Metric assessment protocol in the context of answer fluctuation on MCQ tasks

Ekaterina Goliakova [1,2], Xavier Renard[1,2], Marie-Jeanne Lesot[1], Thibault Laugel[1,2], Christophe Marsala[1], Marcin Detyniecki[1,2,3]

[1]Sorbonne University, CNRS, LIP6, Paris, France
[2]AXA, Paris, France
[3]Polish Academy of Science, IBS PAN, Warsaw, Poland
Correspondence: ekaterina.goliakova@lip6.fr

Abstract

Using multiple-choice questions (MCQs) has become a standard for assessing LLM capabilities efficiently. A variety of metrics can be employed for this task. However, previous research has not conducted a thorough assessment of them. At the same time, MCQ evaluation suffers from *answer fluctuation*: models produce different results given slight changes in prompts. We suggest a metric assessment protocol in which evaluation methodologies are analyzed through their connection with fluctuation rates, as well as original performance. Our results show that there is a strong link between existing metrics and the answer changing, even when computed without any additional prompt variants. Using the protocol, the highest association is demonstrated by a novel metric, *worst accuracy*.

1 Introduction

Testing on question answering tasks has become standard in the LLM evaluation field (Rogers et al., 2021). However, assessing models' generations in these conditions is a complex task, due to inapplicability of "traditional" metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang et al., 2020), because of high variation between possible correct answers (He et al., 2022; Sulem et al., 2018). While human evaluation can be used instead, it can be costly (Elangovan et al., 2024) and subjective (Elangovan et al., 2025; Abeysinghe and Circi, 2024). Thus, multiple-choice questions (MCQ) benchmarks have prevailed in LLM evaluation, as a tool that maps all possible responses to a small set of options, with examples such as ARC (Clark et al., 2018), GPQA (Rein et al., 2024), and BigBench-Hard (Suzgun et al., 2022).

Using MCQ tasks allows for the exact matching of answers selected by models and correct ones and for the computation of standard metrics, such as accuracy (Gemma Team et al., 2024; OpenAI et al., 2023; Wang et al., 2024d). While reporting accuracy is typical, the metrics available for MCQ tasks include other possibilities. For instance, continuous metrics such as *probability mass* of correct answer can improve signal-to-noise ratio in evaluations (Madaan et al., 2024) or better track actual performance of models of different sizes during training (Schaeffer et al., 2023; Du et al., 2024). Additionally, new metrics were proposed specifically in the context of MCQ evaluation (e.g. Pezeshkpour and Hruschka, 2024; Zheng et al., 2024). However, previous work has not provided a thorough comparative analysis of these metrics.

In addition, prior research (Pezeshkpour and Hruschka, 2024; Gupta et al., 2024; Li and Gao, 2024; Zheng et al., 2024; Tjuatja et al., 2024) indicates that LLMs are sensitive to changes in MCQ options order: it is possible to elicit a different response from a model simply by rearranging the proposed answers. The phenomenon of LLMs producing different answers given semantically insignificant prompt changes can be called *answer fluctuation* (Wei et al., 2024) or answer floating (Wang et al., 2024b).

A deep understanding of answer fluctuation is crucial since LLMs' reliability remains a concern, especially in sensitive domains (Khatun and Brown, 2023; Amiri-Margavi et al., 2024; Naik, 2024). Nevertheless, discovering all cases of fluctuation leads to significantly higher computation costs, due to the necessity of testing multiple prompts.

We propose to use this factor in order to compare metrics available for the evaluation of MCQ tasks. In particular, we perform the costly calculation of models' responses fluctuation on all possible permutations and then compare those results with metrics computed on smaller subsets of permutations, assessing if any of the metrics could be used as a cost-efficient proxy for the *full fluctuation rates* (computed on all permutations), without losing the

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 302–319

information about the original performance. Our contributions can be summarized as follows:

1. Compilation and formalization of existing metrics used for estimating LLMs' performance on MCQ benchmarks (Section 3).

2. Proposition of a novel metric for MCQ evaluation (Section 3.4).

3. Introduction of a metric assessment protocol in which we analyze how well a given metric correlates with full fluctuation rates, as well as the original accuracy of the model (Section 4).

4. Application of the protocol to the results of 10 models on 17 tasks (Section 5).

We find that most metrics strongly correlate with the full fluctuation rates, even when calculated only on the original version of the benchmark. However, the correlation becomes stronger when adding results from multiple permutations, achieving the coefficient of determination $R^2 > 0.9$ for *partial fluctuation rates* (computed on subsets of permutations) and the novel metric, *worst accuracy*.

2 Context & Related Work

MCQs have been widespread in the education field (Brady, 2005; Moss, 2001). They are characterized by presenting several answer *options* within a question body, typically accompanied by *labels* (e.g. `A/B/C/D`), where a *correct answer* can be one, several, or no labels. In the context of LLMs evaluation, however, MCQ benchmarks come with a single correct label, see an example in Figure 1. The unique correct answer allows for comparing models' responses to it and obtaining accuracy.

As for the extraction of a model's responses, one can compare probabilities of the next token given a question prompt and choose the most probable one as the model's selected label. Another method prominent in the field, though not covered in this paper, is to allow models to generate an answer of arbitrary length and later classify it as one of the labels (Wang et al., 2024c).

Previous research demonstrates that one can cause answer fluctuation by permuting questions, their options and/or labels.

Answer fluctuation Mizrahi et al., 2023 show that even minimal prompt paraphrases, e.g., replacing "have" with "include" in the question, impact models' performance. Liang et al., 2023 indicate

Which of these will form new soil the fastest?	
Labels	**Options**
A	A log rotting in a forest.
B	Water running in a stream.
C	A rock sitting in a garden.
D	Waves breaking on a beach.
Correct label: **A**	

Figure 1: An MCQ example from ARC-C (Clark et al., 2018).

that a different choice of few-shot examples can lead to vast differences in obtained F1 scores. Mina et al., 2025, as well, highlight the effect of few-shot examples, where recency bias (preference towards selecting the last option) is found in the few-shot scenario but not the zero-shot scenario.

Pezeshkpour and Hruschka, 2024 study the effect of option order permutation. Their work shows that the difference between the best and worst possible performance of a model achievable via option reordering can be as high as 70 percentage points for InstructGPT and 50 percentage points for GPT-4, highlighting the fact that the introduction of few-shot examples does not lead to higher robustness.

Zheng et al., 2024 demonstrate that moving all correct answers to one of `A/B/C/D` can cause a performance increase in some models and a decrease in others, serving as an example of *selection bias* (Li and Gao, 2024; Pezeshkpour and Hruschka, 2024; Wang et al., 2024a). Additionally, using different option typography (e.g., `(A)` instead of `A.` or replacing common option labels `A/B/C/D` with rarer ones, e.g. `$/&/#/@`) leads to lower results (Zheng et al., 2024; Alzahrani et al., 2024). Furthermore, a similar drop in performance is achieved (Wei et al., 2024) if one keeps the order of options but reverses the order of labels (e.g., `D/C/B/A`).

Tjuatja et al., 2024 compare LLMs' biases on MCQ with those of people and find no apparent replication of human behavior, while indicating that all tested models show sensitivity to factors not significant for human respondents, such as typos.

Finally, changing the question from MCQ to another format, such as Cloze (Madaan et al., 2024), open-ended generation (Röttger et al., 2024), or True/False questions (Wang et al., 2025) can drastically change models' responses.

LLM evaluation in the fluctuation context Given the answer instability, Wei et al., 2024 propose the *fluctuation rates* metric that compares answers on the original and inverse option orders. It considers that a model's response fluctuates if these answers are different. However, this calculation is not adapted for working with multiple permutations.

To ensure more stable model performance, Zheng et al., 2024 introduce *PriDe* (Li et al., 2024; Wei et al., 2024; Reif and Schwartz, 2024 present other calibration techniques): an approach to adjust models' probabilities of answer tokens (e.g. `A/B/C/D`) by computing their priors, independent from questions, and then using them to debias models' responses. This methodology has only been evaluated in terms of improving the original performance of models, not considering the evaluation of answer robustness.

Sensitivity gap (Pezeshkpour and Hruschka, 2024) is one of the proposed metrics that incorporates the information about both model performance and answer fluctuation. It is computed as the difference between the maximum and minimum accuracies that can be obtained by changing the order of options. However, the paper does not provide the exact formula for this calculation. Similarly, Gupta et al., 2024 introduce an unnamed metric to assess, which we take the liberty to name *strong accuracy*. It compares pair-wise responses from the original option order and a permutation and calculates an average rate of keeping correct answers through permutation pairs. Their approach involves picking random permutations, although the stability of the metric is not addressed.

To the best of our knowledge, the above-mentioned metrics have not been substantially compared to one another, as well as to robustness. The connection of reliability and other metrics has remained underexplored, being demonstrated only for accuracy (Pezeshkpour and Hruschka, 2024; Liang et al., 2023; Wei et al., 2024).

3 Metrics Survey

Given the variety of metrics available for MCQ evaluation, it is essential to provide a coherent formalization for each of them. This section presents our notation and permutation types used for computation. Furthermore, we provide formulas for existing metrics. Finally, we introduce a novel metric, that we call *worst accuracy*.

3.1 Notation

We assume that all benchmarks come with their own set of labels L (such as `A/B/C/D`), as well as a set of questions. We define each metric for a question q and, within our experiments, we average all calculations among questions. However, one can potentially adopt different aggregation strategies.

Each question has an associated set of textual options $O = \{o_1 \dots o_{|L|}\}$, e.g. $\{cat, dog \dots\}$, as well as a correct answer a (e.g. dog). We define a permutation set $\mathcal{R}(O)$ as a set of reordering of set O, e.g. $\mathcal{R}(O) = \{\{o_1, o_2, o_3, o_4\}, \{o_4, o_3, o_2, o_1\}\}$. Given few-shot examples, question q and permuted options $r_j \in \mathcal{R}(O)$, we obtain model answer m_j.

Please note that the labels are not permuted. Therefore, a label of the correct answer might differ among permutations. To keep track of it, we introduce the notation l_{a_j} which stands for the label of the correct answer a on a permutation $r_j \in \mathcal{R}(O)$. Few-shot examples and the question itself remain constant throughout the permutations, and for this reason, they are not presented in subsequent formalization.

3.2 Permutation types

When all possible orders of options are present, we call such a permutation set $\mathcal{R}_{full}$. Since $|\mathcal{R}_{full}| = |L|!$, its calculation is extremely costly. To make computations more efficient, we employ subsets of permutations.

If the permutation set contains only the original options order, we call refer to it as $\mathcal{R}_{original}$. Previous research (Wei et al., 2024), among their other propositions, suggests using a permutation that can be described as $original\ and\ inverse$ order: $\mathcal{R}_{oi} = \{\{o_1 \dots o_{|L|}\}, \{o_{|L|}, o_{|L|-1} \dots o_1\}\}$. Following Zheng et al., 2024, we also utilize $cyclic$ permutations in which all options are moved in a circular manner between permutations. $\mathcal{R}_{cyclic} = \{\{o_1 \dots o_{|L|}\}, \{o_2 \dots o_{|L|}, o_1\}, \dots, \{o_{|L|}, o_1 \dots o_{|L|-1}\}\}$, where $|\mathcal{R}_{cyclic}| = |L|$.

Finally, we assess the importance of picking these particular option orders by creating random subsets[1] $\mathcal{R}_{random2}$ (size $= 2$) and $\mathcal{R}_{randomL}$ (size $= |L|$).

[1] Out of the set of possible permutations select random, using `random.sample` with `seed = 0`.

3.3 Existing metrics

The central notion of this work is fluctuation, for the measurement of which we adjust the fluctuation rates metric introduced by Wei et al., 2024:

$$FR = 1 - \prod_{j=1}^{|\mathcal{R}|} \mathbb{1}[m_1 = m_j] \quad (1)$$

By this definition, we consider a model's answer to fluctuate if at least one response changes throughout permutations. This rigid interpretation allows us to have higher confidence in models' responses.

In the permutation context, one can adapt accuracy by averaging the accuracies obtained in the tested permutations. This change transforms the discrete accuracy into a continuous metric $average\ accuracy$ (which is equivalent to accuracy when computed on $\mathcal{R}_{original}$):

$$AAcc = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \mathbb{1}[m_j = a] \quad (2)$$

Furthermore, we compare the average accuracy results to *strong accuracy*, as introduced by Gupta et al., 2024, strengthening the accuracy with pairwise comparison of answers across permutations. We update the formula to fit our notation:

$$SAcc = \frac{\mathbb{1}[m_1 = a]}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \mathbb{1}[m_1 = m_j] \quad (3)$$

Moreover, we utilize $PriDe$ (Zheng et al., 2024) in its original implementation by the authors. The method involves computing accuracy using debiased probabilities instead of the original ones. See details about the implementation in the original paper.

To adapt the probability mass of the correct answer to the permutation context, we simply average probabilities across permutations:

$$Prob = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} p(l_{a_j}|r_j). \quad (4)$$

We adjust *Brier score* equivalently[2]:

$$BS = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \sum_{l \in L} (\mathbb{1}[l = l_{a_j}] - p(l|, r_j))^2 \quad (5)$$

[2]In this work, we convert the metric to *1 - Brier*, to map all the metrics onto the same interval $[0, 1]$ where 0 is the worst performance and 1 is the best.

Lastly, we modify the *normalized ENtropy* formula from Tjuatja et al., 2024 to incorporate the permutations[3]:

$$EN = \frac{-1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} \sum_{l \in L} \frac{p(l|r_j)) \cdot log_2(p(l|r_j))}{log_2(|L|)} \quad (6)$$

3.4 Metric proposition

Since metrics are averaged across all questions, both average and strong accuracies become hard to interpret. A result of 0.5 can signify both that a model is robust and produces correct answers in all permutations for 50% of the questions, or that the model is not robust and for all questions there is only a 50% chance to get a correct response. We argue that this distinction is important in the context of model reliability, and hence we propose a novel metric, *worst accuracy*, which equals 1 iff a model answers correctly throughout all tested permutations:

$$WAcc = \mathbb{1}[m_1 = a] \prod_{j=1}^{|\mathcal{R}|} \mathbb{1}[m_1 = m_j] \quad (7)$$

One can notice stark similarities between the proposition and Eq. 3. In fact, the metrics are equal if $|\mathcal{R}| = 2$. However, extending the pairwise comparison to include all answers guarantees model robustness on a given question.

In the original paper (Pezeshkpour and Hruschka, 2024), *sensitivity gap* only receives a textual definition: "difference between the maximum and minimum LLMs' performance". We provide an interpretation of the metric[4], using the above-mentioned worst accuracy and an auxiliary metric *best accuracy (BAcc)*, described below.:

$$SensG = BAcc - WAcc \quad (8)$$

BAcc considers a question answered if there is at least one permutation in which the model arrives at the correct answer:

$$BAcc = 1 - \prod_{j=1}^{|\mathcal{R}|} \mathbb{1}[m_j \neq a] \quad (9)$$

[3]Similarly to $Brier$, we use $1 - Entropy$.

[4]Similarly to $Brier$ and $Entropy$, we use *1 - SensG*.

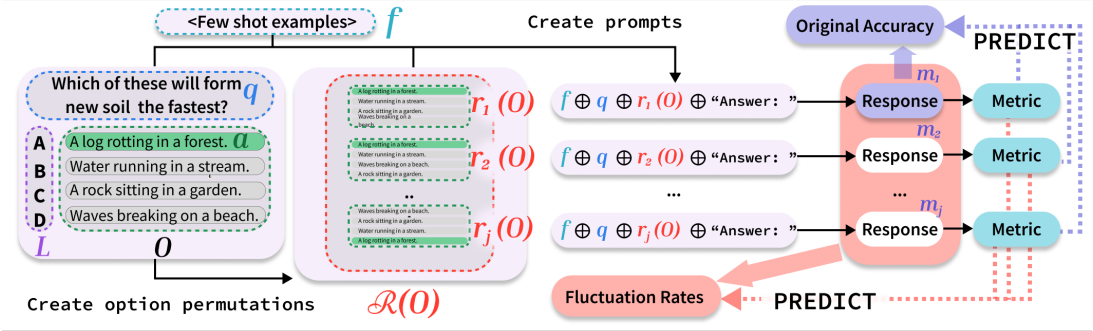


Figure 2: Schematization of the proposed evaluation protocol.

4 Assessment Protocol

Having presented all the metrics, one can choose a multitude of assessment protocols. Since computing all permutations and finding the full fluctuation rates is a costly venture, we argue that an appropriate metric for MCQ evaluation would be highly representative of the full fluctuation rates computed in a lower-cost environment. Therefore, we propose evaluating the correlation of the proposed methodologies with full fluctuation rates. However, a metric should still be illustrative of the model's accuracy on the original option order, since this represents the result of a model on a version it was exposed to. Thus, we additionally propose the following protocol, illustrated in Figure 2:

1. Calculate the accuracy models achieve on the original benchmarks (using the original option order).
2. Calculate fluctuation rates on all possible permutations of option order for each model and benchmark.
3. Calculate the metrics from Section 3 on a smaller subset of permutations for each model on each benchmark.
4. Find the correlation between metrics and full fluctuation rates using R^2.
5. Find the correlation between metrics and original accuracy using R^2.
6. Find the correlation between a metric and both full fluctuation rates and original accuracy using R^2.

4.1 Models

We perform our experiments on 10 LLMs with parameter sizes below 10B. Models of this size are frequently used for fine-tuning[5], thus making their evaluation more impactful. This size also allows us to perform a costly operation of computing all possible permutations. In our experiments we use pre-trained and instruct-tuned versions of Llama-3.1-8B (Dubey et al., 2024), Gemma-2-9B (Gemma Team et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2.5-7B (Qwen et al., 2025), as well as R1-Distill-Llama-8B and R1-Distill-Qwen-7B from DeepSeek (DeepSeek-AI et al., 2025). All models are initialized using HuggingFace's `transformers` library with `bfloat16` precision.

4.2 Benchmarks

Due to potential variability in results coming from slight variations of input text, we choose to use publicly shared Meta's evaluation datasets[6] that contain full final prompts, including instructions, few-shot examples, their order, and option typography for ARC-C (Clark et al., 2018), CSQA (Talmor et al., 2019), MMLU[7] (Hendrycks et al., 2021), AGIEval[8] (Zhong et al., 2024), and Winogrande (Sakaguchi et al., 2021)[9]. All benchmarks' prompts can be generalized to the following format: "<instruction> <few-shot examples> <test question q > <test options r_j> Answer: ".

5 Results

This section presents the results of Steps 4-6 of the protocol introduced above. To begin with, we

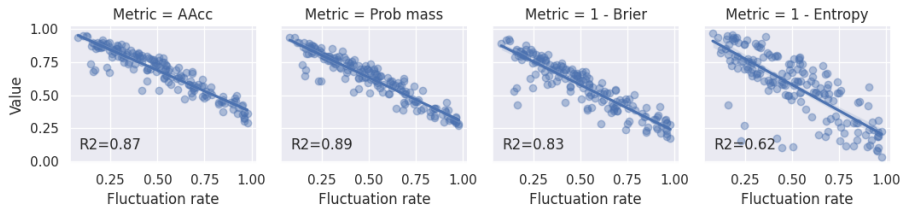
[5]At the time of writing 100-900+ fine-tuned versions are available on HuggingFace for each selected model.

[6]https://huggingface.co/datasets/meta-llama/Llama-3.1-8B-evals

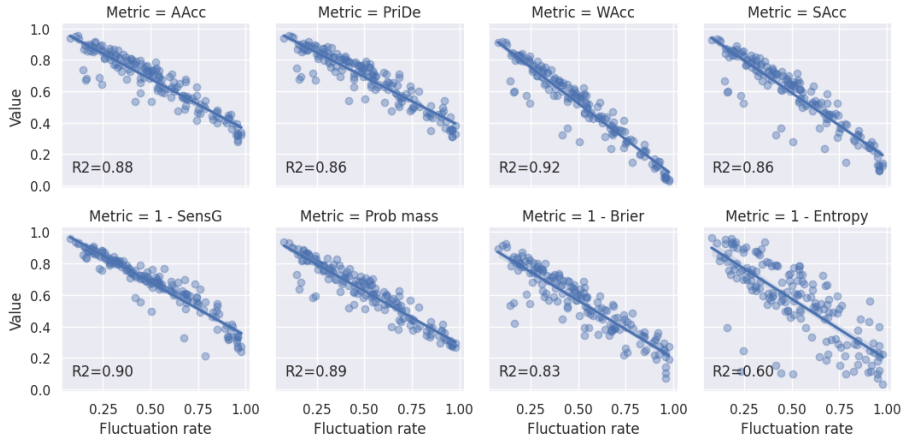
[7]The benchmark contains 57 diverse subtasks, in this work we present results from a sample of 12 subtasks.

[8]Though originally a 5-option benchmark, AGIEval contains questions with nan as the final option. We remove it and consider such questions to be 4-option, thus creating two subsets AGIEval-4 and AGIEval-5.

[9]See Appendix B for more information.



(a) Metrics computed using $\mathcal{R}_{original}$



(b) Metrics computed using $\mathcal{R}_{cyclic}$

Figure 3: Metrics and full fluctuation rates correlation. Each data point represents results obtained by a model on a benchmark using the given metric.

compute the correlation of the metrics with full fluctuation rates using original order and permutation subsets. Second, we compare the results when adding correlation with the original accuracy. Lastly, we assess the impact of picking random permutations for metric calculation[10].

5.1 Correlation with full fluctuation rates

Figure 3a shows that all metrics that could be calculated using only the original option order are representative of full fluctuation rates to a great extent, with the probability mass being the best proxy out of the tested metrics. While entropy appears to have the weakest correlation, the R^2 measure still indicates a certain level of association.

Figure 3b presents the metrics results calculated using each benchmark's cyclic permutations. Interestingly, there is no change in R^2 for probability mass and Brier score when adding extra permutations, thus indicating that additional permutations do not contain more information about fluctuation for these metrics. Worst accuracy appears to have the highest correlation with full fluctuation rates on $\mathcal{R}_{cyclic}$. As seen in the plots of the sensitivity gap and strong and worst accuracies, specific data points appear pretty far from the general fit. These points represent the results of models on Winogrande[11], a benchmark with only two options. One potential explanation for this behavior is that the performance of these metrics is dependent on the size of $|L|$ and, therefore, the number of available permutations.

Seeing these results, we investigate if partial fluctuation rates (computed over subsets of permutations) are associated with full fluctuation rates. In fact, such an approach shows the best performance in $\mathcal{R}_{cyclic}$ and $\mathcal{R}_{randomL}$ setups, exceeding the results of the worst accuracy (see Table 1a). However, such a method appears to be much less stable over just two permutations, with correlation dropping significantly over $\mathcal{R}_{random2}$. Similarly, sensitivity gap performs very poorly on $\mathcal{R}_{random2}$. This can serve as an additional indicator that two permutations are insufficient for calculating these metrics.

5.2 Correlation with original accuracy and full fluctuation rates

As the next step, we find the correlation between the metrics and the accuracy computed on the original benchmark (see the results in Table 1b). Though partial fluctuation rates have a substantial

[10] All metrics are computed on the same randomly picked permutations $\mathcal{R}_{random2}$ and $\mathcal{R}_{randomL}$.

[11] Find more detailed representation in Appendix A.3.

	AAcc	PriDe	WAcc	SAcc	1 - SensG	Prob mass	1 - Brier	1 - Entropy	1 - FR (partial)
$\mathcal{R}_{oi}$	0.873	0.863	0.870	0.870	0.640	**0.893**	0.833	0.605	0.829
$\mathcal{R}_{random2}$	0.881	0.877	0.831	0.831	0.235	**0.894**	0.836	0.594	0.479
$\mathcal{R}_{cyclic}$	0.877	0.863	0.923	0.863	0.896	0.894	0.832	0.602	**0.953**
$\mathcal{R}_{randomL}$	0.880	0.868	0.914	0.864	0.866	0.894	0.835	0.600	**0.941**

(a) Target feature = full fluctuation rates.

	AAcc	PriDe	WAcc	SAcc	1 - SensG	Prob mass	1 - Brier	1 - Entropy	1 - FR (partial)
$\mathcal{R}_{oi}$	0.990	**0.993**	0.960	0.960	0.647	0.960	0.943	0.686	0.844
$\mathcal{R}_{random2}$	**0.979**	0.978	0.930	0.930	0.275	0.957	0.937	0.674	0.508
$\mathcal{R}_{cyclic}$	0.987	**0.994**	0.961	0.963	0.827	0.960	0.941	0.682	0.897
$\mathcal{R}_{randomL}$	**0.988**	0.985	0.964	0.958	0.813	0.959	0.941	0.681	0.903

(b) Target feature = accuracy on original order.

	AAcc	PriDe	WAcc	SAcc	1 - SensG	Prob mass	1 - Brier	1 - Entropy	1 - FR (partial)
$\mathcal{R}_{oi}$	**0.932**	0.928	0.915	0.915	0.643	0.927	0.888	0.645	0.836
$\mathcal{R}_{random2}$	**0.930**	0.928	0.881	0.881	0.255	0.926	0.886	0.634	0.494
$\mathcal{R}_{cyclic}$	0.932	0.928	**0.942**	0.913	0.861	0.927	0.887	0.642	0.925
$\mathcal{R}_{randomL}$	0.934	0.926	**0.939**	0.911	0.839	0.927	0.888	0.641	0.922

(c) Target features = full fluctuation rates and original accuracy.

Table 1: R^2 scores for metrics computed on permutation subsets and full fluctuation scores and/or original accuracy. For random subsets, we used the same permutations for all calculations. Best results for each permutation subset are bolded.

correlation with full fluctuation rates, it appears that this strong link comes with less information about original accuracy than other metrics. Similar to the previous results, sensitivity gap and fluctuation rates computed over $\mathcal{R}_{random2}$ demonstrate a drastic drop in comparison to $\mathcal{R}_{oi}$, further suggesting the impact of chosen dimensions on the calculation of the metric.

Curiously, the highest correlation with the original accuracy on $\mathcal{R}_{oi}$ and $\mathcal{R}_{cyclic}$ is achieved by PriDe and not by averaged accuracy. Probability mass, Brier score, worst and strong accuracies are strongly associated with original accuracies, though slightly worse than PriDe and averaged accuracy.

As our final evaluation, we compute the R^2 score for correlation with both targets simultaneously (Table 1c). Worst accuracy arises to be the best approach given $\mathcal{R}_{cyclic}$ or $\mathcal{R}_{randomL}$. In contrast, averaged accuracy appears to be the best on $\mathcal{R}_{oi}$ and $\mathcal{R}_{random2}$, demonstrating the most balanced performance across two target features.

5.3 Permutation choice impact

Considering the differences in performance when adopting $\mathcal{R}_{oi}$ and $\mathcal{R}_{random2}$, we compare the standard deviations of the tested metrics. For this purpose, we choose 100 random pairs of permutations for each benchmark except Winogrande[12], as well as 100 random tuples of size $|L|$, and calculate metrics for each of them. We report an averaged standard deviation of a metric on a benchmark in Figure 4. We find that the standard deviation of the sensitivity gap and partial fluctuation rates computed over random pairs of permutations are the most significant among the metrics, mirroring the observed drops of R^2 when replacing $\mathcal{R}_{oi}$ with $\mathcal{R}_{random2}$. Furthermore, we remark that standard deviations are higher on benchmarks where all models perform worse on the original order[13] (e.g. Global Facts, Machine learning, and High School Math).

Additionally, we notice that within permutations, continuous metrics can increase on some questions, however, to a similar extent decrease on others, and the overall averaged performance stays stable no matter the permutations chosen (reflected by low standard deviation in Figure 4). While this stability allows one to pick random permutations for calculation of the metrics, it appears to be also associated with a capped correlation with fluctuation: R^2 values do not improve when adding more permutations (compare Figures 3a and 3b). Thus, *computing continuous metrics over several permutations might have no benefit over computing them over* $\mathcal{R}_{original}$.

[12] Since only 2 permutations are available for it.

[13] See the details about models' original accuracies in Appendix A.1.

Figure 4: Standard deviation of each metric on a given benchmark, averaged by model. *Left*: standard deviation given random pairs of permutations. *Right:* standard deviation computed on random tuples of permutations of length $|L|$.

While using $|L|$ permutations is associated with lower standard deviation, it remains quite significant for PriDe, worst and strong accuracies, sensitivity gap and fluctuation rates. Consequently, selecting random permutations (as proposed in Gupta et al., 2024) might lead to unstable evaluation.

6 Limitations & Future Work

Selection of permutations As demonstrated in the results, multiple metrics appear sensitive to the permutations chosen to compute them. While we observe this phenomenon, further study is required on the optimal approaches to permutation selection.

Other permutation types While we illustrated how strongly metrics correlate with fluctuation, we only considered option order permutations. As discussed in Section 2, fluctuation can occur with question paraphrasing, changing option typography, replacing option labels, etc. Further work needs to include these types of permutations in the assessment.

Model sizes All experiments were performed using similar-sized models. Including models of other sizes is essential to understanding whether the demonstrated correlation of tested metrics is characteristic only of the models of this size or whether a more general pattern exists.

Text generation vs next token prediction In our experiments, models' answers were decided by the next token with the highest probabilities, but as previous research has demonstrated (Wang et al., 2024b,c), it might be associated with higher fluctuation rates of responses than text generation. Further research needs to incorporate and analyze both approaches.

7 Conclusion

In this paper, we presented a new protocol for metric comparison in the context of answer fluctuation that LLMs exhibit when options of MCQ tasks are permuted. To achieve this, we reviewed, formalized, and computed existing metrics applicable to such benchmarks, and introduced a new metric, worst accuracy. When applying the evaluation framework, we discovered that:

1. Most existing metrics appear to correlate strongly with fluctuation rates.
2. When only having access to the results of a model on the original order of options, one might employ probability mass for a substantial correlation with full fluctuation rates. However, computing the same metric over multiple permutations does not appear to yield better results.
3. If information about the original model performance is not of high importance, computing fluctuation rates on cyclic permutations comes to be the best indicator of fluctuation on all possible permutations.
4. However, if it is essential for the evaluation to represent the original accuracy, the worst accuracy shows the best performance.

Further research is required to extend these findings to different approaches to answer generation by models, a variety of sizes, and other types of permutations that lead to answer fluctuation.

References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. In *ArXiv*.

Norah A. Alzahrani, Hisham A. Alyahya, Sultan Yazeed Alnumay, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal A. Mirza, Nouf M. Alotaibi, Nora Altwairesh, Areeb Alowisheq, Saiful Bari, Haidar Khan, A. Jeddah, B. Makkah, C. Paris, Djafri Riyadh, Bekkai Riyadh, D. Makkah, Peter Clark, Isaac Cowhey, and 189 others. 2024. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. In *Annual Meeting of the Association for Computational Linguistics, ACL*.

Alireza Amiri-Margavi, Iman Jebellat, Ehsan Jebellat, and Seyed Pouyan Mousavi Davoudi. 2024. Enhancing Answer Reliability Through Inter-Model Consensus of Large Language Models. In *ArXiv*.

Anne-Marie Brady. 2005. Assessment of learning with multiple-choice questions. In *Nurse Education in Practice*, volume 5, pages 238–242. Elsevier.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. In *ArXiv*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. In *ArXiv*.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding Emergent Abilities of Language Models from the Loss Perspective. In *Conference on Neural Information Processing Systems, NeurIPS*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The Llama 3 Herd of Models. In *ArXiv*.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-The-Human Evaluation Framework: Rethinking Human Evaluation for Generative Large Language Models. In *Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1137–1160.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *International Conference on Learning Representations, ICLR*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. In *arXiv*.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing Answer Order Can Decrease MMLU Accuracy. In *ArXiv*.

Tianxing He, Jingyu (Jack) Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James R. Glass, and Yulia Tsvetkov. 2022. On the Blind Spots of Model-Based Evaluation Metrics for Text Generation. In *Annual Meeting of the Association for Computational Linguistics, ACL*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations, ICLR*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. In *ArXiv*.

Aisha Khatun and Daniel Brown. 2023. Reliability check: An analysis of GPT-3's response to sensitive topics and prompt wording. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 73–95. Association for Computational Linguistics.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2024. CalibraEval: Calibrating Prediction Distribution to Mitigate Selection Bias in LLMs-as-Judges. In *ArXiv*.

Ruizhe Li and Yanjun Gao. 2024. Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions. In *ArXiv*.

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 R'e, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic Evaluation of Language Models. In *Annals of the New York Academy of Sciences*, volume 1525, pages 140 – 146.

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

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Oluwasanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. Quantifying Variance in Evaluation Benchmarks. In *ArXiv*.

Mario Mina, Valle Ruíz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor González-Agirre. 2025. Cognitive Biases, Task Complexity, and Result Intepretability in Large Language Models. In *International Conference on Computational Linguistics, ICCL*, pages 1767–1784.

Moran Mizrahi, Guy Kaplan, Daniel Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of What Art? A Call for Multi-Prompt LLM Evaluation. In *Transactions of the Association for Computational Linguistics*, volume 12, pages 933–949.

Edward Moss. 2001. Multiple choice questions: their value as an assessment tool. In *Current Opinion in Anesthesiology*, volume 14, pages 661–666. LWW.

Ninad Naik. 2024. Probabilistic Consensus through Ensemble Validation: A Framework for LLM Reliability. In *ArXiv*.

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, and 261 others. 2023. GPT-4 Technical Report. In *arXiv*.

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

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

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. In *arXiv*.

Yuval Reif and Roy Schwartz. 2024. Beyond Performance: Quantifying and Mitigating Label Bias in LLMs. In *Proc. of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 6784–6798.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *Conference on Language Modeling, COLM*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. In *ACM Computing Surveys*, volume 55, pages 1 – 45.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schutze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Annual Meeting of the Association for Computational Linguistics, ACL*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. In *Communications of the ACM*, volume 64, pages 99–106. ACM New York, NY, USA.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? In *Conference on Neural Information Processing Systems, NeurIPS*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. In *Conference on Empirical Methods in Natural Language Processing*.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Annual Meeting of the Association for Computational Linguistics, ACL*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 4149–4158.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. In *Transactions of the Association for Computational Linguistics, TACL*, volume 12, pages 1011–1026. MIT Press.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. LLMs May Perform MCQA by Selecting the Least Incorrect Option. In *International Conference on Computational Linguistics, ICCL*, pages 5852–5862.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu,

Tianyu Liu, and Zhifang Sui. 2024a. Large Language Models are not Fair Evaluators. In *Annual Meeting of the Association for Computational Linguistics, ACL*, pages 9440–9450.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024b. Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. In *Conference on Language Modeling, COLM*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL*, pages 7407–7416.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024d. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Conference on Neural Information Processing Systems, NeurIPS*.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL*, pages 5598–5621.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations, ICLR*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. In *International Conference on Learning Representations, ICLR*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL*, pages 2299–2314.

A Metric Results

In this section we present detailed results, indicating individual model performance on tested benchmarks. Section A.1 demonstrates original accuracies for benchmark pairs. Section A.2 includes full fluctuation rates for model-benchmark pairs. Section A.3 presents correlation plots of a metric and full fluctuation rates, detailed by model and benchmark.

A.1 Original Accuracy

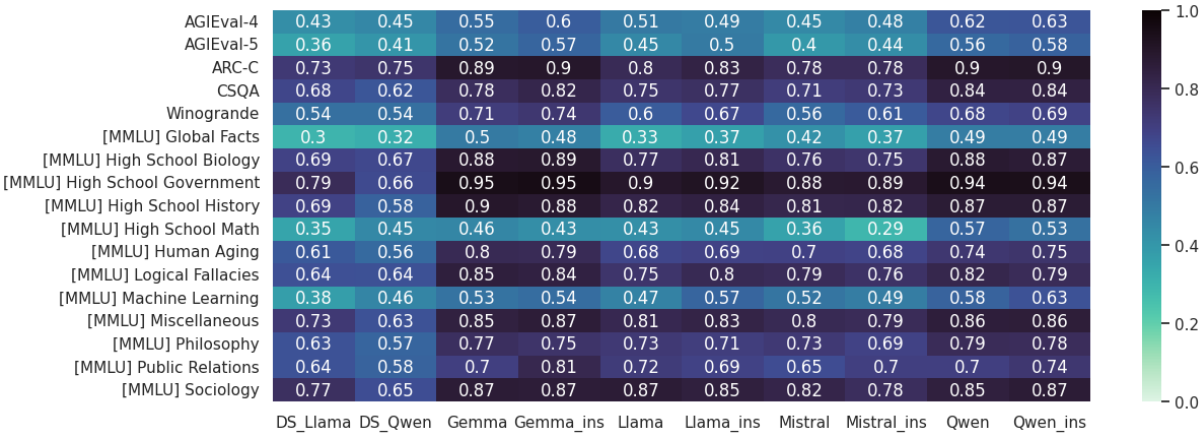


Figure 5: Accuracies obtained by the models on the benchmarks using the original option order.

A.2 Full Fluctuation Rates

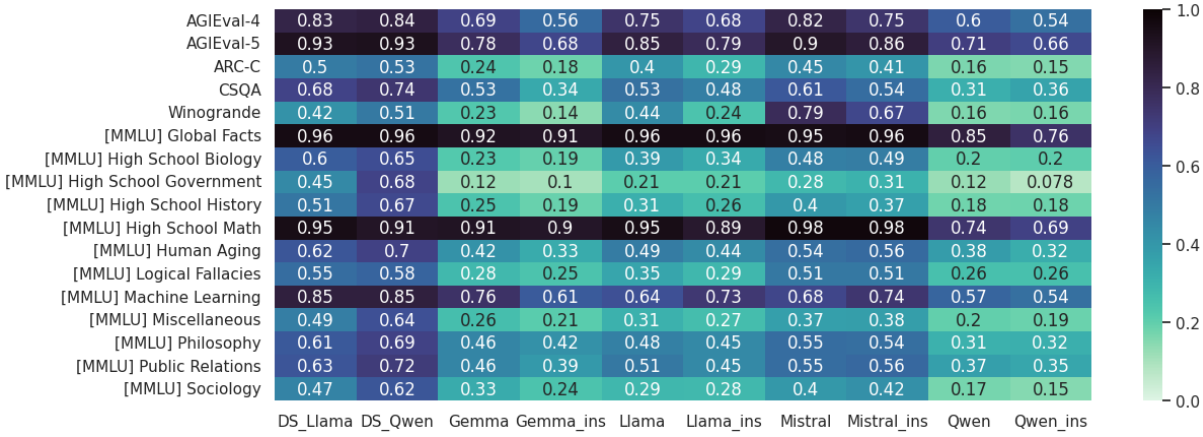


Figure 6: Fluctuation rates of the models on the benchmarks calculated using all permutations.

A.3 Metrics on Different Permutations

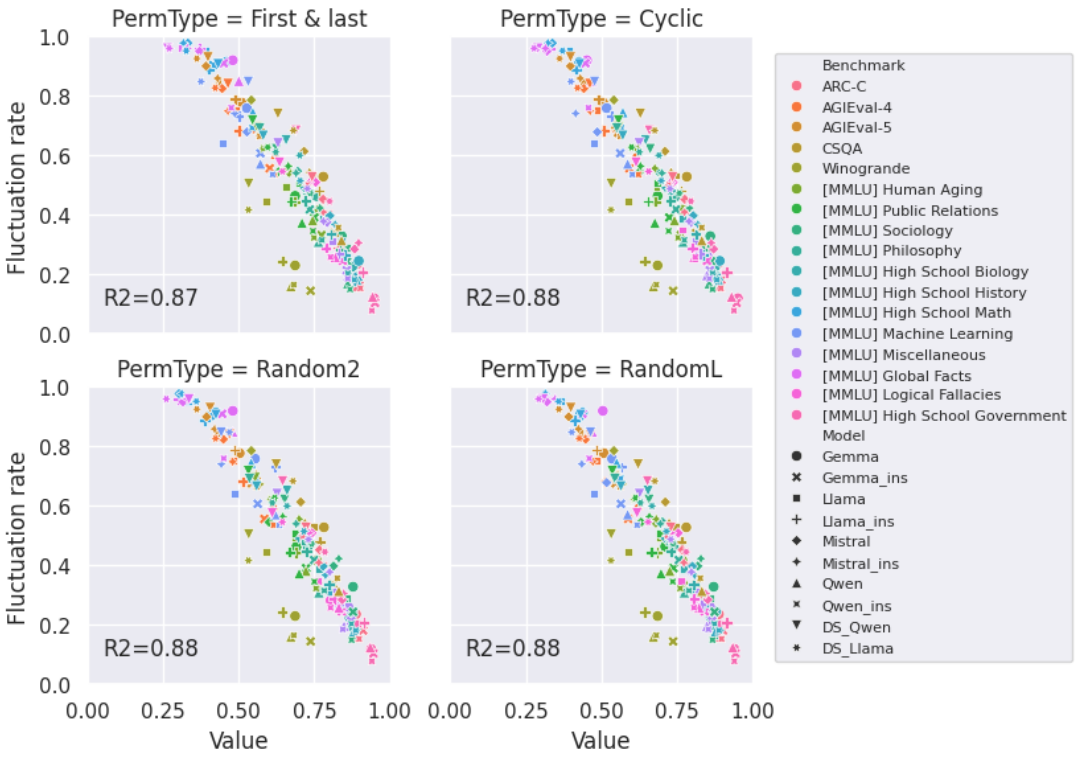


Figure 7: Average accuracy on permutation subsets and full fluctuation rates for all tested models and benchmarks.

Figure 8: PriDe on permutation subsets and full fluctuation rates for all tested models and benchmarks.

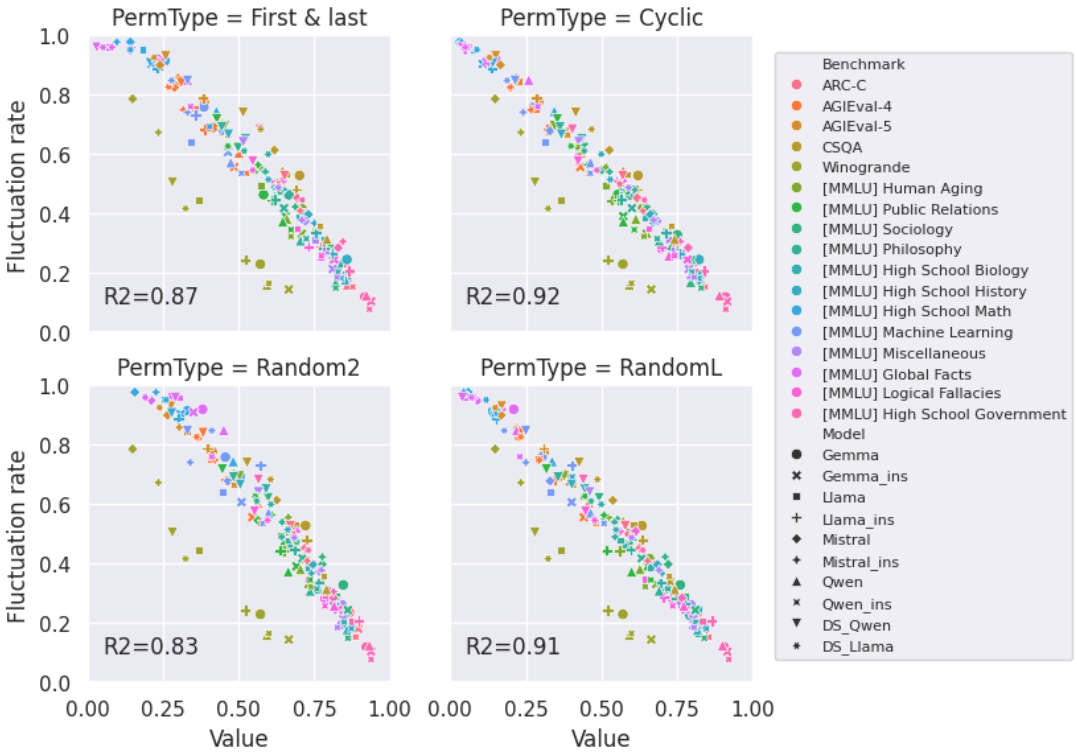


Figure 9: Worst accuracy on permutation subsets and full fluctuation rates for all tested models and benchmarks.

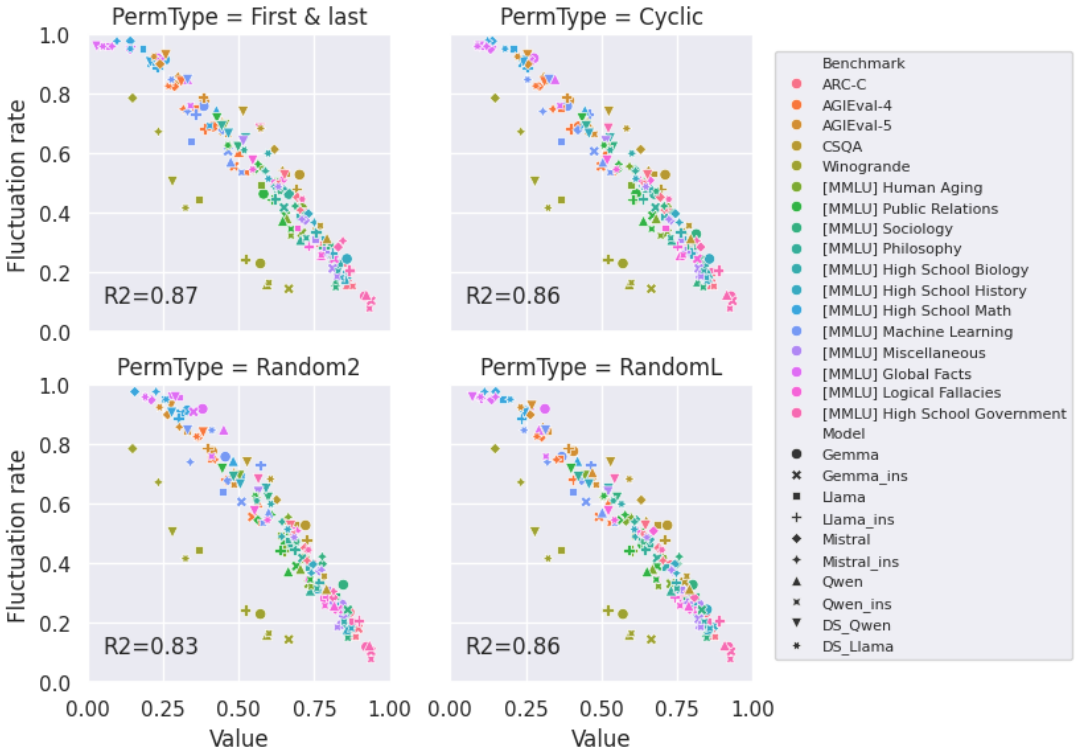


Figure 10: Strong accuracy on permutation subsets and full fluctuation rates for all tested models and benchmarks.

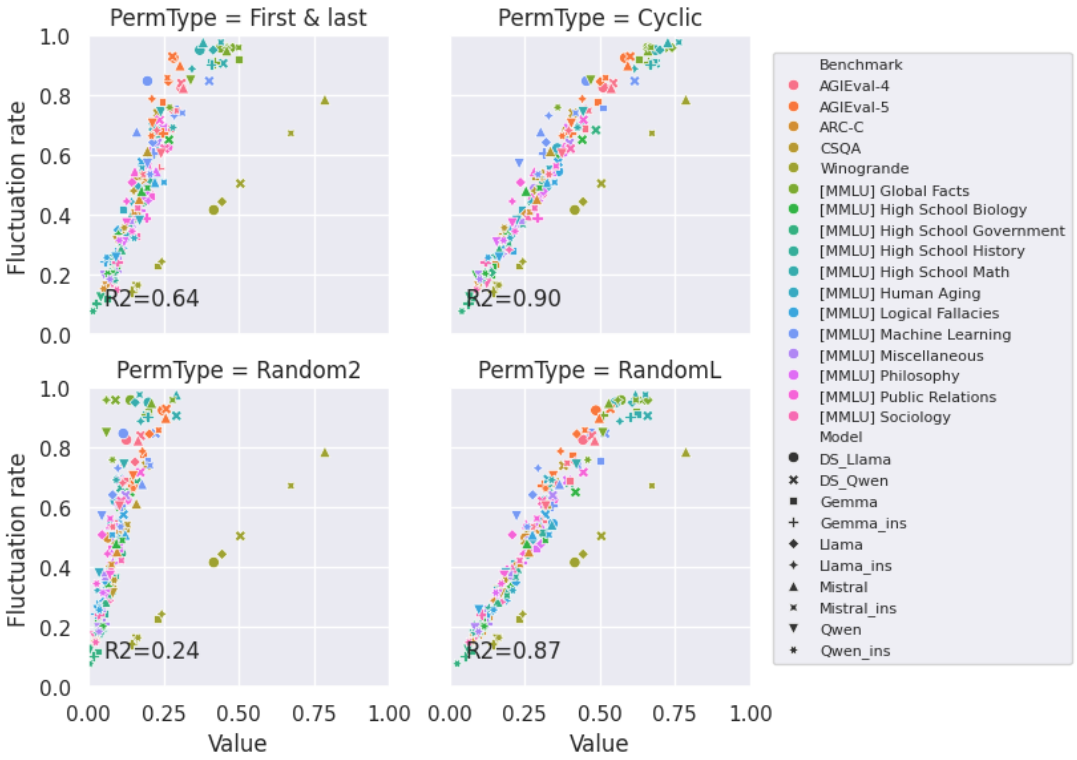


Figure 11: Sensitivity gap on permutation subsets and full fluctuation rates for all tested models and benchmarks.

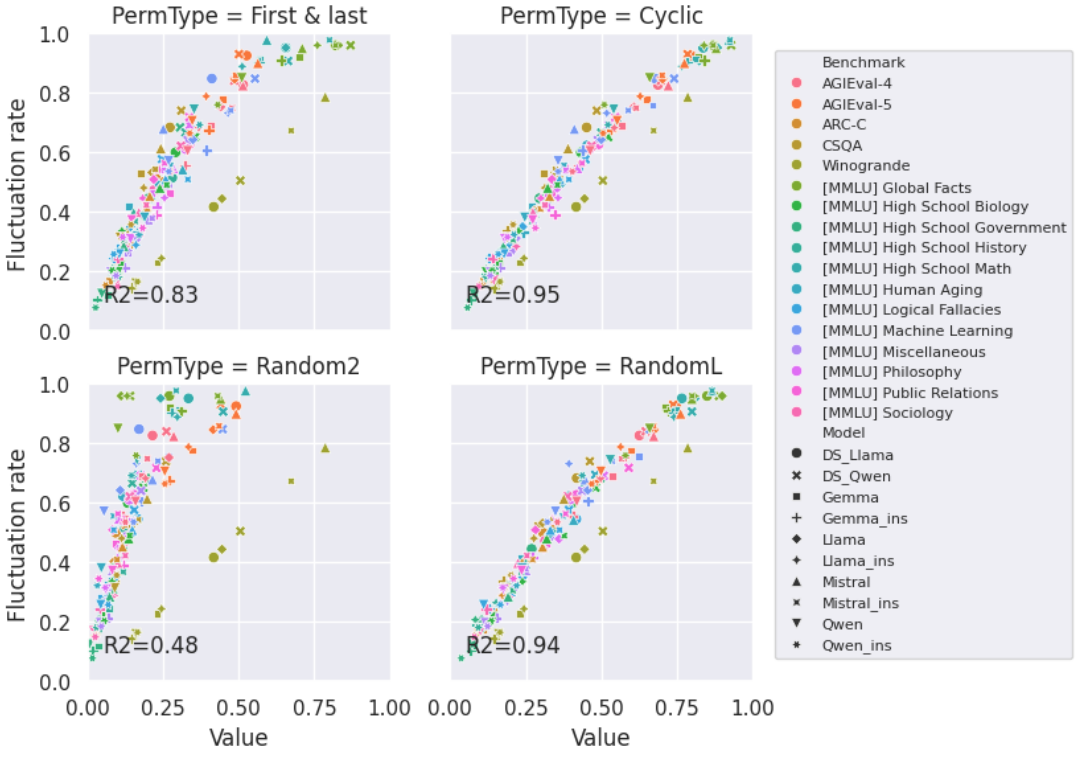


Figure 12: Fluctuation rates on permutation subsets and full fluctuation rates for all tested models and benchmarks.

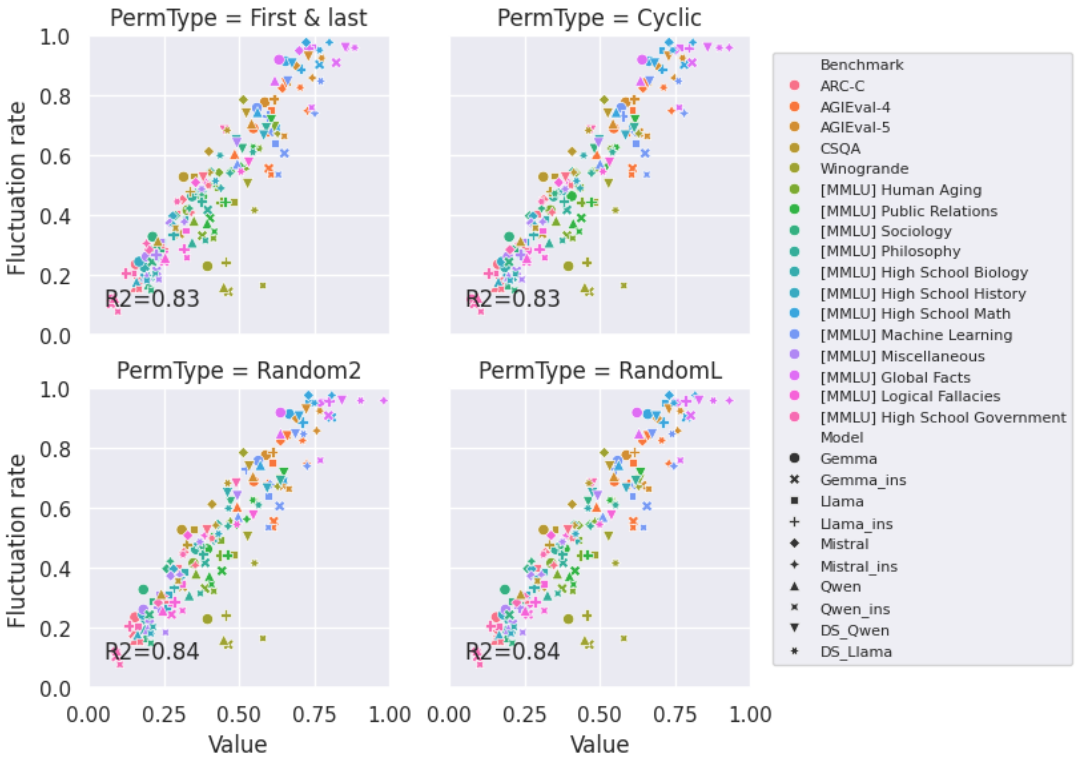


Figure 13: Brier score on permutation subsets and full fluctuation rates for all tested models and benchmarks.

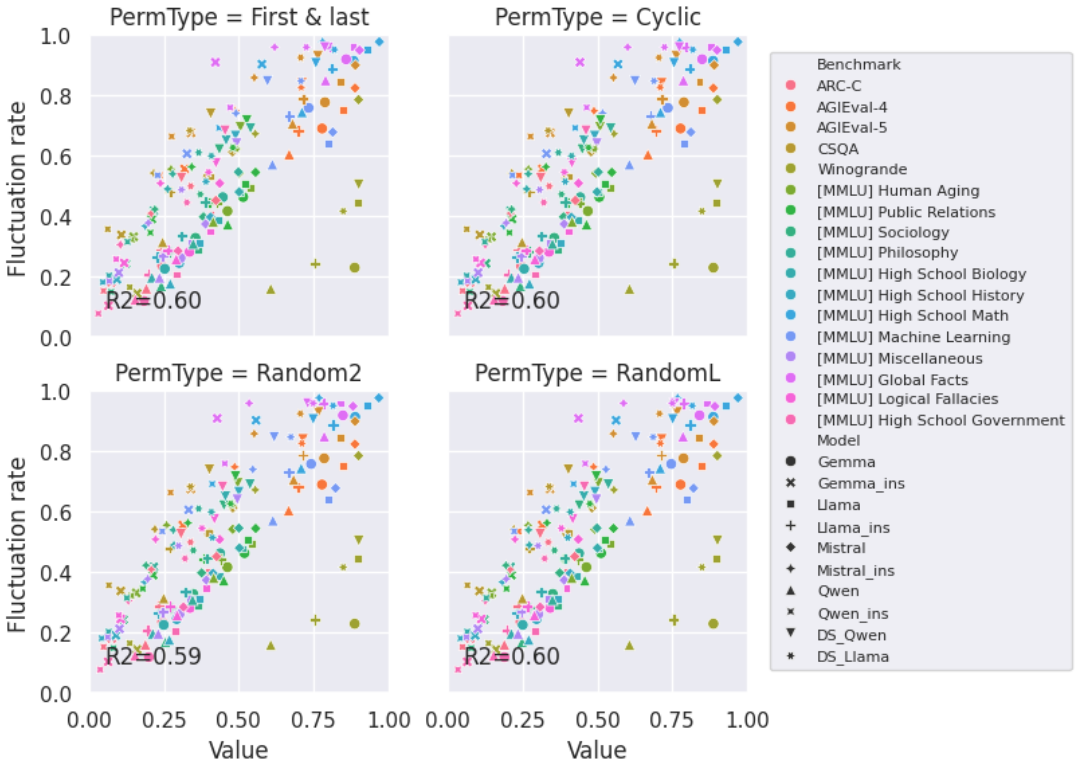


Figure 14: Entropy on permutation subsets and full fluctuation rates for all tested models and benchmarks.

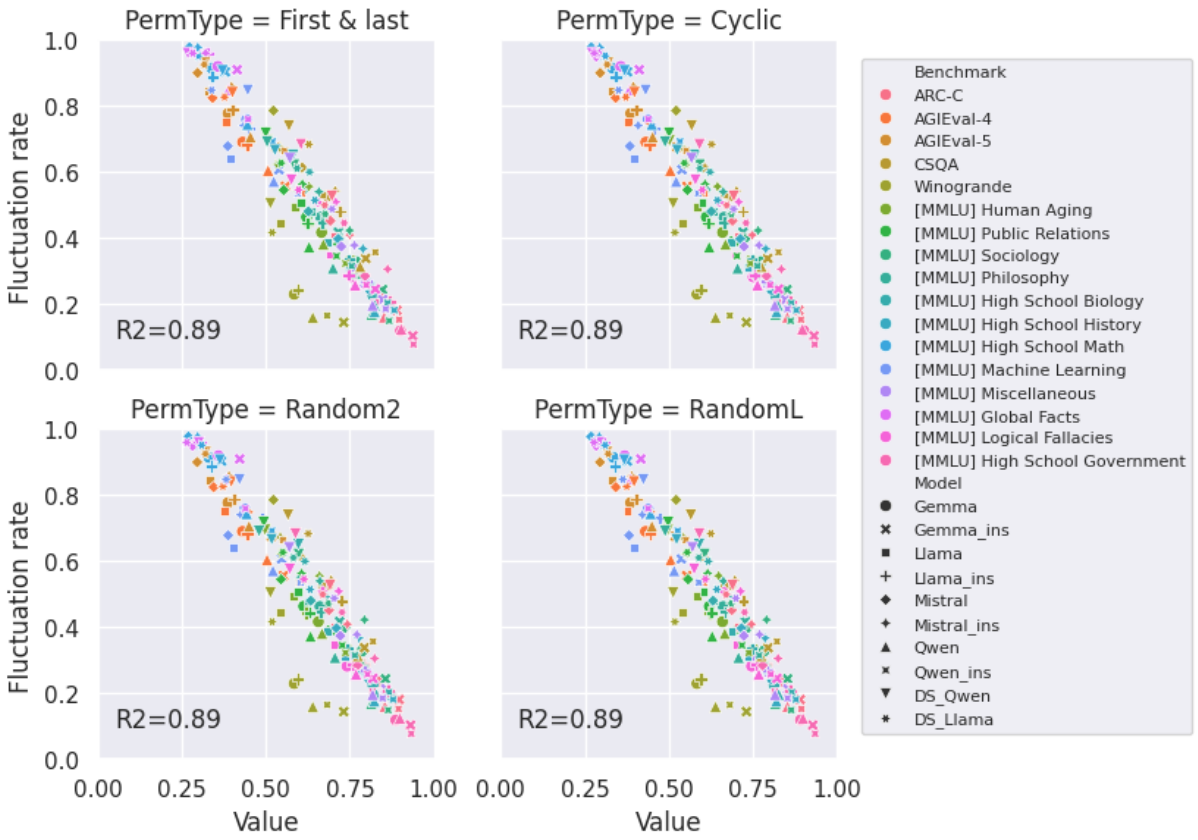


Figure 15: Probability of correct answer on permutation subsets and full fluctuation rates for all tested models and benchmarks.

B Benchmark Details

Benchmark	# Questions	# Options
ARC-C	1165	4
AGIEval-4	1283	4
AGIEval-5	1263	5
CSQA	1221	5
Winogrande	1267	2
MMLU - Human Aging	223	4
MMLU - Public Relations	110	4
MMLU - Sociology	201	4
MMLU - Philosophy	311	4
MMLU - High School Biology	310	4
MMLU - High School History	204	4
MMLU - High School Math	270	4
MMLU - Machine Learning	112	4
MMLU - Miscellaneous	783	4
MMLU - Global Facts	100	4
MMLU - Logical Fallacies	163	4
MMLU - High School Government	193	4

Table 2: Benchmarks used in the experiments, along with the number of questions in each benchmark and the number of options in each question.