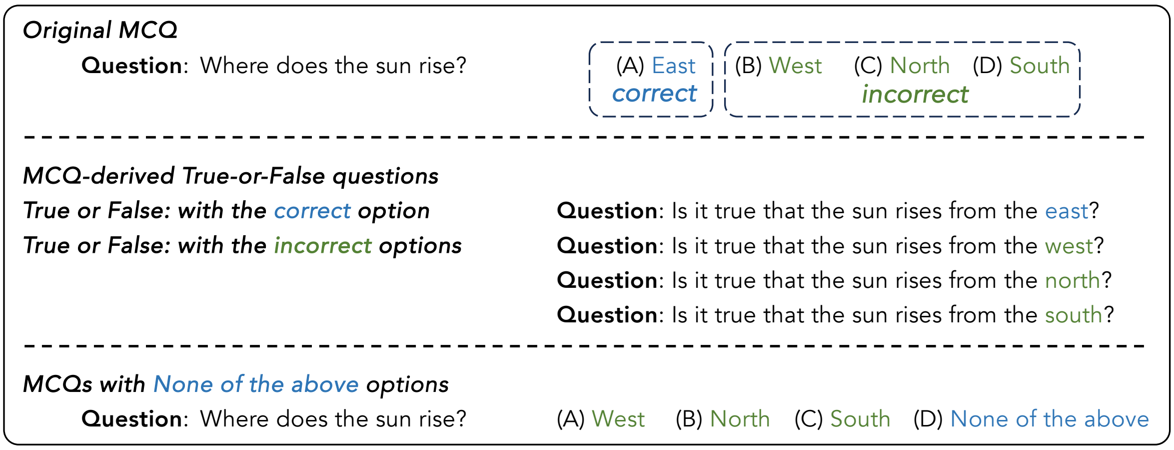

Figure 3: A case of an original MCQ, True-or-false questions derived from the MCQ and the MCQ with the correct options replaced by "None of the above".

answer questions derived from the original MCQ because LLMs may perform MCQA by selecting which is the least incorrect.

3 Does Invariability Imply Reliability?

As discussed previously, prior research has demonstrated that LLMs may exhibit variability in their responses across different permutations of answer options in MCQs (Robinson and Wingate, 2022; Pezeshkpour and Hruschka, 2024). Various techniques have been explored to ensure that LLMs exhibit invariability in response to such permutations, with the assumption that invariability could serve as a proxy for model reliability in MCQ tasks. However, this raises an important question: does invariability truly equate to reliability?

3.1 Models and Datasets

In this study, we focus on evaluating several prominent generative models that have garnered significant attention within both academic and public domains. These include LLaMA models (Touvron et al., 2023b) (LLaMA 3 8B, LLaMA 2 13B[1], LLaMA 3 70B), Mixtral (Jiang et al., 2024) (Mixtral 8×7B), and ChatGPT (OpenAI, 2022) (ChatGPT-3.5, ChatGPT-4o, and ChatGPT-4o-mini). For the datasets, we sample from two widely recognized benchmarks: the first is MMLU (Hendrycks et al., 2021), a general-domain benchmark widely used in MCQA evaluation for LLMs; the second is MedMCQA (Pal et al., 2022), which is specific to the medical domain and requires extensive domain-specific knowledge, presenting a significant challenge for most LLMs.

3.2 Invariability Dataset Preparation

Due to our work aiming to demonstrate certain deficiencies in using MCQs to test LLMs, and to facilitate subsequent experiments, we first conduct tests on MCQs with option permutations on subsets of MMLU and MedMCQA datasets with questions testing for knowledge instead of reasoning (like math). Then, we filter out the subsets MMLU$^{\blacklozenge}$ and MedMCQA$^{\blacklozenge}$ where the LLMs show invariability.

3.3 Transforming to True-or-False Format

We transform the original MCQAs in only MMLU$^{\blacklozenge}$ and MedMCQA$^{\blacklozenge}$ into a True-or-False (T/F) format to explore how the LLMs behave on the questions that they have predicted correctly with invariability in MCQA-format. For every MCQA instance, we generate T/F-format questions, including one with the correct option (T/F: correct) and other questions with the incorrect options (T/F: incorrect) [2] as depicted in Figure 3, anticipating that the LLMs will respond accurately with "Yes" and "No" respectively.

Table 1 presents an analysis of LLM performance on T/F questions. If consistency were a reliable indicator of accuracy, we would expect LLMs to achieve near-perfect performance in this format on both the "T/F: correct" and "T/F: incorrect" datasets. In the few-shot scenario, we provide the LLMs with two examples as demonstrations, one with the answer "correct" and the other with

[1]LLaMA 3 currently comprises models with 8B and 70B parameters

[2]"None of the above"-like options are not transformed into T/F format.

Table 1: Accuracy of LLMs on the True-or-False questions derived from MCQs. ♦ means the subsets of datasets where LLMs have answered correctly across all re-ordered answer options of the MCQs.

		0-shot		few-shot	
		MMLU♦	MedMCQA♦	MMLU♦	MedMCQA♦
LLaMA 3 8B	`T/F: correct`	93.3	92.4	95.1	93.0
	`T/F: incorrect`	43.7	52.7	48.9	55.5
LLaMA 2 13B	`T/F: correct`	70.3	70.5	73.1	71.8
	`T/F: incorrect`	52.5	55.8	55.4	55.6
LLaMA 3 70B	`T/F: correct`	97.2	92.3	97.8	92.8
	`T/F: incorrect`	41.7	42.3	44.1	36.8
Mixtral 8×7B	`T/F: correct`	90.7	83.2	91.2	82.7
	`T/F: incorrect`	58.8	54.7	59.2	55.5
ChatGPT-3.5	`T/F: correct`	87.6	76.6	87.8	75.6
	`T/F: incorrect`	68.5	69.9	68.9	71.6
ChatGPT-4o-mini	`T/F: correct`	93.6	89.0	94.1	88.7
	`T/F: incorrect`	65.9	72.4	65.2	73.1
ChatGPT-4o	`T/F: correct`	95.1	88.7	96.4	89.9
	`T/F: incorrect`	73.0	80.2	70.2	80.4

Table 2: Accuracy of LLMs on the MCQA♦ datasets with the correct options replaced with the "None of the Above" option. ♦ denotes the subsets of datasets where LLMs have shown invariability across re-ordered answer options before the alteration of "None of the above".

	0-shot		few-shot	
	MMLU♦	MedMCQA♦	MMLU♦	MedMCQA♦
LLaMA 3 8B	18.3	20.6	19.1	20.4
LLaMA 2 13B	17.2	12.3	6.7	0.0
LLaMA 3 70B	22.6	24.9	31.3	29.8
Mixtral 8×7B	24.7	31.1	40.3	30.1
ChatGPT-3.5	23.7	36.0	48.5	41.7
ChatGPT-4o-mini	43.6	42.6	50.6	51.4
ChatGPT-4o	60.3	60.1	68.6	67.5

that of "incorrect". In practice, LLMs demonstrate varying levels of accuracy on the "True/False: correct" datasets, ranging from 70.3% to 96.4%. This suggests that LLMs can generally perform well on T/F questions derived from MCQs with correct options. However, a notable performance decline occurs when the T/F questions include incorrect options. For instance, LLaMA 3 70B achieves an accuracy as low as 36.8% on the "T/F: incorrect" datasets based on MedMCQA♦. Similar trends are observed across almost all tested LLMs. This highlights a critical limitation: while LLMs tend to be consistent when handling questions with both MCQA and T/F formats with correct options, they frequently misclassify statements containing incorrect options as correct.

3.4 "None of the Above" Options

Kadavath et al. (2022) examined the potential impact of "None of the above" options on certain close-source LLMs with the *entire* MCQA datasets, without considering the confounding factor of variability. In this study, we extend the analysis to a broader range of LLMs, focusing exclusively on variability-free sub-datasets, that are MMLU♦ and MedMCQA♦, as illustrated in Figure 3. For the few-shot scenarios, demonstrations involve MCQs with the correct answer of "None of the above". As presented in Table 2, the substitution of correct options with the "None of the above" leads to a substantial decline in model performance. With the exception of scenarios involving ChatGPT-4o and ChatGPT-4o-mini, the LLMs consistently fail to select the "None of the Above" option in place of the

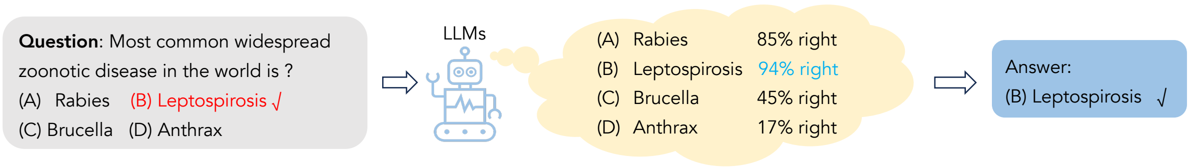

Figure 4: Illustration of the hypothesis: LLMs may perform MCQA by selecting the least incorrect option.

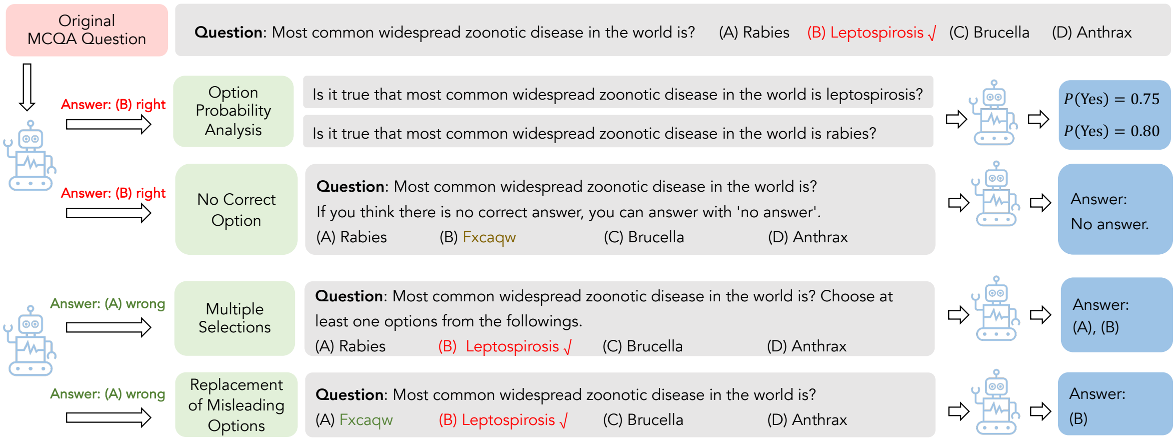

Figure 5: The validation experiments consist of (1) "Option probability analysis" (2) "No correct option" on the MCQA$^{\blacklozenge}$, along with (3) "Multiple selections" and (4) "Replacement of misleading options" on the original MCQs where the LLMs made incorrect predictions. (1) assesses the confidence of LLMs when handling T/F questions with correct and incorrect answer options from the MCQA$^{\blacklozenge}$ datasets; (2) assesses whether LLMs can respond with "no answer" when presented with MCQs that do not contain correct options; (3) prompts the LLMs to selected all the answers options they consider correct in the MCQs where they previously made incorrect prediction. (4) replaces the incorrect options previously chosen by the model with non-semantic tokens.

correct answer, with accuracy not exceeding 48.5% across all other scenarios. Even the ChatGPT-4o model demonstrates a failure rate exceeding 31% on the MCQA$^{\blacklozenge}$ datasets.

3.5 Analysis

Despite the invariability in LLM performance on MCQs, modifying the MCQA datasets to include (1) T/F questions derived from incorrect options in the original MCQs, and (2) scenarios where the correct option is replaced by "None of the Above," results in a pronounced performance decline. This finding highlights a critical issue: the invariability exhibited by LLMs in handling multiple-choice questions does not necessarily signify reliability. Specifically, the models demonstrate unexpected behavior when confronted with questions involving options other than the correct answer options, raising concerns about the robustness of their decision-making processes in such contexts.

4 LLMs May Do MCQA by Selecting Which Is the Least Incorrect

Through our experiments with modified datasets only with the incorrect options in the original MCQs, we have shown that invariability in LLM responses does not necessarily equate to reliability. Based on these observations, we propose the following hypothesis:

While LLMs demonstrate invariability on specific MCQs with a consistent answer option, they may not regard this option as uniquely correct. Rather, LLMs may treat the selected option as the most accurate among the choices, without dismissing the potential partial correctness of other, incorrect options—albeit to a lesser degree than the chosen one.

which is visually illustrated in Figure 4.

If this hypothesis holds true, it suggests that LLMs may recognize some of the unselected options as partially correct. This could offer a plausible explanation for the observed model behavior in the aforementioned experiments. To further inves-

Table 3: Confidence of answer options in MCQA tasks with open-source LLMs. $C_{correct}$: mean confidence of the correct options. $C_{incorrect^*}$: mean confidence of the incorrect options that score with the highest confidence. R_c: relative confidence score. ♦ denotes the experiments on the sub-datasets where the LLMs predict correctly with the original MCQA settings.

	LLaMA 3 8B		LLaMA 2 13B		LLaMA 3 70B		Mixtral 8×7B	
	MMLU♦	MedMCQA♦	MMLU♦	MedMCQA♦	MMLU♦	MedMCQA♦	MMLU♦	MedMCQA♦
$C_{correct}$	16.4	18.7	35.0	36.0	23.6	25.5	31.2	34.6
$C_{incorrect^*}$	16.3	18.5	32.5	30.1	20.3	21.9	29.1	33.1
R_c	99.4%	98.9%	92.9%	83.6%	86.0%	85.9%	93.3%	95.7%

Table 4: Ratio of the instances where the LLMs can generate "no answer" on the MCQs with no correct option under the few-shot settings.

	MMLU♦	MedMCQA♦
ChatGPT 4o-mini	32.8	32.6
ChatGPT 4o	59.5	60.0

tigate this phenomenon, we explore the behavior of the models from the following four perspectives, as depicted in Figure 5.

4.1 Option Probability Analysis

Leveraging the MMLU♦ and MedMCQA♦ datasets, we investigate the confidence of LLMs by examining the distribution of token probabilities for answer option tokens, as discussed in Chen et al. (2023). To do this, we convert MCQs into T/F format as illustrated in Figure 5. The confidence scores for each answer option are derived based on the "yes" or "no" token probabilities assigned by the LLMs. For T/F questions that include the correct answer options from the original MCQs, we compute the confidence score $C_{correct}$ using instances where the LLMs made correct predictions. This score quantifies the degree of confidence the LLMs exhibit when recognizing a claim with the correct option as accurate. Conversely, for T/F questions containing incorrect answer options from the MCQs, we compute the confidence score $C_{incorrect}$ based on cases where the LLMs made incorrect predictions. This score measures how confidently the LLMs mistakenly identify a claim with an incorrect option as correct. For a specific MCQ, we consider the incorrect* option with the highest confidence in corresponding T/F questions for $C_{incorrect^*}$.

$$C_{correct} = \frac{1}{N} \sum_{z_{\text{"yes"}} > z_{\text{"no"}}} \frac{e^{z_{\text{"yes"}}}}{\sum_{t \in V} e^{z_t}}$$

$$C_{incorrect} = \frac{1}{M} \sum_{z_{\text{"no"}} > z_{\text{"yes"}}} \frac{e^{z_{\text{"no"}}}}{\sum_{t \in V} e^{z_t}}$$

where z_t is the logit for each token t in the vocabulary, and V denotes the full vocabulary set. N and M is the number of corresponding questions.

Table 3 demonstrates the confidence of LLMs for the correct options ($C_{correct}$) and the incorrect* options ($C_{incorrect^*}$), along with the relative confidence scores. The experimental results show that while LLMs consistently consider the incorrect* options as less correct than the correct options, demonstrated by all relative confidence being below 100%, the incorrect* options still achieve substantial confidence ranging from 83.6% to 99.4% of those for the correct options. Consequently, despite invariability, LLMs may still perceive certain incorrect options as correct, though to a lesser extent compared to the correct ones.

4.2 MCQA with No Correct Option

In our previously described scenarios for evaluating LLMs on MCQs, there has always been a correct answer among the candidate options. However, when no correct answer is present, we expect LLMs to recognize that the question is flawed. To guide the model in such cases, we prompt the LLMs with "If you think there is no correct answer, you can respond with 'no answer'," to observe whether the model generates a "no answer" response. Empirically, even large-scale open-source models such as LLaMA 3 70B struggle to effectively follow this instruction. Consequently, our analysis focuses on the ChatGPT-4 series models. Using the MCQA♦ dataset, where the models exhibit invariability, we replace the correct options with non-semantic tokens. As shown in Table 4, with appropriately designed prompts, ChatGPT-4o successfully identifies that there is no correct answer in approximately 60% of the MCQs. However, for the remaining

Table 5: Experiments on the altered MCQA datasets with multiple selections. $\texttt{Recall}_{correct}$: recall of the correct options. $\texttt{Recall}_{misleading}$: recall of the misleading options (the incorrect options LLMs have chosen). † denotes the subsets where the LLMs have generated incorrect answers on the original MCQA datasets.

	MMLU†		MedMCQA†	
	$\texttt{Recall}_{correct}$	$\texttt{Recall}_{misleading}$	$\texttt{Recall}_{correct}$	$\texttt{Recall}_{misleading}$
LLaMA 3 8B	85.1	70.1	82.3	74.2
LLaMA 2 13B	92.5	67.5	78.9	70.3
LLaMA 3 70B	94.2	72.2	84.0	80.1
Mixtral 8×7B	91.5	76.4	84.6	85.4
ChatGPT-3.5	85.1	66.7	81.3	69.7
ChatGPT-4o-mini	85.0	89.2	84.4	91.6
ChatGPT-4o	89.9	90.8	83.2	92.1

40% of the questions, it still selects one option as correct, indicating that LLMs do not fully recognize all incorrect options as incorrect.

4.3 MCQA with Multiple Selections

For the MCQA datasets involved in this study, LLMs are tasked with identifying only one correct option per MCQ. We collect the instances where LLMs incorrectly predict the answers, denoted as MMLU† and MedMCQA†. In the above instances, the incorrect options which LLMs have regarded as the correct ones mistakenly are defined as *misleading* options. Then, LLMs are prompted to recognize all plausible correct options among all answer options. Table 5 showcases the recall for the correct and misleading options for the instances where LLMs render multiple selections. The results reveal that the correct options are included in the selections in over 78.9% of instances, reaching up to 94.2%. This indicates that the LLMs also recognize the correct options as correct but less correct than the misleading ones.

4.4 MCQA with the Misleading Option Replacement

Apart from the multi-selection scenario, we explore the impact of replacing misleading options with arbitrary non-semantic tokens in MMLU† and MedMCQA†. Table 6 elucidates that the LLMs correctly identify the correct options in 30.9% to 58.0% of instances, highlighting the influence of misleading options on their predictions. For cases where LLMs continue to make incorrect predictions, a fundamental deficit in relevant knowledge likely underpins the LLM incapability to generate the correct answers.

Table 6: Ratio of the instances where the LLMs turn to predict correctly with the replacement of misleading options (the incorrect options LLMs have chosen).

	MMLU†	MedMCQA†
LLaMA 3 8B	42.1	30.9
LLaMA 2 13B	58.0	41.0
LLaMA 3 70B	46.3	41.1
Mixtral 8×7B	50.4	34.8
ChatGPT-3.5	53.6	44.5
ChatGPT-4o-mini	39.7	41.2
ChatGPT-4o	41.1	48.6

4.5 Summary

The analyses conducted across the four experimental scenarios provide substantial support for the validity of the proposed hypothesis. This leads to a critical observation that highlights a fundamental limitation of using MCQA-based evaluations to assess the capabilities of LLMs: *In the context of MCQA, while LLMs may select the correct answer, there remains a possibility that they also attribute correctness to other, incorrect options.*

5 MCQA+ for Robust Evaluation

Dataset Preparation Experimental analyses have revealed significant limitations in using the MCQA benchmark to evaluate LLMs, highlighting that LLMs may consider options they did not select in MCQs as correct. To address this, we propose an augmentation approach based on the original MCQA dataset, informed by the empirical findings from the above experiments. Each MCQ is transformed into one of the following settings: (a) Original MCQs; (b) MCQs with re-ordered answer options; (c) True-or-False questions derived from correct answer options; (d) True-or-False questions derived from incorrect answer options; (e) MCQs

Table 7: Model performance on the original MCQA, MCQA+, MCQA+^{hard} and MCQA+ ($\times$1) datasets.

	MMLU				MedMCQA			
	MCQA	MCQA+	MCQA+^{hard}	MCQA+ ($\times$1)	MCQA	MCQA+	MCQA+^{hard}	MCQA+ ($\times$1)
LLaMA 3 8B	75.1	56.8	40.5	58.4	47.9	36.3	24.4	34.1
LLaMA 2 13B	72.7	46.1	21.2	45.3	43.2	38.8	16.5	40.0
LLaMA 3 70B	78.9	60.2	46.8	57.1	53.1	42.8	29.1	44.4
Mixtral 8$\times$7B	71.2	58.6	43.7	58.5	51.4	43.5	28.7	42.7
ChatGPT-3.5	65.0	63.2	57.8	64.0	56.9	53.9	49.6	54.1
ChatGPT-4o-mini	79.0	70.9	63.0	72.4	68.3	64.4	60.2	63.8
ChatGPT-4o	82.4	80.7	73.1	79.7	72.7	69.7	64.0	71.3

where the correct options are replaced with "None of the above"; and (f) MCQs with no correct options, where LLMs are expected to generate "no answer" as the response. Using these settings, we propose three dataset augmentation approaches: (1) **MCQA+**: Encompasses all of the above settings; (2) **MCQA+**^{hard}: Includes only the settings (b, d, e, f), serving as a much more challenging benchmark for LLMs. (3) **MCQA+ ($\times$1)**: Samples one question from the MCQA+ settings as an efficient approximation to MCQA+.

The mean accuracy across all settings is adopted as the evaluation metric for LLMs. For settings with multiple questions (e.g., (b)), accuracy$_b$ is measured as the mean accuracy across all questions in setting (b). Table 7 illustrates the comparative performance of LLMs on the original MCQA dataset, MCQA+, MCQA+^{hard}, and MCQA+ ($\times$1). Performance on the MCQA+ dataset shows a significant decline across all LLMs compared to the original MCQA dataset. For example, accuracy for LLaMA 3 8B dropped from 75.1% to 56.8% on the MMLU dataset. Performance on the MCQA+^{hard} benchmark is even lower, likely for reasons discussed in Section 4. Even ChatGPT-4o experienced a performance decline of 9.3% from the original MCQA to MCQA+^{hard} on the MMLU dataset.

Although MCQA+ and MCQA+^{hard} provide a more accurate reflection of LLM capabilities, they entail significantly higher computational costs compared to the original MCQA. Therefore, MCQA+ ($\times$1), which samples from MCQA+ for each MCQ, requires no additional computational cost compared to the original MCQA. As shown in Table 7, this cost-efficient approach still effectively reveals the true capabilities of LLMs.

Discussion The MCQA+ strategy offers an efficient and refined approach to augmenting existing MCQ datasets, enabling a more accurate assessment of model capability. However, ensuring consistent performance across tasks not addressed by MCQA+, such as generative tasks, remains a challenge. Based on the results of this study, we hypothesize that the observed performance decline may be linked to the training strategies of LLMs in generative tasks during pre-training and instruction-tuning, that is predicting the next token based on the ranking of probability. LLMs have been only instructed to choose the best options but not to treat those options as exclusively correct. While reinforcement learning aligns the model's outputs with human preferences, it does not fully resolve the issue where incorrect options might still receive high probabilities in different contexts. This could explain the discrepancies in model performance between discriminative and generative tasks, as noted by West et al. (2024). As such, the reliability of evaluating LLMs using MCQs necessitates further scrutiny and attention.

6 Conclusion

In this study, we investigated the limitations of using MCQA as a benchmark for evaluating the performance of LLMs through a comprehensive series of experiments. Our findings suggest that LLMs may not always select the distinctly correct option, but instead opt for the least incorrect option. This behavior raises concerns about the robustness and reliability of MCQA-based evaluations. To address these issues, we proposed the MCQA+ dataset augmentation method, which provides a more refined evaluation framework by challenging LLMs to demonstrate a deeper level of understanding. Our work underscores the importance of continued efforts to develop more comprehensive evaluation methodologies for LLMs, ensuring that their true capabilities are accurately reflected, not only in discriminative tasks, such as MCQA but also in broader, more complex contexts.

Limitations

In this study, we analyzed a probable issue that LLMs may face when answering MCQs. Building on previous research on the variability of large models, we conducted experiments demonstrating that although LLMs can achieve impressive results on MCQA benchmarks, their treatment of incorrect options may be ambiguous, potentially recognizing incorrect options in other contexts. This issue may stem from negative impacts introduced by different stages of the training objectives of LLMs, such as instruction-tuning and RL-based alignment, presenting a broader challenge for the entire NLP field. Therefore, we propose a method to improve the reliability of model evaluations through diversity testing, which represents a trade-off between efficiency and accuracy, without fundamentally addressing the core challenges of evaluating LLMs based on MCQs. We aim to draw attention from the community to the potential long-term impacts of this issue and to collaboratively work towards resolving it.

Acknowledgements

We thank the anonymous reviewers for their insightful and constructive comments and gratefully acknowledge the support of the National Natural Science Foundation of China [62206079]; and the Heilongjiang Provincial Natural Science Foundation of China [2023ZX01A11]. We also appreciate the support from China Mobile Group Heilongjiang Co., Ltd. @ on our research, the research is jointly completed by both parties.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2022. Chatgpt. `https://chat.openai.com`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Joshua Robinson and David Wingate. 2022. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

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

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2024. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.