When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards

Norah A. Alzahrani*, **Hisham Abdullah Alyahya***, **Yazeed Alnumay**, **Sultan Alrashed**, **Shaykhah Z. Alsubaie**, **Yousef Almushayqih**, **Faisal Abdulrahman Mirza**, **Nouf M. Alotaibi** **Nora Al-Twairesh**, **Areeb Alowisheq**, **M Saiful Bari**, **Haidar Khan***†

National Center for AI (NCAI), Saudi Data and AI Authority (SDAIA)
Riyadh, Saudi Arabia

Abstract

Large Language Model (LLM) leaderboards based on benchmark rankings are regularly used to guide practitioners in model selection. Often, the published leaderboard rankings are taken at face value — we show this is a (potentially costly) mistake. Under existing leaderboards, the relative performance of LLMs is highly sensitive to (often minute) details. We show that for popular multiple-choice question benchmarks (e.g., MMLU), minor perturbations to the benchmark, such as changing the order of choices or the method of answer selection, result in changes in rankings up to 8 positions. We explain this phenomenon by conducting systematic experiments over three broad categories of benchmark perturbations and identifying the sources of this behavior. Our analysis results in several best-practice recommendations, including the advantage of a *hybrid* scoring method for answer selection. Our study highlights the dangers of relying on simple benchmark evaluations and charts the path for more robust evaluation schemes on the existing benchmarks. The code for this paper is available at https://github.com/National-Center-for-AI-Saudi-Arabia/lm-evaluation-harness.

1 Introduction

The advent of transformer-based Large Language Models (LLMs) (OpenAI, 2023; Deepmind, 2023; Anthropic, 2023; Anil et al., 2023; Touvron et al., 2023) has led to a generational leap in generative models, enabling interaction with computing devices through natural language. This advancement encompasses improvements that have rendered many earlier benchmarks and leaderboards obsolete (Laskar et al., 2023; Shen et al., 2023), leading to the compilation of more challenging and comprehensive tests. However, the current generation of leaderboards still does not satisfy many of the requirements of researchers and practitioners looking to build on LLMs (Ethayarajh and Jurafsky, 2021; Dehghani et al., 2021). Since LLMs are extremely expensive to both train and inference, selecting the LLM (or LLM training recipe) is often the most costly decision for the entire project. Stable leaderboards are critical to making the right decision.

Leaderboards based on multiple choice questions (MCQ) for evaluation (Wang et al., 2018, 2019; Nie et al., 2019; Zhong et al., 2023; Hendrycks et al., 2020) present both convenience and significant limitations (Pezeshkpour and Hruschka, 2023; Zheng et al., 2023). While MCQs offer an *automated* and *quantifiable* means to assess certain aspects of model ability (e.g., knowledge), they fall short as a stable means to measure performance. Figure 1 demonstrates the instability of the leaderboard ranking of one popular benchmark, Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), under small perturbations.

Moreover, the reliance on MCQs raises concerns about the models being *overfit* to these benchmarks, potentially excelling in structured tests while lacking real-world applicability. This discrepancy highlights the need for more holistic and diverse evaluation methods that transcend the simplicity of MCQs (Liang et al., 2023). It also prompts critical reflection on how these models might inadvertently be trained to achieve high scores through spurious correlations, pattern recognition, and optimization for specific question formats rather than genuine language comprehension or knowledge. As LLMs continue to evolve, it is imperative to develop evaluation frameworks that can more accurately assess their abilities in a way that mirrors the complexity of real-world use.

Despite being widely used, benchmarking with MCQs has turned out to be anything but simple. It requires the full synchronization of eval-

*Core contributor
†Corresponding author: haidark@sdaia.gov.sa

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13787–13805

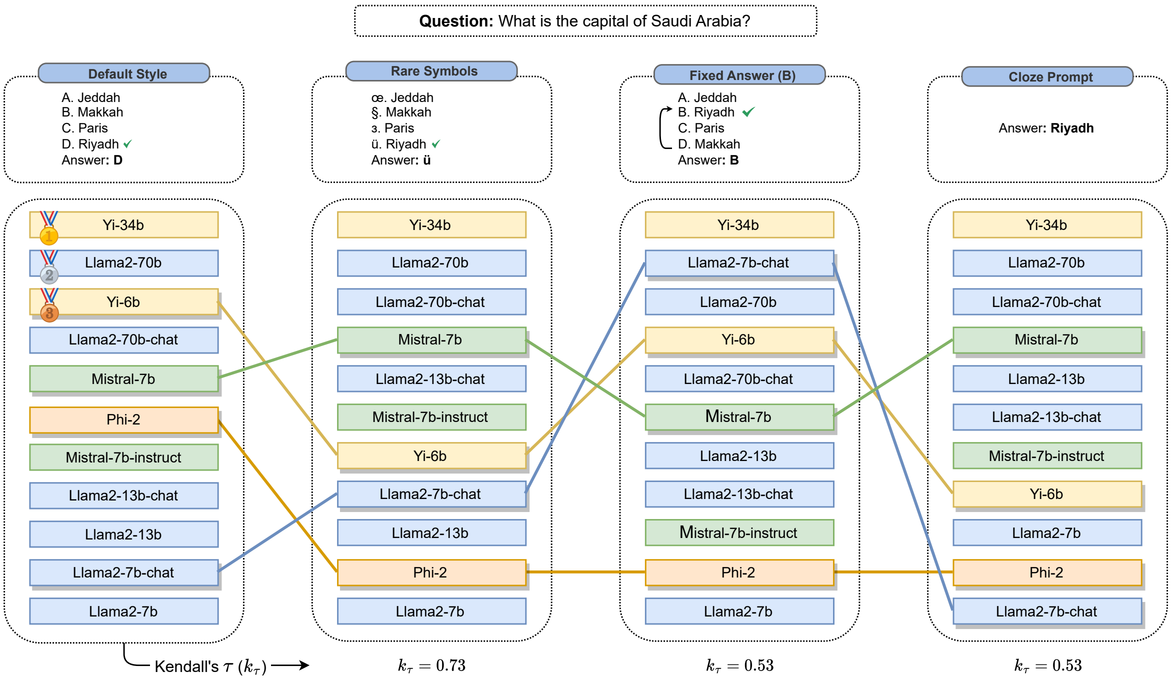

Figure 1: Minor perturbations cause major ranking shifts on MMLU (Hendrycks et al., 2020). Models can move up or down up to eight positions on the leaderboard under small changes to the evaluation format. Columns (from left): 1) Original ranking given by MMLU using answer choice symbol scoring (a common default). 2) Ranking under an altered prompt for the same questions, where answer choice symbols are replaced with a set of rare symbols. 3) Setting where the correct answer choice is fixed to a certain position (in this case, B). 4) Using the cloze method for scoring answer choices. Under each new ranking, we report Kendall's τ (Kendall, 1938) with respect to the original ranking (lower k_τ indicates more disagreement between rankings)

uation frameworks and results often vary wildly due to nuanced differences. For example, minor changes in prompting and scoring can produce invalid results for particular LLMs[1]. Recent studies have investigated the issue of sensitivity in evaluating LLMs, with some demonstrating that LLMs are susceptible to the ordering of answer choices and bias towards specific tokens/symbols (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023; Lu et al., 2022), and others exploring different prompt modifications and their effects in benchmarking LLMs (Mizrahi et al., 2024; Sclar et al., 2023; Weber et al., 2023; Zhao et al., 2021).

In this work, we conduct a broad range of minor perturbation experiments to MCQ benchmarks and observe the disruption it causes to model rankings on leaderboards. We also take additional steps to precisely identify the limitations of LLMs on this measurement approach.

The contributions of this paper can be summarized as follows:

[1]https://huggingface.co/blog/evaluating-mmlu-leaderboard

1. Existing model rankings on popular benchmarks **break down under slight perturbations**, particularly in the medium to small model sizes.

2. This behavior can be explained by the susceptibility of all tested LLMs to various forms of bias in MCQ.

3. Some families of LLMs have an over-reliance on format, pointing to potential benchmark leakage.

4. We find that LLMs also exhibit bias to the scoring method for answer choices in MCQ.

5. We demonstrate that some categories of modifications do not affect the benchmark rankings.

2 LLM Evaluation with MCQs

Evaluating LLMs with MCQs has rapidly become a standard for measuring the knowledge and reasoning capabilities of the model (OpenAI, 2023; Anil et al., 2023; Deepmind, 2023; Jiang et al., 2023).

Many such MCQ benchmarks have been used to measure LLMs, including Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), Ai2 Reasoning Challenge (ARC) (Clark et al., 2018a), and Common-sense Question Answers (CSQA) (Saha et al., 2018).

Mechanically, testing LLMs with MCQs is accomplished by presenting the question along with the answer choices to the model and selecting the choice deemed most probable by the model. Although this setup appears straightforward, LLMs react in unpredictable ways to formatting and other minor changes to the questions or the answers. LLM performance on an MCQ test can change with the introduction of an extra space (e.g., between the question and answer) or adding an additional instructional phrase (e.g., "Choices:"). In addition to this brittleness, Pezeshkpour and Hruschka (2023) found changes to the order in which answer choices are presented to GPT4 and instructGPT can change the model's prediction.

These findings lead us to take a deeper look at how MCQ-based benchmark results are affected by small perturbations to question formats, LLM prompts, presentation of few-shot examples, and other dimensions. In particular, we introduce variations in three categories:

- **Answer choice format and ordering**: testing the limits of LLM sensitivity to ordering and formatting (Section 3.1).
- **Prompt and scoring modifications**: changing text included in the prompt and analyzing different scoring schemes (Section 3.2).
- **In-context knowledge manipulation**: inserting relevant/irrelevant information in the prompt and/or few-shot examples (Section 3.3).

Our main aim is to quantify how these small perturbations/variations **change the rankings** of a set of models on a particular benchmark. As MCQ benchmarks-based leaderboards are often used to compare models and guide model selection, we investigate the robustness of benchmarks for this purpose. Figure 1 demonstrates how existing benchmarks exhibit significant undesirable shifts in rankings under small perturbations.

3 Methods

In this section, we describe and justify the perturbations we apply in each category. We note that some MCQ test changes, like modifying the order of answer choices, can change performance even for humans but the effect is typically not pronounced (Lions et al., 2021). In general, our modifications are designed to be small perturbations to the MCQ and prompts that *should not* affect performance. The exception to this is some of the **in-context knowledge manipulations** described in Section 3.3, which are designed to improve or degrade performance drastically.

3.1 Answer choice format and ordering

In light of earlier findings related to selection bias (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023; Lu et al., 2022), we investigate the effects of changes to the presented order of answer choices and changes to the symbols associated with answer choices.

Random choice order Our first study aims to uncover how dependent MCQ benchmark performance and rankings are on the original ordering of the answer choices. We apply two simple schemes to randomly change the order of answer choices presented to the model: (i) swapping choices using a fixed set of swaps for all questions and (ii) randomly assigning new positions to each choice while ensuring each choice is moved to a different position.

Biased choice order In this setting, the correct answer choice is set to a fixed position across the entire test to measure bias toward predicting answers at particular positions. For zero-shot, we simply set the correct answer choice to each of the positions in turn.

In the few-shot case, we examine the influence of biasing the correct answers in the examples to the model's inherent bias to particular positions. For each question, we fix the correct answer of the examples to each position in turn. We then modify the test question in two ways: (i) unchanged answer choices and (ii) correct choice fixed to the same position as examples. This setup is shown in Figure 2.

Answer choice symbols The symbols used for the answer choices (e.g. A, B, C, D) also play a role in model bias (Zheng et al., 2023). Thus, we experiment with replacing the symbols with alternative and less common tokens. This aims to decouple the bias to particular positions from the bias to symbols or the relative ordering in natural

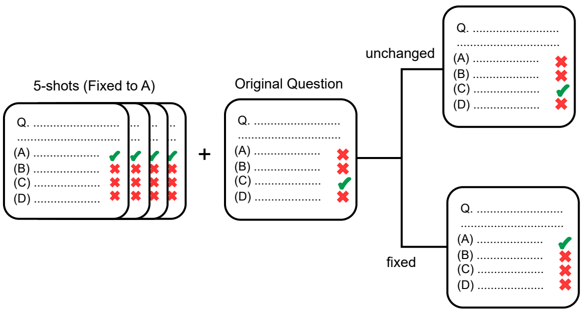

Figure 2: Experiment setup for probing position bias with few-shot examples.

symbols. We replace *['A', 'B',' C', 'D']* with the following two sets of symbols:

- Set 1: *["$", "&", "#", "@"]* comprising of common tokens that are language-independent.
- Set 2: *["æ", "§", "Ze (Cyrillic)", "ü"]* consisting of rare tokens in the vocabulary without any implicit relative order.

In the few-shot setting, we test both assigning fixed ordering for the replaced symbols in the examples as well as changing the ordering across examples.

3.2 Prompt and scoring modifications

LLMs exhibit high sensitivity to variations in prompt formatting (Sanh et al., 2021; Mishra et al., 2022), forcing benchmark developers to unify prompt templates within the same evaluation scheme. However, it remains unknown if certain models have an affinity towards any specific prompt templating style. It is unclear how benchmarking prompt choices advantage/disadvantage different models. In addition to that, the scoring style may change depending on how we are prompting the context of a query. We distinguish three major categories of scoring methods for MCQs.

- **Symbol scoring**: Prompt template is structured as question followed by answer choices. The model chooses the answer based on the likelihood scores for the answer choice symbol. Used in Hendrycks et al. (2020).
- **Hybrid scoring**: Prompt template is structured as a question followed by answer choices. The model chooses the answer based on the likelihood scores for the answer choice content normalized by length. Used in Raffel et al. (2020); Sanh et al. (2021); Chowdhery et al. (2022)
- **Cloze scoring**: Prompt templates are structured as a question followed by a single answer choice. Maximum normalized likelihood scores over all answer choices define the prediction. Used in Clark et al. (2018a).

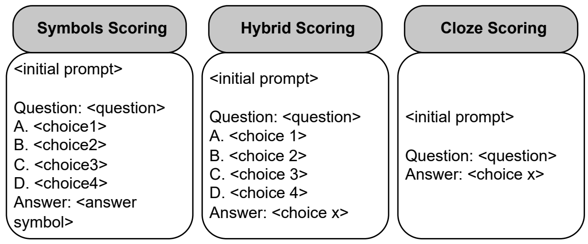

Figure 3: Answer choice scoring methods for LLMs. The symbols and hybrid scoring methods are most similar, sharing identical prompts. Cloze scoring does not reflect a "true" MCQ style, as the model is not shown all the options. However, due to its prevalence we compare it to the other methods as a baseline.

Figure 3 gives an overview of each scoring method. In addition, we also investigate further modification of instruction and sentinel tokens in the prompt template.

Prompt instructions To assess the impact of subtle token alterations in prompt instructions, we conduct experiments on *(i)* removing question subject information and *(ii)* adding "Correct" alongside the answer. These targeted changes aim to identify the robustness in response to certain tokens, particularly when they carry crucial information, as well as to evaluate the influence of contextual bias introduced by minor modifications of the instruction text.

3.3 In-context knowledge manipulation

In this category, we attempt to measure model and benchmark robustness in the few-shot setting by testing the entire spectrum of knowledge injected in the few-shot examples. In particular, we experiment under the following settings:

Correct answer provided We provide the target question and the correct answer in the prompt as an example to the model. This corresponds to the simplest setting for the model, where it only needs to look up the answer in the context.

Incorrect answer provided This setting is the opposite of the former. The target question is provided with an incorrect answer as an example. It is challenging as the model must ignore the context and determine the correct answer independently.

Trivial examples We replace few-shot examples with simple questions the model is known to be able to answer (typically related to the language/text of the question itself). The only information the examples convey is related to formatting (Soltan et al., 2022). We create three versions of these questions and answers using GPT-4 and ensure the model can answer them correctly (as shown in Figure 8).

Out of domain examples Instead of providing examples from the same subject as the target question, we add out-of-domain questions (from another subject) as the few-shot examples. This setting corresponds to a difficulty level between the original format and providing trivial examples.

4 Experiments

In the bulk of our experiments, we focus on the MMLU benchmark due to the extensive nature of our experiments (11 models, 22+ settings), and extend some experiments to ARC-challenge to show generalizability.

MMLU (Hendrycks et al., 2020) is a commonly used benchmark for comparing LLMs, consisting of 57 subjects spanning four domains: humanities, STEM, social sciences, and others. Each subject includes at least 100 multiple-choice questions with 4 answer choices. The entire benchmark contains 14,042 questions (Tables A.29 and A.30 have a breakdown of the MMLU subjects and their distributions).

Ai2 Reasoning Challenge (Clark et al., 2018b) is a benchmark consisting of 7787 grade school science questions. The benchmark is split into two sets: Easy and Challenge. We conduct experiments on the Challenge set (ARC-C) which is proven to contain harder questions for existing models. The questions in ARC-C have 3-5 answer choices.

Unless otherwise stated, the reported score for each experiment/model combination on MMLU is the mean accuracy across all 14,042 questions. All tested model tokenizers encode the multiple-choice answers as single tokens. Hence, the accuracy is equivalent to the normalized accuracy. All baseline and modified MMLU benchmarks were performed using the LM Evaluation Harness (Gao et al., 2023) library. Their implementation of MMLU measures the log-likelihood of each of the answer tokens *['A', 'B', 'C', 'D']* after the input prompt and chooses the letter with the highest probability as the model's answer.

Some of our experiments require permuting the answer choice order (Table 1, A.8, A.9, and A.10), however, this can be confusing for questions where the answer choices are dependent on their position, such as *"D. All of the above."*, or reference other choices, such as *"C. Both A and B."*. To circumvent this dependency, we manually inspected and modified the questions from three subjects to ensure their answers are permutation-independent for a subset of our experiments. The modified subjects are college chemistry, college mathematics, and global facts.

For each variation introduced to the MCQ benchmarks, we calculate the change in accuracy (ΔAcc) and recall standard deviation (RStd) for each model. RStd measures the bias of a model to a particular answer choice by computing the standard deviation of recalls for each answer choice (Zheng et al., 2023). This metric quantifies how much the model favors particular positions for the correct answer choice. We typically observe whether RStd changes (ΔRStd) are significant across experimental settings.

To measure the change in ranking induced by an applied perturbation to a benchmark, we measure the normalized Kendall's τ distance between two rankings of n models (Kendall, 1938). Kendall's τ computes the number of swapped pairs between two rankings normalized by the total number of pairs $\frac{n(n-1)}{2}$. We report $k_\tau = \frac{\tau+1}{2}$, where $k_\tau = 1.0$ indicates total agreement between rankings, and $k_\tau = 0$ indicates complete disagreement by reversing the original rankings.

5 Results & Analysis

In this section, we highlight the major findings of our work and combine the results of multiple lines of experimentation (detailed in Section 3) into concise observations. Additional observations and complete experimental results can be found in the appendix (Section A.1).

5.1 MCQ benchmarks are not robust to perturbations

As shown in Figure 1, there exist perturbations that cause dramatic shifts in the order of models with respect to commonly accepted leaderboard rankings. We find a significant number of small perturbations demonstrate this effect, while other perturbations are more benign.

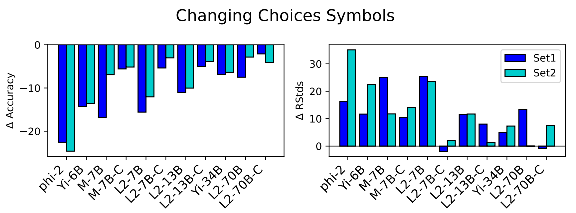

Figure 4: Change in accuracy and bias (RStd) on zero-shot MMLU after swapping answer choice symbols with two different sets of symbols (described in Section 3.1). While accuracy always decreased, most models exhibited even more selection bias with the new symbols. k_τ for Set1 and Set2 were 0.689 and 0.733 respectively

Sensitive perturbations Shuffling/changing the presented order of the choices, swapping choice symbols, and alternative scoring methods all cause major shifts to the rankings (determined by thresholding $k_\tau \leq 0.75$). For example, in a controlled experiment using a subset of MMLU, we randomly shuffle the answer choices presented to the models (Table 1). 5 out of 11 models change in ranking after the perturbation and k_τ drops to 0.564. A similar pattern is seen for perturbations like fixing the correct answer to a particular position (Table 2), replacing the default choice symbols with other sets (Figure 4), and alternative scoring methods (Figure 7).

Some models elicit this behavior much more strongly. For example, we observe that Yi-6b drops from 3rd place to 7th or 8th place under some benchmark perturbations in the group of 11 models we tested (namely, the rare symbol and cloze perturbations). Other models in the same size range are more stable (e.g., Mistral-7b, Llama2-7b), not shifting more than one or two ranks under all perturbations. The reasons for this are unclear but could indicate overfitting to aspects of the benchmark style. Since training data for these models is not public, it is difficult for us to verify this hypothesis.

Unsensitive perturbations Changes that have little effect on the model rankings are discussed in Section 5.4.

Model	Rank	Acc (ΔAcc)	RStd (ΔRStd)
phi-2	(7→7)	34.6 (-3)	14.2 (7.4)
Yi-6b	(3→9)	33.0 (-8.3)	11.9 (1.8)
Mistral-7b	(4→3)	40.0 (1.0)	9.8 (0.7)
Mistral-7b-Instruct	(8→8)	33.3 (-1.7)	16.7 (3.5)
Llama2-7b	(11→11)	24.3 (-5.0)	13.2 (-0.4)
Llama2-7b-chat	(9→10)	28.6 (-3.7)	27.7 (7.9)
Llama2-13b	(6→6)	37.0 (0.7)	22.7 (5.7)
Llama2-13b-chat	(9→5)	37.6 (6.0)	26.7 (0.0)
Yi-34b	(1→1)	45.0 (-5.0)	9.2 (-2.3)
Llama2-70b	(2→2)	40.3 (-1.7)	9.07 (-5.5)
Llama2-70b-chat	(5→4)	37.6 (0.3)	13.4 (-6.2)

Table 1: We show that model rankings can shift under shuffling of the order of answer choices. The largest change in rank is 5 positions (Yi-6b) followed by 4 positions (Llama2-13b-chat). This zero-shot experiment is done on a subset of MMLU subjects (college chemistry, college mathematics, and global facts) which we manually verified maintained correctness after shuffling answer choice order (i.e. did not contain cross references between answer choices). $k_\tau = 0.564$ for this experiment, indicating a significant disagreement in rankings.

5.2 Revisiting selection bias: token bias vs. position bias

Prior and concurrent work finds that LLMs answering MCQs are highly sensitive to the order that choices are presented (Pezeshkpour and Hruschka, 2023; Robinson et al., 2023) (position bias) as well as the symbols used as choice IDs (Zheng et al., 2023) (token bias). We find selection bias is apparent in **all** LLMs we test both in 0 and 5-shot setups, as shown in Tables 2 and A.6. This confirms earlier findings and highlights a major weakness of the current methods of evaluating LLMs on MCQs.

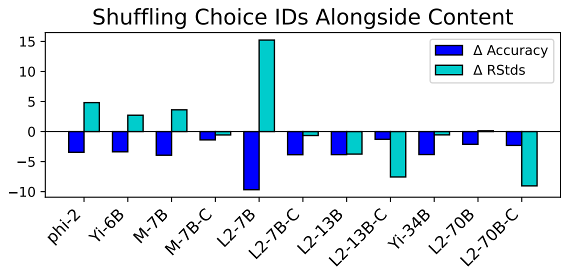

Figure 5: Accuracy and RStd change after randomly shuffling the order of the choices alongside their option IDs. Although (Zheng et al., 2023) use this experiment as evidence that position bias has minimal effect on selection bias, we find it inconclusive as variance in ΔRStd is large.

To disentangle these two sources of bias, we first measure the change in bias (measured by RStd) as we randomly shuffle the entire choice and symbols together, as performed in (Zheng et al., 2023). We find that simply shuffling entire choices is inconclusive in ruling out the effect of position bias (vs. token bias) as there is a wide variance in the bias change across LLMs (Figure 5, Table A.7). In light of this, we opt to isolate token bias from position bias by replacing the default symbols (A/B/C/D) with new/rare symbols from the LLM's vocabulary (without an implicit relative ordering) and shuffling them. This experiment, displayed in Figure 6 and Table A.8, shows that (i) LLMs always bias toward the symbols representing the choice IDs and (ii) even after shuffling the symbols, bias changes in unpredictable ways.

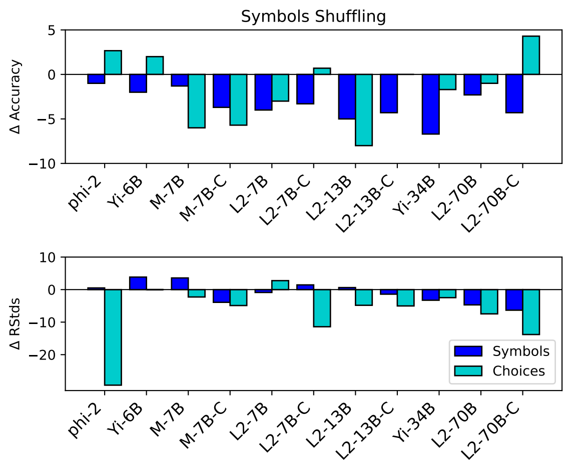

Figure 6: Using a set of rare symbols (Set2) we test two modes of shuffling answer choices: shuffling the symbols only (blue bars) and shuffling the answer choice text only while fixing the symbols set (cyan bars). Even using rare symbols, model selection bias changes unpredictably, indicating token and position bias are difficult to mitigate. This experiment was conducted using the three-subject subset of MMLU.

5.3 Another source of bias: scoring bias

Beyond the ordering of choices and the symbols associated with them, LLMs exhibit varying amounts of bias under the choice of scoring method for MCQs. We studied the three scoring methods described in Section 3.2: symbol scoring, cloze scoring, and hybrid scoring. Symbol scoring has become the dominant method for evaluating LLMs on MCQs, largely due to the high accuracy achieved by LLMs (Robinson et al., 2023). This, however, comes at the cost of high selection bias. Cloze scoring can essentially eliminate bias since the choices are never presented to the model, but LLMs tend to score poorly when using this method. This also does not reflect a true MCQ setting. Figure 7 and

Model	Baseline	A	B	C	D
`phi-2`	54.47	52.31 (-2.16)	56.53 (+2.07)	56.30 (+1.83)	50.19 (-4.28)
`Yi-6b`	61.12	62.53 (+1.41)	64.44 (+3.32)	58.59 (-2.53)	63.13 (+2.02)
`Mistral-7b`	59.56	52.19 (-7.38)	60.98 (+1.42)	63.84 (+4.27)	60.43 (+0.86)
`Mistral-7b-Instruct`	53.48	49.77 (-3.71)	54.67 (+1.18)	49.99 (-3.49)	57.74 (+4.26)
`Llama2-7b`	41.81	66.36 (+24.55)	30.40 (-11.42)	36.28 (-5.53)	23.37 (-18.44)
`Llama2-7b-chat`	46.37	30.84 (-15.53)	69.41 (+23.04)	50.05 (+3.68)	28.23 (-18.14)
`Llama2-13b`	52.08	35.82 (-16.26)	57.24 (+5.16)	68.65 (+16.57)	44.08 (-8.00)
`Llama2-13b-chat`	53.12	36.73 (-16.39)	56.72 (+3.60)	71.81 (+18.69)	42.63 (-10.49)
`Yi-34b`	73.38	66.16 (-7.22)	75.22 (+1.84)	78.07 (+4.69)	73.88 (+0.50)
`Llama2-70b`	65.44	56.47 (-8.97)	67.38 (+1.95)	69.92 (+4.48)	66.47 (+1.03)
`Llama2-70b-chat`	61.11	41.78 (-19.34)	62.24 (+1.13)	75.07 (+13.96)	57.71 (-3.41)
k_τ	-	0.455	0.527	0.527	0.855

Table 2: Performance on zero-shot MMLU when placing the correct answer at each possible position. All the LLMs tested showed a clear preference for specific positions/answer choice symbols, although the position varied among models and even in model families. These results corroborate the findings in (Zheng et al., 2023).

Table A.13 detail the results of these experiments.

Hybrid scoring, where cloze scoring is combined with a prompt that reveals all answer choices to the model, represents an acceptable balance between the two, reducing bias over symbol scoring on MMLU and ARC-C, as shown in Figure 7. In light of this, we recommend practitioners to replace symbol scoring with hybrid scoring to mitigate the effects of bias on model rankings.

5.4 Minor few-shot and prompt changes have little effect on benchmark rankings

We ran several experiments to assess the effect of the initial prompt on model performance and rankings. We find that changing the informativeness of in-context examples, e.g. providing irrelevant/trivial examples (Figure 8, Tables A.17-A.19) or examples from subjects other than the target subject (Figure 9, Tables A.24- A.25), slightly changes performance across models and reduces bias compared to zero-shot settings but does not change rankings drastically. This finding leads us to conclude that adding few shot examples to benchmark evaluations can help reduce, but not eliminate, leaderboard sensitivity.

We also experiment with removing subject information from instructions and ending the prompt with "Correct Answer:" instead of "Answer:" (Figure 10, Tables A.20-A.23). We see little changes

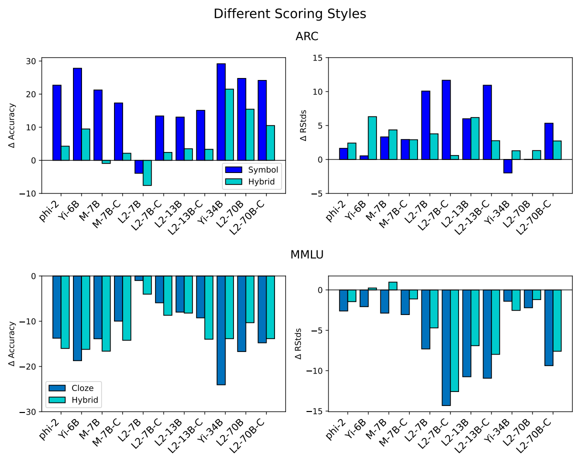

Figure 7: Comparing scoring method *{symbol, cloze, hybrid}* across two tasks, MMLU and ARC-Challenge. Note the baseline method for MMLU is **symbol** while the baseline method for ARC-C is **cloze**. The general trend for accuracy across models and tasks is symbol scoring (highest accuracies) followed by hybrid scoring/cloze depending on the model. The measured selection bias also follows this trend, with symbol scoring resulting in the highest bias across models.

Orignal prompt template

Instruction
"The following are multiple choice questions (with answers)"
Prompt
"*[question][choices]*Answer:"

Version 1

Prompt
trivial_examples = ["The capital of France is A. Paris. B. Berlin. C. Madrid. D. Rome Answer: A", ...]
"*[trivial_examples][question][choices]*Answer:"

Version 2

Prompt
trivial_examples = ["Which language the previous sentence is written in? A. Russian. B. English. C. Spanish. D. Japanese Answer: B", ...]
"*[trivial_examples][question][choices]*Answer:"

Version 3

Prompt
trivial_examples = ["The first word in this sentence is A. The. B. first. C. sentence. D. word Answer: A", ...]
"*[trivial_examples][question][choices]*Answer:"

Figure 8: Illustration of the three versions of the trivial examples.

($k_\tau > 0.9$) in these prompt modification experiments.

5.5 LLMs readily reference knowledge provided in-context (even if it is misleading)

In our study of in-context knowledge injection, we find that LLMs can, expectedly, read off answers

Orignal prompt template

Instruction
The following are multiple choice questions (with answers) about *[subject]*.
{{ few shot examples from *[subject]* }}
Prompt
*[question][choices]*Answer:

Modification 1: Remove subject name

Instruction
The following are multiple choice questions (with answers) ~~about *[subject]*.~~
{{ few shot examples from *different subject* }}
Prompt
*[question][choices]*Answer:

Modification 2: Mention various subjects

Instruction
"The following are 5 multiple choice questions (with answers) on various subjects, followed by a question about *[subject]*."
{{ few shot examples from *different subject* }}
Prompt
*[question][choices]*Answer:

Figure 9: Illustration of subject independent few-shot prompting experiment. We ensure that we do not sample from similar domains to the one being evaluated (e.g. sampling college mathematics few-shots for high school mathematics questions). (results are in Table A.24 & A.25).

Orignal prompt template

Instruction
"The following are multiple choice questions (with answers) about *[subject]*"
Prompt
"*[question][choices]*Answer:"

Experiment 1: Removing subject name

Instruction
"The following are multiple choice questions (with answers) ~~about *[subject]*~~"
Prompt
"*[question][choices]*Answer:"

Experiment 2: Prompt Modification

Instruction
"The following are multiple choice questions (with answers) about *[subject]*"
Prompt
"*[question][choices]*Correct Answer:"

Figure 10: Illustration of minor prompt modifications. Experiment 1 showcases the removal of the subject name from the instruction. Experiment 2 shows the prompt change by specifying "Correct Answer" instead of "Answer". (results are in table A.20, A.22, A.23)

to questions when the answer is provided in the context (Table A.27). However, when the question is answered incorrectly in the LLM's context (Table A.26), all models (regardless of size) are unable to reason correctly.

To investigate whether this behavior is due to in-context knowledge acquisition or "incorrect answer" pattern following, previous work has shown that smaller models tend to rely more on priors learned during the pretraining stage while larger models tend to be more influenced by knowledge given in-context (Wei et al., 2023). This result suggests that we would find that the magnitude of changes in accuracies for small models is smaller than that of larger models. However, we do not observe a significant effect to support that finding conclusively (when comparing A.1 with A.26).

This behavior is further studied in Wang et al. (2023); Xie et al. (2023); Min et al. (2022); Yoo et al. (2022) and indicates answer leakage in this way could affect benchmark results.

To test whether LLMs can infer subtler patterns in the few-shots examples, we fix all answers in the few-shot examples to each of the positions A/B/C/D. The results (Table 3) suggest that LLMs also bias their answers to these kinds of (potentially inadvertent) patterns in the context.

While we have not observed these vulnerabilities in current benchmarks, we highlight them here as (potential) sources of benchmark instability.

	5-shot Baseline	A	B	C	D
phi-2	56.77	36.67 (-20.11)	41.33 (-15.44)	40.67 (-16.11)	41.67 (-15.11)
Yi-6B	63.22	36.67 (-26.56)	36.33 (-26.89)	37.67 (-25.56)	39.33 (-23.89)
Mistral-7B	62.36	34.67 (-27.70)	41.33 (-21.03)	43.00 (-19.36)	40.33 (-22.03)
Llama-2-7b	45.88	22.00 (-23.88)	31.00 (-14.88)	30.67 (-15.22)	34.33 (-11.55)

Table 3: Results of fixing the 5 few-shot example answers to positions A/B/C/D on one model from each family, averaged over 3 selected subjects. We can see that performance drops across all cases/models, suggesting that models refer to subtle patterns in the context while answering. Full results are reported in Table A.28

6 Related Work

Benchmarks for the evaluation of LLMs (Chang et al., 2023) such as MMLU (Hendrycks et al., 2020), HELM (Liang et al., 2023), and BigBench (Suzgun et al., 2022) have seen widespread adoption recently. Depending on the ability that is being assessed (e.g., language generation, knowledge understanding, complex reasoning) some benchmarks are designed in the form of close-ended problems like MCQs. To facilitate comparisons among LLMs, a number of leaderboards aggregating these benchmarks have been established, such as the OpenLLM Leaderboard (Beeching et al., 2023) and OpenCompass (Contributors, 2023).

However, issues with the leaderboards and the underlying benchmarks have emerged. In a case study, Deng et al. (2023) discovered contamination/leakage of the MMLU benchmark in the training sets of multiple models. A significant portion of models memorized benchmark questions and was able to perfectly reconstruct the removed part of some benchmark questions or answers. For instance, GPT-4 correctly completed the questions in 29% of the prompts with URL hinting.

Even under the assumption of uncontaminated data, the performance of models on the underlying benchmarks are not robust to minor perturbations. Pezeshkpour and Hruschka (2023) showed that specific orderings of MMLU answer choices resulted in up to $\pm 30\%$ deviations in GPT-4 performance on various subjects. Similarly, Zheng et al. (2023) demonstrate that models are biased to certain answer letters. On `llama-30B`, they showed a 27% difference in MMLU accuracy by forcing all correct answers to either position `A` or `D`. As well, (Robinson et al., 2023) find that the accuracy of LLMs improve (without regard to bias) when evaluating using a pure multiple choice question style vs a cloze question answering style.

While prior work has highlighted weaknesses in LLMs themselves (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023), evaluation method (Robinson et al., 2023), or the contents of benchmarks (Dehghani et al., 2021) in our work we thoroughly study the effects these factors have on existing leaderboards and demonstrate where leaderboards lack robustness.

7 Conclusion

Building robust leaderboards is a major challenge for the community, as leaderboards help practitioners select the best methods and models for continued research. Given this importance, it is critical to address the breakdown of existing leaderboards to the slight perturbations we demonstrated in our work. In addition to building our understanding of the causes of this sensitivity (e.g., bias in LLMs and bias in scoring methods), future work should aim to adopt and design benchmark practices that avoid these pitfalls.

8 Limitations

The limitations of our work fall into two main categories: (i) understanding the causes of LLM bias and (ii) our limited success at overcoming leaderboard sensitivity.

To explain LLM bias, we attempted to design experiments that isolate each source of bias under MCQ but were unable to quantify the relative effects of bias or conclude why they occur. This was further complicated by our inability to access the pretraining datasets of the LLMs to rule out benchmark contamination. Future work in this direction will most likely require tools from interpretability research (e.g. mechanistic interpretability).

One of our main contributions was to highlight where MCQ-based leaderboards fail to deliver stable rankings. Although we succeeded in showing this, we were unable to demonstrate a robust solution to this problem. Our recommendation to, for example, use hybrid scoring methods is still not completely robust to perturbations.

9 Potential Risks

In this work, we do not present a new leaderboard. There is a risk that Figure 1 is interpreted as a leaderboard or used for model selection. Our intention was to demonstrate how minor changes can affect model rankings.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Anthropic. 2023. Anthropic. model card and evaluations for claude models.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

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*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018a. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018b. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Google Deepmind. 2023. Gemini: A family of highly capable multimodal models.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.

Kawin Ethayarajh and Dan Jurafsky. 2021. Utility is in the eye of the user: A critique of nlp leaderboards.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Séverin Lions, Carlos Monsalve, Pablo Dartnell, María Inés Godoy, Nora Córdova, Daniela Jiménez, María Paz Blanco, Gabriel Ortega, and Julie Lemarié. 2021. The position of distractors in multiple-choice test items: The strongest precede the weakest. In *Frontiers in Education*, volume 6, page 731763. Frontiers.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

OpenAI. 2023. Gpt-4 technical report.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv e-prints*, pages arXiv–2309.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A.1 Appendix

We present a comprehensive collection of tables containing the results of all our experiments. The often complex nature of the observed behavior warrants a closer look that may inspire novel interpretations for future studies. We believe providing these detailed results will help researchers conduct further analysis and generate hypotheses to help drive research in LLM-benchmarking robustness forward.

A.1.1 Baselines

This section lists the baselines referenced in different experiments throughout the paper.

Model	Acc 0shot	RStd 0shot	Acc 5shot	RStd 5shot
phi-2	54.47	4.01	56.77	2.65
Yi-6B	61.12	3.57	63.23	2.54
Mistral-7B	59.56	4.13	62.36	1.64
Mistral-7B-Instruct	53.48	4.58	53.95	4.78
Llama-2-7b	41.81	8.49	45.88	8.92
Llama-2-7b-chat	46.37	16.11	47.22	12.15
Llama-2-13b	52.08	12.04	55.06	4.42
Llama-2-13b-chat	53.12	12.80	53.53	8.32
Yi-34B	73.38	5.17	76.39	2.16
Llama-2-70b	65.44	3.20	68.78	1.56
Llama-2-70b-chat	61.11	10.95	63.17	8.06

Table A.1: The baseline accuracies and RStd values for the original MMLU implementation which uses the Symbols scoring style mentioned in section 3.2. All the models performed better in five-shot settings; the highest model was Yi-34B model in both settings.

Model	Acc 0shot	RStd 0shot	Acc 5shot	RStd 5shot
phi-2	54.096	2.558	58.874	2.509
Yi-6B	50.512	2.114	55.034	0.737
Mistral-7B	53.584	2.578	59.556	1.037
Mistral-7B-Instruct	52.048	1.443	54.778	2.022
Llama-2-7b	46.331	4.094	53.072	0.837
Llama-2-7b-hf	44.283	2.175	51.877	1.399
Llama-2-13b	48.976	2.923	56.997	0.799
Llama-2-13b-chat	50.256	1.841	57.594	2.991
Yi-34B	61.519	2.537	64.505	1.48
Llama-2-70b	57.253	2.657	66.126	1.926
Llama-2-70b-chat	54.266	1.505	64.078	2.084

Table A.2: The baseline accuracies and RStd values for ARC-C using the Cloze scoring style mentioned in section 3.2 which is considered as the original ARC-C implementation. As the table shows, the RStd values are relatively low in both settings. Yi-34B has the highest values on zero-shot while Llama-2-70b was the highest on five-shots

A.1.2 Answer choice format and ordering

The following tables provide details on the choice formatting manipulation on the three selected MMLU subjects.

Model	Acc 0shot	RStd 0shot	Acc 5shot	RStd 5shot
phi-2	37.67	6.78	41.00	5.02
Yi-6B	41.33	10.17	40.67	14.07
Mistral-7B	39.0	9.17	41.00	12.08
Mistral-7B-Instruct	35.0	13.31	36.00	15.75
Llama-2-7b	29.33	13.64	33.33	17.69
Llama-2-7b-chat	32.33	19.83	33.33	21.39
Llama-2-13b	36.33	17.05	35.67	13.85
Llama-2-13b-chat	31.67	26.78	32.67	24.69
Yi-34B	50.00	11.49	49.33	9.35
Llama-2-70b	42.00	14.58	44.67	6.21
Llama-2-70b-chat	37.33	19.63	41.00	18.46

Table A.3: The selected three domains baseline average results on zero-shot and five-shot using Symbols scoring style on MMLU. MMLU mostly uses this scoring style. This baseline was utilized in most experiments to analyze and comprehend the influence of each experiment compared with this baseline in the selected domains subset (it was used in A.4, A.5 and 1).

Model	Task Acc (ΔAcc)	Task RStd (ΔRStd)
phi-2	26.33(-11.3)	41.85 (35.0)
Yi-6B	32.60 (-8.7)	22.80 (12.7)
Mistral-7B	35.30 (-3.7)	18.79 (9.6)
Mistral-7B-Instruct	34.00 (-1.0)	26.90 (13.7)
Llama-2-7b	29.60 (0.3)	25.80 (12.2)
Llama-2-7b-chat	31.30 (-1.0)	27.00 (7.2)
Llama-2-13b	34.30 (-2.0)	26.10 (9.1)
Llama-2-13b-chat	34.00 (2.3)	21.90 (-4.8)
Yi-34B	42.60 (-7.3)	22.7 (11.3)
Llama-2-70b	39.60 (-2.3)	15.10 (0.5)
Llama-2-70b-chat	36.00 (-1.3)	29.50 (10.0)
$k_\tau = 0.527$		

Table A.4: The baseline average zero-shot results for the selected domains using symbols Set2 which replaced the A/B/C/D choices symbols with œ/§/Ze (Cyrillic)/ü as options as described in section 3.1 (it was used as a baseline in A.9 and A.8). The deltas are calculated compared with A.3. In this particular experiment, all models encountered a decline in accuracy, coupled with a significant increase in RStds values, except Llama-13b-chat.

A.1.3 Prompt and scoring modifications

The following tables provide results on the effect of different scoring styles of MCQs task on MMLU and ARC-C.

A.1.4 In-context Knowledge Manipulation

This section provides the results from experimentation on in-context manipulation.

A.1.5 MMLU Overview

Model	Acc 0shot (ΔAcc)	RStd 0shot (ΔRStd)	Acc 5shot (ΔAcc)	RStd 5shot (ΔRStd)
phi-2	28.3 (-9.3)	6.0 (-0.7)	34.6 (-6.3)	5.7 (-1.04)
Yi-6B	35.0 (-6.3)	11.5 (1.4)	39.0 (-1.7)	13.5 (-0.6)
Mistral-7B	34.3 (-4.7)	10.7 (1.6)	44.0 (3.0)	16.3 (4.2)
Mistral-7B-Instruct	35.0 (0.0)	14.0 (0.7)	38 (2.0)	15.7 (0.0)
Llama-2-7b	31.3 (2.0)	12.6 (-1.0)	32.6 (-0.7)	16.9 (-0.7)
Llama-2-7b-chat	27.0 (-5.3)	12.5 (-7.3)	32.6 (-0.7)	13.3 (-8.0)
Llama-2-13b	37.0 (0.7)	14.0 (-3.0)	40.0 (4.3)	15.7 (1.9)
Llama-2-13b-chat	33.0 (1.3)	9.1 (-17.7)	37.6 (5.0)	17.33 (-7.4)
Yi-34B	46.6 (-3.3)	12.8 (1.4)	47.6 (-1.7)	10.1 (0.8)
Llama-2-70b	41.3 (-0.7)	10.5 (-4.0)	49.0 (4.3)	10.3 (4.2)
Llama-2-70b-chat	39.3 (2.0)	7.8 (-11.8)	42.6 (1.7)	11.9 (-6.5)
k_τ	0.564		0.6	

Table A.5: The average zero-shot results on the three selected domains baseline using the Hybrid style mentioned in section 3.2. The deltas are compared with A.3 where the Rstd values exhibited a decrease and the accuracies remained relatively stable, except phi-2, which demonstrated the most significant decline in accuracy.

Model	**Baseline**	**A**	**B**	**C**	**D**
phi-2	54.47	57.33 (+2.87)	44.00 (-10.47)	25.00 (-29.47)	32.33 (-22.13)
Yi-6B	61.12	49.67 (-11.45)	23.67 (-37.45)	18.33 (-42.78)	44.67 (-16.45)
Mistral-7B	59.56	77.00 (+17.44)	46.33 (-13.23)	48.33 (-11.23)	68.00 (+8.44)
Mistral-7B-Instruct	53.48	78.33 (+24.85)	42.33 (-11.15)	18.67 (-34.82)	49.33 (-4.15)
Llama-2-7b	41.81	79.00 (+37.19)	57.33 (+15.52)	24.67 (-17.14)	23.67 (-18.14)
Llama-2-7b-chat	46.37	16.67 (-29.70)	66.33 (+19.97)	38.67 (-7.70)	14.33 (-32.04)
Llama-2-13b	52.08	33.67 (-18.41)	37.33 (-14.75)	45.33 (-6.75)	39.33 (-12.75)
Llama-2-13b-chat	53.12	20.00 (-33.12)	23.00 (-30.12)	61.33 (+8.21)	15.67 (-37.45)
Yi-34B	73.38	59.00 (-14.38)	45.67 (-27.71)	53.67 (-19.71)	48.00 (-25.38)

Table A.6: Performance on five-shot MMLU when placing the correct answer at each possible position, for both the examples and the question asked. Similar to the zero-shot case mentioned in Section 5, all the LLMs tested showed a clear preference for specific positions/answer choice symbols, although the position varied among models and even in model families.

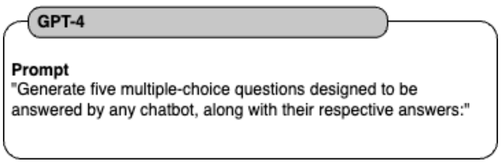

Figure A.1: Illustration of the prompt that was used to generate the trivial examples version 1 using GPT4.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	51.01	-3.45	8.82	4.82
Yi-6B	57.75	-3.37	6.29	2.72
Mistral-7B	55.63	-3.94	7.75	3.62
Mistral-7B-Instruct	52.09	-1.39	4.02	-0.57
Llama-2-7b	32.13	-9.68	23.72	15.23
Llama-2-7b-chat	42.52	-3.85	15.45	-0.66
Llama-2-13b	48.24	-3.84	8.29	-3.75
Llama-2-13b-chat	51.83	-1.29	5.24	-7.56
Yi-34B	69.56	-3.82	4.62	-0.55
Llama-2-70b	63.32	-2.12	3.33	0.13
Llama-2-70b-chat	58.80	-2.31	1.91	-9.04

Table A.7: Reproducing shuffling ablation experiment from (Zheng et al., 2023). Randomly shuffling the order in which the options are presented. Surprisingly, all models demonstrated a decrease in accuracy, suggesting a lack of decisiveness in the experiment. However, these variations indicate a potential bias in the benchmark.

Model	Task Acc (ΔAcc)	Task RStd (ΔRStd)
phi-2	25.33 (-12.33)	42.35 (35.57)
Yi-6B	30.66 (-2.0)	26.68 (3.8)
Mistral-7B	34.00 (-1.3)	22.37 (3.6)
Mistral-7B-Instruct	30.33 (-3.7)	23.08 (-3.9)
Llama-2-7b	25.66 (-4.0)	24.98 (-0.9)
Llama-2-7b-chat	28.00 (-3.3)	28.49 (1.4)
Llama-2-13b	29.33 (-5.0)	26.75 (0.6)
Llama-2-13b-chat	29.66 (-4.3)	20.58 (-1.4)
Yi-34B	36.00 (-6.7)	19.48 (-3.3)
Llama-2-70b	37.33 (-2.3)	10.41 (-4.7)
Llama-2-70b-chat	31.66 (-4.3)	23.25 (-6.3)
$k_\tau = 0.564$		

Table A.8: The average zero-shot results on the three selected domains using Symbols Set2 (mentioned in section 3.1 and shuffling the choices while fixing the order of the choices symbols. The deltas are measured compared with A.4. As displayed in the table, mostly all the models faced a decrease in accuracy while the RStds values were not decisive. The most affected model in this experiment was phi-2.

Model	Task Avg Acc (ΔAcc)	Task Avg RStd (ΔRStd)
phi-2	29.00(-8.6)	12.4 (5.6)
Yi-6B	34.67 (2.0)	22.84 (0.0)
Mistral-7B	29.33 (-6.0)	16.52 (-2.3)
Mistral-7B-Instruct	28.33 (-5.7)	22.10 (-4.9)
Llama-2-7b	26.67 (-3.0)	28.62 (2.8)
Llama-2-7b-chat	32.00 (0.7)	15.64 (-11.4)
Llama-2-13b	26.33 (-8.0)	21.31 (-4.8)
Llama-2-13b-chat	34.00 (0.0)	16.97 (-5.0)
Yi-34B	41.00 (-1.7)	20.28 (-2.5)
Llama-2-70b	38.67 (-1.0)	7.65 (-7.5)
Llama-2-70b-chat	40.33 (4.3)	15.78 (-13.8)
$k_\tau = 0.455$		

Table A.9: The average zero-shot results on the three selected domains using Symbols Set2 mentioned in section 3.1. This experiment focused on shuffling the symbols while maintaining the original listing order of the choices. Compared with A.4,Most of the models were impacted in terms of accuracy and RStds, indicating that randomization affects the models even after changing the symbols.

Model	Acc 0shot (ΔAcc)	RStd 0shot (ΔRStd)	Acc 5shot (ΔAcc)	RStd 5shot (ΔRStd)
phi-2	30.6 (2.3)	12.8 (6.8)	32.6(-2)	13.6 (7.8)
Yi-6B	30.3 (-4.7)	12.0 (0.5)	34.3 (-4.7)	11.4 (-2.1)
Mistral-7B	31.6 (-2.7)	12.5 (1.8)	39 (-5)	11.1 (-5.2)
Mistral-7B-Instruct	32.66 (-2.3)	11.18 (-2.9)	37 (-1)	7.94 (-7.8)
Llama-2-7b	28.6 (-2.7)	11.4 (-1.2)	33.3 (0.7)	15.1 (-1.8)
Llama-2-7b-chat	29.3 (2.3)	16.2 (3.7)	35 (2.3)	16.6 (3.3)
Llama-2-13b	35.3 (-1.7)	10.1 (-3.9)	37.6 (-2.3)	12.1 (-3.6)
Llama-2-13b-chat	29.6 (-3.3)	10.9 (1.9)	35.3 (-2.3)	17.0 (-0.3)
Yi-34B	43 (-3.7)	5.4 (-7.4)	48.3 (0.7)	11.7 (1.6)
Llama-2-70b	40 (-1.3)	9.0 (-1.5)	48 (-1)	10.5 (0.1)
Llama-2-70b-chat	35 (-4.3)	11.1 (3.4)	41.3 (-1.3)	6.8 (-5.1)
k_τ	0.527		0.382	

Table A.10: The selected domains results after randomizing the choices using Hybrid style mentioned in section 3.2, the deltas are calculated from this table A.5 where it showed more consistency compared to the results of other randomization settings (1,A.8, and A.9).

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	31.92	-22.55	20.23	16.22
Yi-6B	46.87	-14.25	15.24	11.67
Mistral-7B	42.68	-16.88	29.07	24.94
Mistral-7B-Instruct	47.90	-5.58	15.06	10.48
Llama-2-7b	26.23	-15.58	33.78	25.29
Llama-2-7b-chat	41.01	-5.36	14.17	-1.94
Llama-2-13b	41.05	-11.03	23.54	11.50
Llama-2-13b-chat	48.09	-5.03	20.82	8.02
Yi-34B	66.56	-6.82	10.13	4.96
Llama-2-70b	57.94	-7.50	16.52	13.32
Llama-2-70b-chat	59.00	-2.11	10.09	-0.86
$k_\tau = 0.6$				

Table A.11: The zero-shot results of MMLU on Symbols Set1 mentioned in section 3.1. All of the models demonstrated reduced accuracies, while most of them showed an increase in RStds values compared with A.1.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	29.85	-24.62	39.10	35.09
Yi-6B	47.58	-13.54	26.09	22.52
Mistral-7B	52.63	-6.94	15.87	11.74
Mistral-7B-Instruct	48.33	-5.15	18.70	14.12
Llama-2-7b	29.76	-12.05	32.09	23.60
Llama-2-7b-chat	43.34	-3.03	18.20	2.09
Llama-2-13b	42.06	-10.02	23.75	11.70
Llama-2-13b-chat	49.23	-3.89	14.07	1.28
Yi-34B	67.03	-6.35	12.48	7.31
Llama-2-70b	62.60	-2.84	3.21	0.01
Llama-2-70b-chat	57.01	-4.10	18.53	7.59
$k_\tau = 0.636$				

Table A.12: The zero-shot results of MMLU on Symbols Set2 mentioned in section 3.1. Compared with the original MMLU implementation that used A/B/C/D as symbols(A.1), the majority of models in this experiment had notably lower accuracies while the RStd values increased.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	40.714	-13.751	1.398	-2.607
Yi-6B	42.40	-18.72	1.49	-2.08
Mistral-7B	45.69	-13.87	1.26	-2.87
Mistral-7B-Instruct	43.51	-9.98	1.53	-3.05
Llama-2-7b	40.81	-1.00	1.19	-7.30
Llama-2-7b-chat	40.44	-5.93	1.79	-14.32
Llama-2-13b	44.09	-7.99	1.29	-10.75
Llama-2-13b-chat	43.87	-9.25	1.86	-10.93
Yi-34B	49.33	-24.05	3.76	-1.41
Llama-2-70b	48.74	-16.70	0.99	-2.21
Llama-2-70b-chat	46.34	-14.77	1.57	-9.38
$k_\tau = 0.527$				

Table A.13: The zero-shot results of MMLU using the Cloze style mentioned in 3.2. As anticipated, employing this style led to significantly low RStd values compared with the Symbols scoring style in table A.1, but it also had a considerable impact on accuracy, resulting in a noticeable decrease in most models.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	38.47	-16.01	2.55	-1.45
Yi-6B	44.90	-16.22	3.80	0.23
Mistral-7B	42.94	-16.62	5.09	0.96
Mistral-7B-Instruct	39.27	-14.21	3.46	-1.12
Llama-2-7b	37.79	-4.02	3.79	-4.70
Llama-2-7b-chat	37.68	-8.69	3.52	-12.58
Llama-2-13b	43.88	-8.20	5.14	-6.91
Llama-2-13b-chat	39.14	-13.98	4.82	-7.98
Yi-34B	59.52	-13.86	2.63	-2.54
Llama-2-70b	55.11	-10.33	2.00	-1.20
Llama-2-70b-chat	47.26	-13.85	3.35	-7.60
$k_\tau = 0.709$				

Table A.14: The zero-shot results of MMLU using the Hybrid style mentioned in 3.2. This style resulted in decreased accuracy but demonstrated more stability and lower RStd values when comparing it with the Symbols scoring style A.1. This style may help reduce the selection and token bias seen in prior experiments.

Model	Task Acc (ΔAcc)	Task RStd (ΔRStd)
phi-2	76.8 (22.7)	4.2 (1.6)
Yi-6B	78.3 (27.8)	2.6 (0.5)
Mistral-7B	74.8 (21.2)	5.9 (3.3)
Mistral-7B-Instruct	69.3 (17.3)	4.3 (2.9)
Llama-2-7b	42.4 (-3.9)	14.1 (10.0)
Llama-2-7b-chat	57.6 (13.3)	13.8 (11.6)
Llama-2-13b	62.0 (13.0)	8.9 (6.0)
Llama-2-13b-chat	65.3 (15.1)	12.7 (10.9)
Yi-34B	90.7 (29.1)	0.5 (-1.9)
Llama-2-70b	81.9 (24.7)	2.6 (0.025)
Llama-2-70b-chat	78.4 (24.1)	6.8 (5.3)
$k_\tau = 0.855$		

Table A.15: The table displays the results of zero-shot on ARC-C with Symbols scoring style mentioned in 3.2. Compared with A.2, all models, except Llama-2-7b, showed higher accuracies. An increase in Rstds values was observed, particularly in the Llama-2 7b, 7b-chat, and 13b models. This proves that if we provide choices in the prompt, models will perform better.

Model	Task Acc (ΔAcc)	Task RStd (ΔRStd)
phi-2	58.4 (4.3)	4.9 (2.4)
Yi-6B	59.9 (9.4)	8.4 (6.3)
Mistral-7B	52.6 (-0.9)	6.9 (4.3)
Mistral-7B-Instruct	54.1 (2.1)	4.3 (2.9)
Llama-2-7b	38.7 (-7.5)	7.8 (3.7)
Llama-2-7b-chat	46.6 (2.3)	2.7 (0.5)
Llama-2-13b	52.4 (3.4)	9.1 (6.1)
Llama-2-13b-chat	53.5 (3.3)	4.6 (2.7)
Yi-34B	83.0 (21.5)	3.8 (1.2)
Llama-2-70b	72.6 (15.4)	3.9 (1.3)
Llama-2-70b-chat	64.7 (10.4)	4.2 (2.7)
$k_\tau = 0.782$		

Table A.16: The zero-shot results of ARC-C using the Hybrid style discussed in 3.2. In some models, it exhibits higher accuracy than the baseline (Table A.2) and more stable RStd values (compared to A.15). The deltas are calculated using this table A.2.

Model	Task acc	ΔAcc
phi-2	54.21	-0.26
Yi-6B	60.11	-1.00
Mistral-7B	58.45	-1.11
Mistral-7B-Instruct	51.14	-2.34
Llama-2-7b	42.77	0.96
Llama-2-7b-chat	46.35	-0.02
Llama-2-13b	51.72	-0.36
Llama-2-13b-chat	50.94	-2.18
Yi-34B	72.28	-1.10
Llama-2-70b	65.25	-0.18
Llama-2-70b-chat	59.79	-1.32
$k_\tau = 0.927$		

Table A.17: Trivial examples few-shot results using the version 1 examples with respect to zero-shot baseline accuracy.

Model	Task acc	ΔAcc
phi-2	53.18	-1.28
Yi-6B	60.28	-0.84
Mistral-7B	59.41	-0.15
Mistral-7B-Instruct	50.95	-2.53
Llama-2-7b	43.52	1.71
Llama-2-7b-chat	46.82	0.46
Llama-2-13b	52.51	0.44
Llama-2-13b-chat	51.84	-1.27
Yi-34B	72.29	-1.09
Llama-2-70b	65.13	-0.31
Llama-2-70b-chat	60.28	-0.83
$k_\tau = 0.891$		

Table A.18: Trivial examples few-shot results with version 2 examples with respect to zero-shot baseline accuracy.

Model	Task acc	ΔAcc
phi-2	53.22	-1.25
Yi-6B	60.46	-0.66
Mistral-7B	59.27	-0.30
Mistral-7B-Instruct	50.58	-2.91
Llama-2-7b	44.47	2.66
Llama-2-7b-chat	46.90	0.53
Llama-2-13b	52.24	0.16
Llama-2-13b-chat	51.88	-1.24
Yi-34B	73.16	-0.22
Llama-2-70b	65.42	-0.02
Llama-2-70b-chat	60.02	-1.09
$k_\tau = 0.891$		

Table A.19: Trivial examples few-shot results with version 3 examples, with respect to zero-shot baseline accuracy.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	53.92	-0.54	4.07	0.07
Yi-6B	60.80	-0.31	3.43	-0.14
Mistral-7B	59.02	-0.54	3.73	-0.40
Mistral-7B-Instruct	53.29	-0.19	4.74	0.16
Llama-2-7b	41.80	-0.01	4.51	-3.99
Llama-2-7b-chat	46.68	0.31	14.93	-1.17
Llama-2-13b	51.92	-0.16	12.05	0.00
Llama-2-13b-chat	53.27	0.15	12.83	0.03
Yi-34B	72.94	-0.44	5.52	0.35
Llama-2-70b	64.83	-0.60	2.81	-0.40
Llama-2-70b-chat	61.14	0.03	10.94	-0.00
k_τ=0.964				

Table A.20: Zero-shot results of removing the subject name from the prompt. (experiment 1 from figure 10). There are minimal changes in performance when applying this perturbation.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	54.21	-0.26	4.21	0.20
Yi-6B	61.06	-0.06	2.33	-1.24
Mistral-7B	60.16	0.60	2.08	-2.06
Mistral-7B-Instruct	53.67	0.19	4.03	-0.56
Llama-2-7b	41.42	-0.39	15.05	6.56
Llama-2-7b-chat	47.22	0.85	14.22	-1.88
Llama-2-13b	53.46	1.38	10.46	-1.59
Llama-2-13b-chat	53.20	0.08	11.09	-1.71
Yi-34B	73.64	0.26	5.68	0.51
Llama-2-70b	65.48	0.04	3.51	0.30
Llama-2-70b-chat	61.20	0.09	10.31	-0.63
k_τ=0.927				

Table A.21: Zero-shot results on adding the “Correct" token in the prompt. (experiment 2 from figure 10). There are minimal changes in performance when applying this perturbation.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	56.69	-0.08	2.57	-0.08
Yi-6B	63.69	0.46	3.22	0.68
Mistral-7B	62.60	0.23	2.98	1.33
Mistral-7B-Instruct	53.99	0.04	4.62	-0.16
Llama-2-7b	45.80	-0.09	8.75	-0.17
Llama-2-7b-chat	47.42	0.20	12.03	-0.11
Llama-2-13b	55.47	0.41	5.04	0.62
Llama-2-13b-chat	53.58	0.05	8.32	0.00
Yi-34B	76.36	-0.02	2.14	-0.02
Llama-2-70b	68.71	-0.07	1.63	0.06
Llama-2-70b-chat	63.14	-0.03	8.49	0.43
k_τ=1.0				

Table A.22: Few-shot results of removing the subject name from the prompt. (experiment 1 from figure 10). There are minimal changes in performance when applying this perturbation.

Model	Task acc	ΔAcc	Task RStd	ΔRStd
phi-2	56.57	-0.21	3.95	1.30
Yi-6B	63.20	-0.03	4.01	1.47
Mistral-7B	62.79	0.43	3.51	1.87
Mistral-7B-Instruct	53.85	-0.10	5.51	0.73
Llama-2-7b	46.21	0.33	7.14	-1.78
Llama-2-7b-chat	47.48	0.26	10.42	-1.73
Llama-2-13b	55.18	0.11	4.79	0.37
Llama-2-13b-chat	53.75	0.23	6.58	-1.74
Yi-34B	75.98	-0.41	1.71	-0.46
Llama-2-70b	69.10	0.32	0.83	-0.73
Llama-2-70b-chat	62.86	-0.31	7.20	-0.86
k_τ=1.0				

Table A.23: Few-shot results on adding the “Correct" token in the prompt. (experiment 2 from figure 10). There are minimal changes in performance when applying this perturbation.

Model	Task acc	ΔAcc
phi-2	54.94	-1.84
Yi-6B	61.51	-1.72
Mistral-7B	59.56	-2.80
Mistral-7B-Instruct	51.72	-2.24
Llama-2-7b	44.10	-1.79
Llama-2-7b-chat	46.92	-0.30
Llama-2-13b	52.61	-2.46
Llama-2-13b-chat	52.63	-0.90
Yi-34B	73.89	-2.50
Llama-2-70b	66.26	-2.52
Llama-2-70b-chat	60.85	-2.31
k_τ = 0.927		

Table A.24: Subject independent five-shots example results with the first prompt. (follow Figure 9 for details). With few exceptions, most models exhibit a 2% drop from changing the few shots example domains. For models that are not fine-tuned, we noticed a performance that is halfway between the standard zero-shot and five-shot. Indicating that these models utilize the few shots for both formatting and knowledge domain information.

Model	Task acc	ΔAcc
phi-2	55.25	-1.52
Yi-6B	61.15	-2.08
Mistral-7B	59.68	-2.69
Mistral-7B-Instruct	52.12	-1.84
Llama-2-7b	44.12	-1.76
Llama-2-7b-chat	46.74	-0.48
Llama-2-13b	52.91	-2.16
Llama-2-13b-chat	52.19	-1.33
Yi-34B	73.62	-2.76
Llama-2-70b	66.06	-2.72
Llama-2-70b-chat	60.64	-2.53
$k_\tau = 0.964$		

Table A.25: Subject independent five-shot example results with the second prompt. (follow figure 9 for details). Changes in the initial prompt only result in negligible differences when compared to the first prompt in Table A.24.

Model	Task Acc 1-shot	Task Acc 5-shot
phi-2	71.778	92.366
Yi-6B	90.91	97.09
Mistral-7B	97.45	98.99
Mistral-7B-Instruct	98.64	99.25
Llama-2-7b	61.00	63.82
Llama-2-7b-chat	87.77	80.15
Llama-2-13b	96.60	99.79
Llama-2-13b-chat	87.02	92.69
Yi-34B	99.10	98.50
Llama-2-70b	93.45	99.09
Llama-2-70b-chat	98.25	93.86
k_τ	0.491	0.382

Table A.27: Results of the one-shot and five-shot MMLU in-context cheating experiment. Performance expectedly increases, indicating that models are readily able to "cheat" from the given few-shot examples in both five-shot and one-shot cases. However, no model achieved 100% accuracy, so we encourage the investigation of misclassified samples to validate their correctness.

Model	Task Acc 1-shot	Task Acc 5-shot
phi-2	33.59	13.91
Yi-6B	36.13	17.97
Mistral-7B	19.51	13.20
Mistral-7B-Instruct	10.71	4.59
Llama-2-7b	24.25	23.63
Llama-2-7b-chat	16.24	28.11
Llama-2-13b	12.76	4.50
Llama-2-13b-chat	31.49	26.30
Yi-34B	32.08	37.42
Llama-2-70b	26.27	21.54
Llama-2-70b-chat	26.26	37.23
k_τ	0.382	0.164

Table A.26: Providing the incorrect answer in-context. Performance drastically drops across the board, indicating that models are easily influenced by the answers given in-context, even when they are incorrect.

	5-shot Baseline	**A**	**B**	**C**	**D**
phi-2	56.77	36.67 (-20.11)	41.33 (-15.44)	40.67 (-16.11)	41.67 (-15.11)
Yi-6B	63.23	36.67 (-26.56)	36.33 (-26.89)	37.67 (-25.56)	39.33 (-23.89)
Mistral-7B	62.36	34.67 (-27.70)	41.33 (-21.03)	43.00 (-19.36)	40.33 (-22.03)
Mistral-7B-Instruct	53.95	32.67 (-21.29)	33.33 (-20.62)	30.67 (-23.29)	35.33 (-18.62)
Llama-2-7b	45.88	22.00 (-23.88)	31.00 (-14.88)	30.67 (-15.22)	34.33 (-11.55)
Llama-2-7b-chat	47.22	31.00 (-16.22)	30.67 (-16.56)	28.67 (-18.56)	31.00 (-16.22)
Llama-2-13b	55.06	35.33 (-19.73)	36.33 (-18.73)	37.67 (-17.40)	32.67 (-22.40)
Llama-2-13b-chat	53.53	31.67 (-21.86)	33.00 (-20.53)	34.67 (-18.86)	33.67 (-19.86)
Yi-34B	76.39	49.67 (-26.72)	49.33 (-27.05)	50.33 (-26.05)	48.67 (-27.72)
Llama-2-70b	68.78	42.67 (-26.11)	44.67 (-24.11)	43.33 (-25.45)	44.33 (-24.45)
Llama-2-70b-chat	63.17	40.33 (-22.84)	42.33 (-20.84)	42.00 (-21.17)	41.33 (-21.84)
k_τ	-	0.855	0.818	0.782	0.636

Table A.28: Results of fixing the five-shot example answers to positions A/B/C/D, averaged over the three selected subjects. We can see that performance drops across the board, suggesting that models get confused when there is a clear pattern in the correct answers of the few-shot examples.

subject	choice_A	choice_B	choice_C	choice_D	total
abstract_algebra	22	26	31	21	100
anatomy	25	34	45	31	135
astronomy	27	28	46	51	152
business_ethics	30	26	23	21	100
clinical_knowledge	57	71	58	79	265
college_biology	37	32	37	38	144
college_chemistry	20	21	18	41	100
college_computer_science	26	15	26	33	100
college_mathematics	21	23	25	31	100
college_medicine	36	36	43	58	173
college_physics	22	20	22	38	102
computer_security	28	24	30	18	100
conceptual_physics	62	76	48	49	235
econometrics	27	32	28	27	114
electrical_engineering	35	32	43	35	145
elementary_mathematics	79	97	101	101	378
formal_logic	36	25	19	46	126
global_facts	18	31	33	18	100
high_school_biology	55	79	78	98	310
high_school_chemistry	31	55	60	57	203
high_school_computer_science	25	23	33	19	100
high_school_european_history	36	40	47	42	165
high_school_geography	35	43	50	70	198
high_school_government_and_politics	38	40	44	71	193
high_school_macroeconomics	79	86	83	142	390
high_school_mathematics	57	71	71	71	270
high_school_microeconomics	50	55	50	83	238
high_school_physics	30	30	41	50	151
high_school_psychology	105	129	121	190	545
high_school_statistics	33	35	46	102	216
high_school_us_history	51	48	53	52	204
high_school_world_history	64	62	63	48	237
human_aging	70	84	45	24	223
human_sexuality	34	30	30	37	131
international_law	29	30	45	17	121
jurisprudence	28	32	25	23	108
logical_fallacies	36	40	49	38	163
machine_learning	35	32	27	18	112
management	18	26	20	39	103
marketing	68	60	60	46	234
medical_genetics	30	26	20	24	100
miscellaneous	186	225	212	160	783
moral_disputes	86	85	101	74	346
moral_scenarios	213	217	221	244	895
nutrition	69	70	77	90	306
philosophy	58	85	93	75	311
prehistory	70	86	95	73	324
professional_accounting	66	72	76	68	282
professional_law	377	367	415	375	1534
professional_medicine	50	55	45	122	272
professional_psychology	153	157	169	133	612
public_relations	24	38	23	25	110
security_studies	46	42	59	98	245
sociology	49	48	50	54	201
us_foreign_policy	28	21	25	26	100
virology	47	53	34	32	166
world_religions	55	36	50	30	171
total	3222	3462	3582	3776	14042

Table A.29: The MMLU subjects statistics in the test split.

subject	choice_A	choice_B	choice_C	choice_D	total
abstract_algebra	2	2	1	0	5
anatomy	0	1	2	2	5
astronomy	3	0	0	2	5
business_ethics	1	1	1	2	5
clinical_knowledge	1	2	2	0	5
college_biology	0	3	1	1	5
college_chemistry	2	0	1	2	5
college_computer_science	0	2	0	3	5
college_mathematics	0	2	1	2	5
college_medicine	2	1	1	1	5
college_physics	3	1	0	1	5
computer_security	1	1	1	2	5
conceptual_physics	2	1	2	0	5
econometrics	1	0	3	1	5
electrical_engineering	1	2	1	1	5
elementary_mathematics	1	4	0	0	5
formal_logic	0	1	3	1	5
global_facts	2	3	0	0	5
high_school_biology	1	0	1	3	5
high_school_chemistry	1	0	3	1	5
high_school_computer_science	0	1	3	1	5
high_school_european_history	2	1	1	1	5
high_school_geography	1	2	1	1	5
high_school_government_and_politics	1	0	2	2	5
high_school_macroeconomics	0	0	3	2	5
high_school_mathematics	0	1	2	2	5
high_school_microeconomics	0	1	2	2	5
high_school_physics	0	2	0	3	5
high_school_psychology	1	2	1	1	5
high_school_statistics	0	0	1	4	5
high_school_us_history	0	2	2	1	5
high_school_world_history	1	3	0	1	5
human_aging	1	2	2	0	5
human_sexuality	3	1	1	0	5
international_law	2	2	1	0	5
jurisprudence	2	0	0	3	5
logical_fallacies	0	1	2	2	5
machine_learning	1	1	2	1	5
management	3	0	1	1	5
marketing	1	1	0	3	5
medical_genetics	3	0	1	1	5
miscellaneous	1	2	2	0	5
moral_disputes	3	1	0	1	5
moral_scenarios	1	1	2	1	5
nutrition	1	1	2	1	5
philosophy	1	0	3	1	5
prehistory	2	1	1	1	5
professional_accounting	2	2	0	1	5
professional_law	2	2	0	1	5
professional_medicine	0	0	1	4	5
professional_psychology	2	0	0	3	5
public_relations	1	0	2	2	5
security_studies	0	2	2	1	5
sociology	1	3	1	0	5
us_foreign_policy	1	0	2	2	5
virology	2	1	1	1	5
world_religions	1	3	0	1	5
total	67	69	71	78	285

Table A.30: The MMLU subjects statistics in the dev split, which is a fixed set of questions per subject used in fewshots evaluation.