Changing Answer Order Can Decrease MMLU Accuracy

Anonymous submission

Abstract

As large language models (LLMs) have grown in prevalence, particular benchmarks have become essential for the evaluation of these models and for understanding model capabilities. Most commonly, test accuracy averaged across multiple subtasks is used to rank models on leaderboards, to determine which model is best for our purposes. In this paper, we investigate the robustness of the accuracy measurement on a widely used multiple choice question answering dataset, MMLU. When shuffling the answer label contents, we find that all explored models decrease in accuracy on MMLU, but not every model is equally sensitive. These findings suggest a possible adjustment to the standard practice of leaderboard testing, where we additionally consider the percentage of examples each model answers correctly by random chance.

Introduction

One of the largest outstanding issues with interpreting the results of model evaluation pertains to the robustness of accuracy measurements. For example, the accuracy of natural language processing models has been shown to be fairly brittle. For example, accuracy can drop when researchers apply input alterations based on paraphrasing (Gan and Ng 2019), word order changes (Gauthier and Levy 2019; Ribeiro et al. 2020; Sinha et al. 2021a, 2022; Allen-Zhu and Li 2023a,b; Berglund et al. 2023; Golovneva et al. 2024; Kitouni et al. 2024; Sugawara et al. 2020), or other minor, largely meaning-preserving input variations or perturbations (Belinkov and Bisk 2018; Ebrahimi et al. 2018; Jiang et al. 2020; Gao, Fisch, and Chen 2021; Li et al. 2021; Sinha et al. 2021b; Moradi and Samwald 2021; Papakipos and Bitton 2022; Qian et al. 2022; Goodarzi et al. 2023; Sinha et al. 2023). If many models fail to be robust on a benchmark, regardless of their initially measured accuracy, we may need to reconsider how we use it as the basis for a leaderboard that actually ranks models.

While there are many approaches to investigating robustness, our approach relies on the intuition that a test-taker, human or model, should always select the right answer regardless of its label, i.e. whether it is listed as answer 'A' or 'C'. Surely, if the right answer is unknown to the test-taker and they make an uneducated guess, they still could happen upon the right answer by chance, but, in an ideal scenario, a true expert should achieve the same score when tested multiple times on versions of a test where only the order that answers are presented in changes.

In humans, this performance stability, often called **test-retest reliability** is an important consideration to determine how to interpret the results of running a test (Bland and Altman 1986). Humans test scores can fluctuate over time, because they are filtered through irrelevant mental or physical factors that affect measurement (Spearman 1910; Dunlap 1933). Such uninformative fluctuations can affect multiple choice tests, for example, when answers are presented in a different order during retest (Krosnick and Fabrigar 1991; Tellinghuisen and Sulikowski 2008; Lions et al. 2022). However, as models do not have the biological limitations of humans, we may expect them to exhibit less variation than humans, or possibly even none at all. Thus, we claim that a model should be robust to answer order changes: if it gets the correct answer to a question when the answer is labeled 'A', it should also always get the correct answer when it is labeled 'C'. Put another way, the model should select the same answer for each question, regardless of its label, for every possible version of a benchmark; its accuracy should be static between test and retest.

In our work, we ask whether shuffling the order of the answer label contents, leaving the order of the labels (A, B, C, D) the same, affects the measurement of accuracy. We keep the question exactly the same while we perform shuffling. We focus our investigation on the MMLU dataset, a popular dataset included on the widely used Hugging Face Open LLM Leaderboard[1], which runs with the Eleuther LM Evaluation Harness (Gao et al. 2023) as its backend.

While testing top performers on the Open LLM Leaderboard, we find that all ten models are affected by our answer shuffling. This indicates that there is serious non-robustness in benchmarking with MMLU. To better rank models on a leaderboard with the MMLU dataset, we may want to take more random shuffles of label contents to better understand the extent to which a model genuinely can output the correct answer. We also found that different categories in MMLU are affected differently by answer order shuffling.

[1] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Methods

MMLU

Massive Multitask Language Understanding (MMLU) is a commonly used benchmark for evaluating LLMs (Hendrycks et al. 2021). It is intended to test a model's world knowledge and problem solving ability, and consists of 57 tasks. Each example in MMLU consists of a question paired with four possible answers, only one of which is correct. Answers are a concatenation of an answer label denoted as a letter, with answer contents (a string of characters). To test the robustness of models to answer choice ordering, we shuffle the answer label contents, with prohibition that the correct answer contents don't change and that we preserve the ordering of MMLU answer labels (A, B, C, D) across different evaluation runs, for example:

original		a possible shuffle	
A. 1		A. 4	
B. 2		B. 2	
C. 3	✓	C. 1	
D. 4		D. 3	✓

We can think of the original orders of answer content labels in each example in MMLU as one of the n (out of 24 possible) shuffles of the example. Given the size of the MMLU dataset, it is not efficient to run all the possible shuffles (as each example has 24 options and there are nearly 14 thousand questions. To do a tractable exploration, we take two random seeds of MMLU, each of which has been shuffled, where each example has been selected from one of the 24 possible answer contents orders to create semantically equivalent versions of MMLU. We utilize the original MMLU implementation (Hendrycks et al. 2021), which uses 5-shot in context learning during evaluation.

Metrics

In essence, we adopt a simplification of the classic formulation of test-retest repeatability from Bland and Altman to match the ML leaderboard setting: an evaluation (the running of a test on a model) is deemed perfectly stable, if and only if the measurements realized at one time of running it produces *the same exact values* when repeated at a later time, when the test is run under the same conditions. We minimally alter the testing conditions when we repeat the test to measure robustness—by changing the order of answer contents—but all other testing parameters remain static. In our setting, we set the number of test takers, n, to 1.

In simple terms, this metric measures how often the model answers the questions correctly in both the original and the shuffled versions. If the model is actually robust, it will select the right answer no matter where it appears, as the answer's meaning doesn't change when you merely change its label and location in the answer string. If the model's accuracy does change in this setting, then we can say the model isn't actually very competent on the task that the test is testing.

To quantify (non-)robustness to answer order shuffling, we define a new metric, *our metric*, which measures how often the model answers the same question(s) correctly in both the original and in a shuffled version of MMLU. We take the average over all the shuffles performed as our metric:

$$\text{Our Metric} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}V_0^i V_j^i, \tag{1}$$

where $V_0^i \in \{0,1\}$ indicates whether the model answers the i^{th} question correctly in MMLU dataset (1 if correct, 0 if incorrect). V_j^i indicates whether the model answers i^{th} question correctly in the j^{th} shuffled version of the answer label content. M is the total number of shuffles in the scope of the experiment (for us 2) and N is the dataset size. We then take the average performance across two such shuffles.

As formulated, our metric tries to capture the true capabilities of the model by reducing the number of questions correctly answered by random chance. Assuming models do not have external memory of earlier queries, enforcing that the model correctly identify the answer M times (for us twice), noticeably lowers the chance of it happening across the correct answer by chance.

Models

In this work, we evaluate 10 state-of-the-art LLMs, ranging in size from 7 billion to 70 billion parameters, most of which have performed very well on the Hugging Face Open LLM leaderboard. The 10 models we use are: Llama3 70B Instruct, Llama3 70B, Llama3 8B Instruct (Meta 2024), Llama2 70B (Touvron et al. 2023), Yi 34B (01.AI et al. 2024), Mixtral 8x7B and Mixtral 8x7B Instruct (Jiang et al. 2024), Falcon 40B Instruct (Almazrouei et al. 2023), Mistral 7B Instruct (Jiang et al. 2023), and Gemma 7B Instruct (Team et al. 2024). All models are openly available, which enables the reproducibility of our findings.

Results

We found that all tested models performed worse according to our metric after answer content shuffling than on the original version of the dataset, as shown in Table 1. After shuffling, we see that models fail to select the correct answer for every question it originally selected correctly, as shown by *our metric* in Figure 1.

Model Name	MMLU	Our Metric	% Drop
Llama-3-70B-it	80.3	75.3	6.2
Llama-3-70B	78.9	72.4	8.2
Yi-34B	75.8	67.7	10.7
Mixtral-8x7B-it	70.6	60.7	14.0
Mixtral-8x7B	70.4	60.9	13.5
Llama-2-70B	69.0	58.8	14.8
Llama-3-8B-it	66.4	58.0	12.7
Mistral-7B-it	59.3	46.5	21.6
Falcon-40B-it	54.7	39.8	27.2
Gemma-7B-it	51.7	38.0	26.5

Table 1: Accuracy drop on MMLU due to changing answer order. Here '-it' marks instruction tuned models.

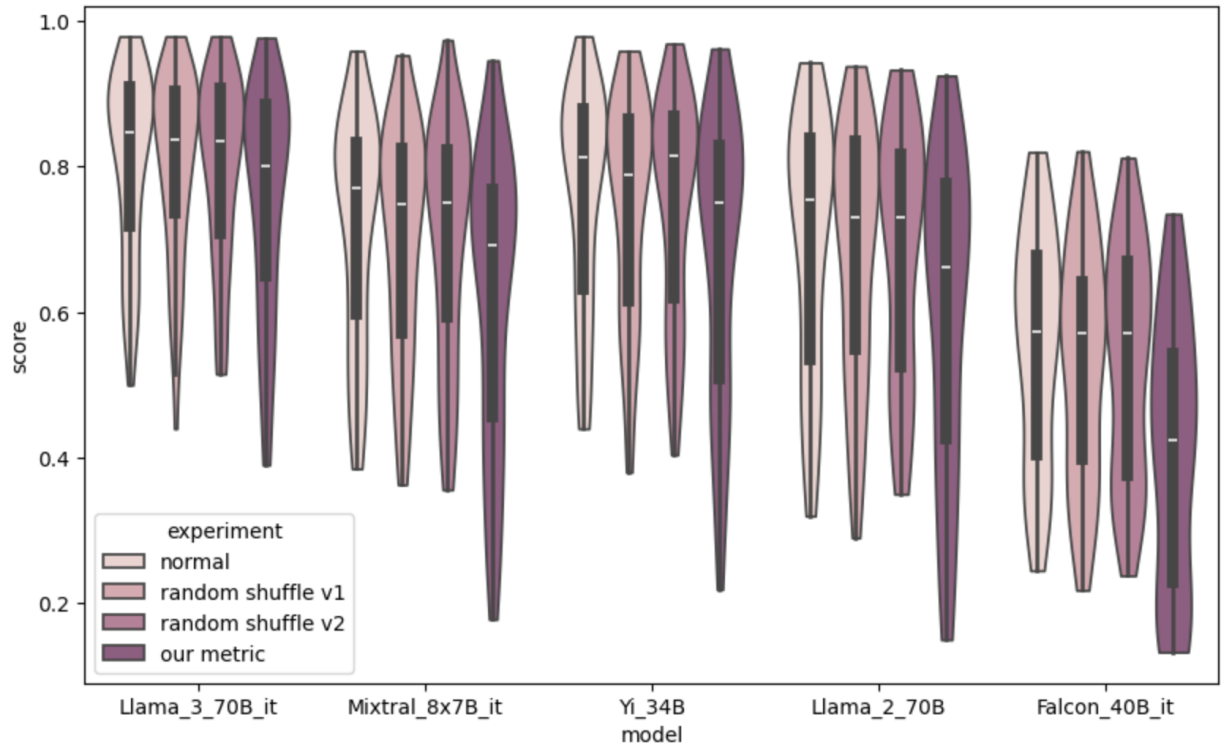


Figure 1: This figure illustrates the performance of a selection of state-of-the-art models that we tested on the original MMLU (v0) and 2 shuffled versions (v1 and v2). Models are ordered by accuracy drop in 'our metric'. Here '-it' denotes an instruction tuned model. The width of the violin corresponds to the number of subdatasets where the model received a particular score. The white indicator marks the median score for subdataset accuracies.

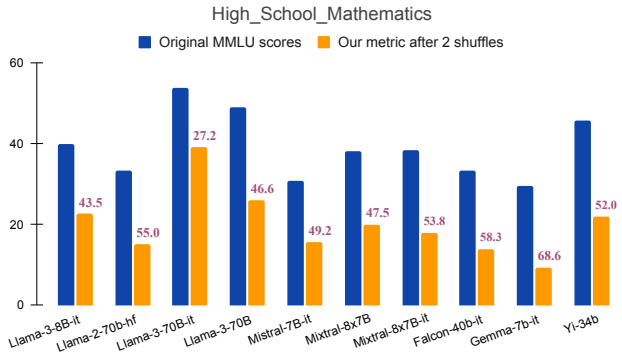
We found that some models had higher retest accuracy than others. Models from the Llama-3 family were the most robust, especially Llama-3-70B for which performance drop was only 6.2%. Interestingly, we found that smaller models can be more robust than larger ones. In particular, we found that Llama-3-8B model was more robust than larger, generally high-performing models such as Mixtral-8x7B and Llama-2-70B. For Llama3-70B and Mixtral-8x7B, we also found that their base and instruction finetuned models were comparably robust. Smaller models, like Mistral-7B and Gemma-7B, were generally more impacted. This result is consistent with findings in (Zhou et al. 2024), which found more inconsistency for smaller models (less than 8B parameters), although in a slightly different setting. Some larger models, such as Falcon-40B-instruct whose score dropped from 54.7 to 39.8 with our approach, were also strongly impacted.

We also analyzed performance drop by subdataset in Table 2, and discovered that the models struggled the most with problem-solving subdatasets, such as high school mathematics. For Gemma-7B and Falcon-40B models, the drop in accuracy on these categories were as high as 40% (as compared to 26% on entire MMLU). As these subdatasets make up a significant portion (over 15%) of original MMLU dataset, this analysis suggests serious robustness issues affecting accuracy scores on problem-solving categories. Additionally, among most impacted subdatasets, such as

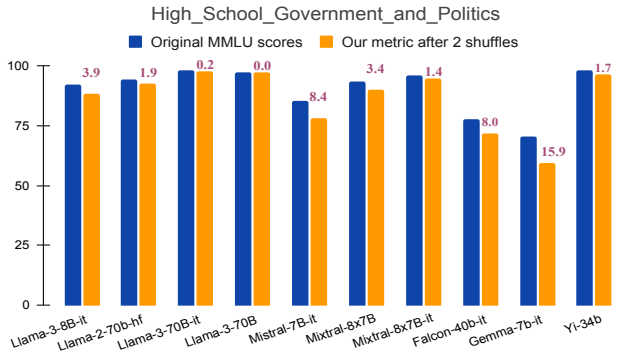
Model Name	MMLU	Our Metric	% Drop
Llama-3-70B-it	72.1	64.5	10.5
Llama-3-70B	68.7	57.7	16.0
Yi-34B	65.6	52.9	19.4
Mixtral-8x7B-it	56.9	43.4	23.7
Mixtral-8x7B	57.0	43.4	23.9
Llama-2-70B	54.6	40.4	26.0
Llama-3-8B-it	54.3	40.9	24.7
Mistral-7B-it	45.2	29.8	34.1
Falcon-40B-it	41.5	24.3	41.4
Gemma-7B-it	38.9	22.2	42.9

Table 2: Accuracy drop on problem solving categories of MMLU dataset due to option text shuffling.

"college mathematics" and "global facts", we investigated whether the drop may be due to the fact that shuffling can ablate the logical order of the original questions. In humans, presenting answer orders in logical order—such as 0,1,2,3 or 3,2,1,0—is recommended by test design research, because random order may pose unnecessary challenge for lower ability students (Huntley and Welch 1993; Haladyna, Downing, and Rodriguez 2002). We discovered that more than 95% of the original MMLU dataset was presented in logical order, which indicates that models may be benefiting from logical answer order and perhaps that they should be



(a) Category of MMLU most affected



(b) Category of MMLU least affected

Figure 2: The most and least affected categories of MMLU with our proposed shuffling. The number above each plot signifies percentage change after shuffling. Here '-it' marks instruction finetuned models.

seen as lower ability test takers.

Related Works

LLMs can be Sensitive to Option Order and Label. Recent works have also shown that the accuracy of models on multiple-choice question datasets can change significantly when the order of answer options is rearranged (Robinson and Wingate 2023; Pezeshkpour and Hruschka 2024; Alzahrani et al. 2024; Wei et al. 2024; Xue et al. 2024; Zong et al. 2024). This suggests that models are sensitive to the order of answer options, which can impact their performance. (Wang et al. 2024a) studies how changing the number of options in mutliple choice question datasets affect the model performance. They found that LLMs have overfitted to multiple choice question datasets with exactly four options.

Other studies have shown that models may exhibit prior biases towards specific option IDs (e.g., 'A') (Wei et al. 2024; Zheng et al. 2023b; Reif and Schwartz 2024; Ross et al. 2024; Li and Gao 2024; Zheng et al. 2023a). Some works have also shown that models can perform surprisingly well above random chance even when question text is removed and only answer options are provided (Balepur, Ravichander, and Rudinger 2024; Shah, Gupta, and Roth 2020; Balepur and Rudinger 2024). Recent works have also shown that replacing the correct option with "None of the Above" leads to a drastic decline in performance across all models (Wang et al. 2024a; Xu et al. 2024). These findings suggest that models may be relying on artifacts or biases in the data rather than truly understanding the questions (Röttger et al. 2024; Raj et al. 2023).

In a concurrent work, McIlroy-Young et al. (2024) proposed a solution fro mitigating the issue of order dependency in LLMs by modifying the self attention matrix of the input sequence. They set the attention scores between different options to be zero, effectively preventing the model from attending to the order of options.

In contrast to above works, our work focuses on category-wise differences in model performance and proposes a new metric that takes into account the variation in model performance across different answer orders. Our approach provides a more nuanced understanding of the impact of answer option ordering on model accuracy.

Evaluation Dataset Validity. For all evaluation datasets, validity is important, and MMLU is no exception. Several recent works have discussed MMLU's validity (Gema et al. 2024; Zheng et al. 2023a; Wang et al. 2024a,b). In particular, Wang et al. (2024b) found trivial and noisy questions in the dataset and proposed an update, MMLU-Pro, which aims to mitigate those issues. Concurrent work on model robustness to question-answering order (Zhou et al. 2024) applies a similar approach to ours that shuffles answer label content and also explores other possible modes of interrogating robustness. While they also find non-robustness to question variants, our work differs from theirs in that our metric can account for the multiplicity of potential orderings of answer labels; we provide further analysis for each category in MMLU in the appendix.

Conclusion

This work tested the robustness of the evaluation benchmark pipeline for the popular leaderboard dataset - MMLU. To separate out the effect of chance on model answers, we apply a largely meaningless change to the datasets by shuffling label contents. We find that this meaning-preserving alteration resulted in a decrease in MMLU accuracy for all models, but not to the same degree. We define a new metric that quantifies the effect of chance and suggest that it is important to take it into consideration during evaluation and leaderboard rankings of models. We also found that different categories in MMLU are affected differently by shuffling label contents.

Limitations

While we explore two possible shuffles of the answer label contents, we restricted ourselves to the M to curtail compute costs. We do acknowledge that there are many more possible shuffles that might be tested, and more would doubtless lead to a better approximation of the non-robustness.

References

01.AI; Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652.

Allen-Zhu, Z.; and Li, Y. 2023a. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

Allen-Zhu, Z.; and Li, Y. 2023b. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Étienne Goffinet; Hesslow, D.; Launay, J.; Malartic, Q.; Mazzotta, D.; Noune, B.; Pannier, B.; and Penedo, G. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867.

Alzahrani, N.; Alyahya, H.; Alnumay, Y.; AlRashed, S.; Alsubaie, S.; Almushayqih, Y.; Mirza, F.; Alotaibi, N.; Al-Twairesh, N.; Alowisheq, A.; Bari, M. S.; and Khan, H. 2024. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13787–13805. Bangkok, Thailand: Association for Computational Linguistics.

Balepur, N.; Ravichander, A.; and Rudinger, R. 2024. Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10308–10330. Bangkok, Thailand: Association for Computational Linguistics.

Balepur, N.; and Rudinger, R. 2024. Is Your Large Language Model Knowledgeable or a Choices-Only Cheater? In Li, S.; Li, M.; Zhang, M. J.; Choi, E.; Geva, M.; Hase, P.; and Ji, H., eds., *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, 15–26. Bangkok, Thailand: Association for Computational Linguistics.

Belinkov, Y.; and Bisk, Y. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2023. The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". *arXiv preprint arXiv:2309.12288*.

Bland, J. M.; and Altman, D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476): 307–310.

Dunlap, J. W. 1933. Comparable tests and reliability. *Journal of Educational Psychology*, 24(6): 442.

Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 31–36. Melbourne, Australia: Association for Computational Linguistics.

Gan, W. C.; and Ng, H. T. 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6065–6075. Florence, Italy: Association for Computational Linguistics.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A framework for few-shot language model evaluation.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830. Online: Association for Computational Linguistics.

Gauthier, J.; and Levy, R. 2019. Linking artificial and human neural representations of language. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 529–539. Hong Kong, China: Association for Computational Linguistics.

Gema, A. P.; Leang, J. O. J.; Hong, G.; Devoto, A.; Mancino, A. C. M.; Saxena, R.; He, X.; Zhao, Y.; Du, X.; Madani, M. R. G.; et al. 2024. Are We Done with MMLU? *arXiv preprint arXiv:2406.04127*.

Golovneva, O.; Allen-Zhu, Z.; Weston, J.; and Sukhbaatar, S. 2024. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*.

Goodarzi, S.; Kagita, N.; Minn, D.; Wang, S.; Dessi, R.; Toshniwal, S.; Williams, A.; Lanchantin, J.; and Sinha, K. 2023. Robustness of Named-Entity Replacements for In-Context Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10914–10931. Singapore: Association for Computational Linguistics.

Haladyna, T. M.; Downing, S. M.; and Rodriguez, M. C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3): 309–333.

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

Huntley, R. M.; and Welch, C. J. 1993. Numerical Answer Options: Logical or Random Order?. In *The Annual of Meeting of the American Educational Research Association*.

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

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8: 423–438.

Kitouni, O.; Nolte, N.; Bouchacourt, D.; Williams, A.; Rabbat, M.; and Ibrahim, M. 2024. The Factorization Curse: Which Tokens You Predict Underlie the Reversal Curse and More. arXiv:2406.05183.

Krosnick, J.; and Fabrigar, L. 1991. The handbook of questionnaire design.

Li, D.; Zhang, Y.; Peng, H.; Chen, L.; Brockett, C.; Sun, M.-T.; and Dolan, B. 2021. Contextualized Perturbation for Textual Adversarial Attack. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5053–5069. Online: Association for Computational Linguistics.

Li, R.; and Gao, Y. 2024. Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions. *arXiv preprint arXiv:2405.03205*.

Lions, S.; Monsalve, C.; Dartnell, P.; Blanco, M. P.; Ortega, G.; and Lemarié, J. 2022. Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education*, 35(2): 133–152.

McIlroy-Young, R.; Brown, K.; Olson, C.; Zhang, L.; and Dwork, C. 2024. Set-Based Prompting: Provably Solving the Language Model Order Dependency Problem. arXiv:2406.06581.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Moradi, M.; and Samwald, M. 2021. Evaluating the Robustness of Neural Language Models to Input Perturbations. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1558–1570. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Papakipos, Z.; and Bitton, J. 2022. Augly: Data augmentations for robustness. *arXiv preprint arXiv:2201.06494*.

Pezeshkpour, P.; and Hruschka, E. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017. Mexico City, Mexico: Association for Computational Linguistics.

Qian, R.; Ross, C.; Fernandes, J.; Smith, E.; Kiela, D.; and Williams, A. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Raj, H.; Gupta, V.; Rosati, D.; and Majumdar, S. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.

Reif, Y.; and Schwartz, R. 2024. Beyond Performance: Quantifying and Mitigating Label Bias in LLMs. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6784–6798. Mexico City, Mexico: Association for Computational Linguistics.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics.

Robinson, J.; and Wingate, D. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*.

Ross, C.; Hall, M.; Romero-Soriano, A.; and Williams, A. 2024. What makes a good metric? Evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Shah, K.; Gupta, N.; and Roth, D. 2020. What do we expect from Multiple-choice QA Systems? In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3547–3553. Online: Association for Computational Linguistics.

Sinha, K.; Gauthier, J.; Mueller, A.; Misra, K.; Fuentes, K.; Levy, R.; and Williams, A. 2023. Language model acceptability judgements are not always robust to context. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6043–6063. Toronto, Canada: Association for Computational Linguistics.

Sinha, K.; Jia, R.; Hupkes, D.; Pineau, J.; Williams, A.; and Kiela, D. 2021a. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2888–2913. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Sinha, K.; Kazemnejad, A.; Reddy, S.; Pineau, J.; Hupkes, D.; and Williams, A. 2022. The Curious Case of Absolute Position Embeddings. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4449–4472. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Sinha, S.; Chen, H.; Sekhon, A.; Ji, Y.; and Qi, Y. 2021b. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. In Bastings, J.; Belinkov, Y.; Dupoux, E.; Giulianelli, M.; Hupkes, D.; Pinter, Y.; and Sajjad, H., eds., *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 420–434. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Spearman, C. 1910. Correlation calculated from faulty data. *British journal of psychology*, 3: 271.

Sugawara, S.; Stenetorp, P.; Inui, K.; and Aizawa, A. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8918–8927.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Sessa, P. G.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.-C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J.-B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; Mao-Jones, J.; Lee, K.; Yu, K.; Millican, K.; Sjoesund, L. L.; Lee, L.; Dixon, L.; Reid, M.; Mikuła, M.; Wirth, M.; Sharman, M.; Chinaev, N.; Thain, N.; Bachem, O.; Chang, O.; Wahltinez, O.; Bailey, P.; Michel, P.; Yotov, P.; Chaabouni, R.; Comanescu, R.; Jana, R.; Anil, R.; McIlroy, R.; Liu, R.; Mullins, R.; Smith, S. L.; Borgeaud, S.; Girgin, S.; Douglas, S.; Pandya, S.; Shakeri, S.; De, S.; Klimenko, T.; Hennigan, T.; Feinberg, V.; Stokowiec, W.; hui Chen, Y.; Ahmed, Z.; Gong, Z.; Warkentin, T.; Peran, L.; Giang, M.; Farabet, C.; Vinyals, O.; Dean, J.; Kavukcuoglu, K.; Hassabis, D.; Ghahramani, Z.; Eck, D.; Barral, J.; Pereira, F.; Collins, E.; Joulin, A.; Fiedel, N.; Senter, E.; Andreev, A.; and Kenealy, K. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

Tellinghuisen, J.; and Sulikowski, M. M. 2008. Does the answer order matter on multiple-choice exams? *Journal of chemical education*, 85(4): 572.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wang, H.; Zhao, S.; Qiang, Z.; Qin, B.; and Liu, T. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *arXiv preprint arXiv:2402.01349*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint arXiv:2406.01574*.

Wei, S.-L.; Wu, C.-K.; Huang, H.-H.; and Chen, H.-H. 2024. Unveiling Selection Biases: Exploring Order and Token Sensitivity in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 5598–5621. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Xu, H.; Lou, R.; Du, J.; Mahzoon, V.; Talebianaraki, E.; Zhou, Z.; Garrison, E.; Vucetic, S.; and Yin, W. 2024. LLMs' Classification Performance is Overclaimed. *arXiv preprint arXiv:2406.16203*.

Xue, M.; Hu, Z.; Zhao, M.; Liu, L.; Liao, K.; Li, S.; Han, H.; and Yin, C. 2024. Strengthened Symbol Binding Makes Large Language Models Reliable Multiple-Choice Selectors. *arXiv preprint arXiv:2406.01026*.

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

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, W.; Wang, Q.; Xu, M.; Chen, M.; and Duan, X. 2024. Revisiting the Self-Consistency Challenges in Multi-Choice Question Formats for Large Language Model Evaluation. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14103–14110. Torino, Italia: ELRA and ICCL.

Zong, Y.; YU, Y. T.; Zhao, B.; Chavhan, R.; and Hospedales, T. 2024. Fool Your Large (Vision and) Language Models with Embarrassingly Simple Permutations.

Appendix

Category Wise Analysis

We analyzed how changing the answer order affects each category in the MMLU dataset. We found that some cate-

gories are more sensitive to these changes than others. Figure 2 shows the impact of answer order changes on eight randomly selected categories.

The MMLU has 57 subcategories, and we observed that some categories are more affected by answer order changes than others. For example, categories such as high school physics, abstract algebra, college mathematics, and moral disputes witnessed a significant decrease in performance after answer order changes. On the other hand, categories such as high school us history, econometrics, and professional law were less affected. In some cases, the impact was highly significant - for instance, the accuracy for Mistral-7B-instruct model on moral scenarios category decreased by 77%, from 31.4 to 7.1, after changing the answer order.

The different plots in Figure 2 highlight that not all categories are equally affected, some parts of MMLU dataset might be good indicator of model performance.

Computation Resources

For all experiments for this work, we utilized 8 V100 32GB GPUs. These GPUs were assembled in a cluster of 8 GPUs in a node. The cumulative computing time required to evaluate all the language models and complete the experiments amounted to approximately 2000 GPU hours.

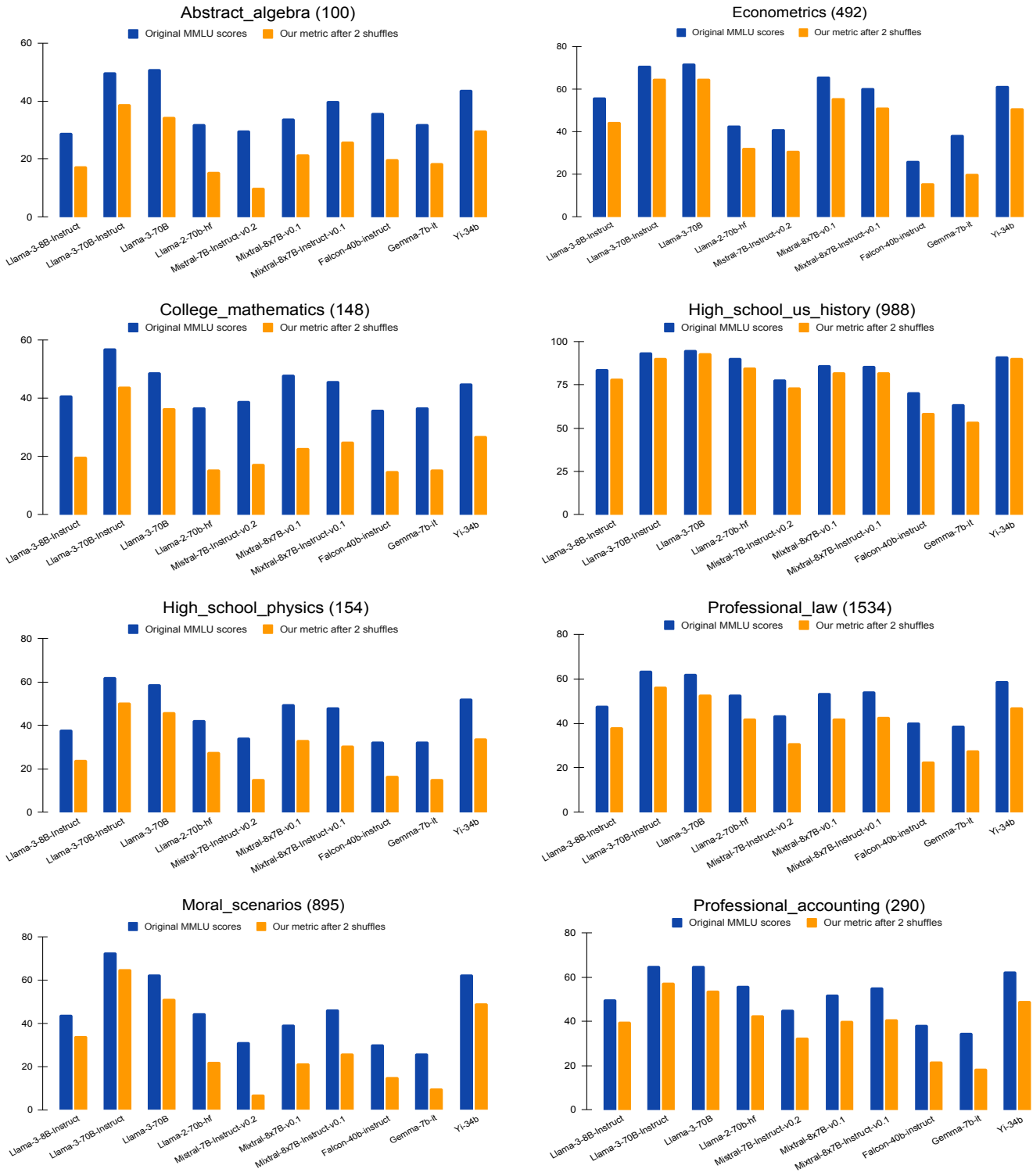


Figure 3: Here we show accuracy scores on random categories of MMLU with our proposed shuffling. The number along with each category name signifies the number of questions for that category in MMLU.