Revisiting the Self-Consistency Challenges in Multi-Choice Question Formats for Large Language Model Evaluation

Wenjie Zhou[1], **Qiang Wang**[2], **Mingzhou Xu**[2], **Xiangyu Duan**[1]
[1]School of Computer Science and Technology, Soochow University, China
[2]Hithink RoyalFlush AI Research Institute
wjzhou223@stu.suda.edu.cn, wangqiangneu@gmail.com, nlp2ct.mzxu@gmail.com
xiangyuduan@suda.edu.cn

Abstract

Multi-choice questions (MCQ) are a common method for assessing the world knowledge of large language models (LLMs), demonstrated by benchmarks such as MMLU and C-Eval. However, recent findings indicate that even top-tier LLMs, such as ChatGPT and GPT4, might display inconsistencies when faced with slightly varied inputs. This raises concerns about the credibility of MCQ-based evaluations. To address this issue, we introduced three knowledge-equivalent question variants: option position shuffle, option label replacement, and conversion to a True/False format. We rigorously tested a range of LLMs, varying in model size (from 6B to 70B) and types—pretrained language model (PLM), supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). Our findings from MMLU and C-Eval revealed that accuracy for individual questions lacks robustness, particularly in smaller models (<30B) and PLMs. Consequently, we advocate that consistent accuracy may serve as a more reliable metric for evaluating and ranking LLMs.

Keywords: Large Language Models, Self-consistency, Evaluation Criterion

1. Introduction

The capacity of world knowledge is an important indicator for evaluating the performance level of Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023). One widely adopted approach for this is the multi-choice questions (MCQ) format, whose typical benchmark representatives are MMLU (Hendrycks et al., 2020), AGIEval(Zhong et al., 2023), and C-Eval(Huang et al., 2023), etc. Using this method, an LLM is presented with a question and four potential answers. If the LLM chooses the correct option, it is typically interpreted as the model having the requisite knowledge.

However, recent studies across various contexts reveal robustness issues in LLMs. They produce inconsistent answers to altered input. This is evident in models like ChatGPT (Jang and Lukasiewicz, 2023; Ohmer et al., 2023) and GPT-4 (Zheng et al., 2023). Consequently, even if an LLM provides a correct answer in the MCQ paradigm, it's uncertain whether the model grasps the content or is swayed by biases (Chang et al., 2023). See Figure 1 for an example.

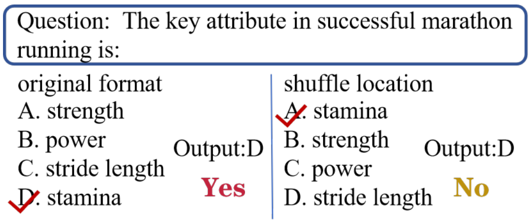


Figure 1: Llama2 13b generates an inconsistent answer by shuffling option positions.

This work spotlights the inconsistency inherent in the MCQ evaluation framework. Given an original question with its corresponding options, we propose three knowledge-equivalent question variants by 1) shuffling the option positions, 2) altering option labels, and 3) transforming the multi-choice question into a judgment question. We prompt the LLM to answer these variants, employing consistent accuracy as our evaluation metric. The goal is to discern whether the model's correct responses stem from genuine knowledge or other factors. Moreover, we comprehensively analyze various LLMs, considering their parameter sizes (ranging from 6 billion to 70 billion) and model types (including PLM, SFT, and RLHF). Results from the MMLU and C-Eval benchmarks show that smaller models and those not enhanced with SFT or RLHF are more prone to inconsistencies in their MCQ evaluations. Notably, platypus2-13b, when assessed in the MMLU task, obtains an accuracy score of 57.1 for the original MCQ. This sharply contrasts with its dip to 37.6 when evaluated for consistency with our tailored variants, leading to a pronounced shift in its comparative ranking. These findings suggest that consistent accuracy could serve as a more reliable metric for evaluating large language models using MCQs.

In summary, our contributions are as follows:

1. Based on MCQ evaluation, we construct

LREC-COLING 2024, pages 14103–14110
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

three types of easily implementable and reproducible question variants from the perspectives of option placement, option symbols, and question format, making the issue of consistency more quantifiable and observable.

2. Our work represents a more comprehensive analysis and evaluation effort, We conducted a more extensive analysis using our proposed consistency metric across three types of models - PLMs, SFTs, and RLHF - and various scales.

3. We analyzed the factors influencing model consistency and explored the effects of consistency, such as its impact on rankings in leaderboards.

2. Related work

LLM evaluation. With the widespread adoption of LLMs, evaluating their capabilities has become especially crucial. MCQ Evaluation, as an assessment method, is preferred due to its definitive answers and straightforward evaluation process. Numerous MCQ benchmarks have been introduced to assess LLMs' capabilities in various areas: world knowledge (e.g., MMLU(Hendrycks et al., 2020) and C-Eval(Huang et al., 2023)), reasoning (e.g., GSM(Cobbe et al., 2021) and BBH(Suzgun et al., 2022)), text toxicity (e.g., Toxigen(Hartvigsen et al., 2022)), and truthfulness (e.g., TruthfulQA(Lin et al., 2022)). Among these, our work specifically emphasizes MMLU and C-Eval with the multi-choice question format due to its streamlined nature compared to free-generation evaluation tasks.

Self-consistency issue. An ideal characteristic of a proficient language understanding model is consistency, the ability to make uniform decisions across semantically equivalent contexts, reflecting its capacity for generalization and comprehension in the face of semantic variations. But numerous studies have found inconsistency problem in LLMs during generation. Elazar and colleagues (Elazar et al., 2021) observed inconsistency in masked LLMs when altering sentence structure without changing meaning and masking the same words. This issue isn't confined to masked LLMs, as models like ChatGPT also face similar problems (Jang and Lukasiewicz, 2023; Ohmer et al., 2023). Differently, this work utilized ChatGPT's robust generative capabilities to create semantically consistent sentences for comparison with the original sentences. However, it was found that even when synonymous sentences were generated by ChatGPT itself, its decisions would often change. Even powerful models like GPT-4 exhibit inconsistency issues. When tasked with evaluating the quality of two sentences, GPT-4 sometimes show a preference for the sentence placed in the first position, regardless of which sentence is positioned there (Zheng et al., 2023).

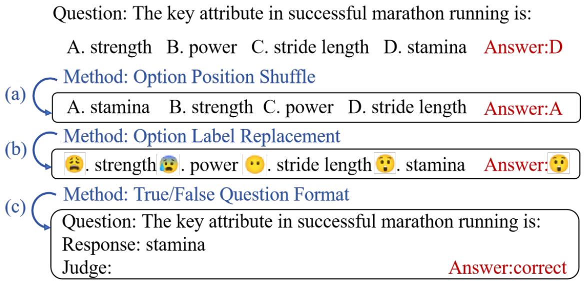

Question: The key attribute in successful marathon running is:
A. strength B. power C. stride length D. stamina Answer:D
(a)
Method: Option Position Shuffle
A. stamina B. strength C. power D. stride length Answer:A
(b)
Method: Option Label Replacement
😩. strength 😰. power 😶. stride length 😲. stamina Answer:😲
(c)
Method: True/False Question Format
Question: The key attribute in successful marathon running is:
Response: stamina
Judge: Answer:correct

Figure 2: Three proposed methods for constructing knowledge-equivalent variants of a multi-choice question.

Unlike previous studies, we introduced three consistency scenarios that include easily measurable consistency metrics, allowing for a broader analysis across various models. Considering the size of parameters and model types, we gained a more detailed understanding of the inconsistency issue.

3. Approach

To assess the self-consistency of LLMs in multi-choice questions, we introduce three strategies for crafting equivalent question variants, as illustrated in Figure 2. For clarity, consider the choices `"A. Venus; B. Mars; C. Jupiter; D. Saturn"`. Here, ”A, B, C, D” are defined as the 'option labels', while the associated responses, such as ”Venus” and ”Mars”, are termed the 'option content'.

(1) Option Position Shuffle. Previous research by (Zheng et al., 2023) indicates a tendency in GPT-4 to favor the first answer when evaluating the quality of two responses. We shuffle the order of choices for each multi-choice question to mitigate potential positional biases. It's important to note that only the option content is altered, leaving the associated labels intact. In the implementation, we assign a distinct random seed for each given question separately to guarantee the same position change across evaluations of different models.

(2) Option Label Replacement. Given the autoregressive generation property of LLMs, the option label precedes the option content. Yet, it remains uncertain whether different option labels influence model behavior. To examine this, we replace common labels like ”A, B, C, D” with uncommon emoji symbols. Compared to shuffle option positions, this method concentrates more on the effect of option labels.

(3) True/False Question Format. While the aforementioned methodologies focus on the presentation of choices, we also explore the question's structure. We posit that if a model genuinely possesses the knowledge, its response accuracy shouldn't waver based on the question format. With this in mind, we convert the standard multi-choice format into individual True/False judgment questions.

(4) Consistency Accuracy Metric Given the above three question variants, we use the consistency-accuracy (CA) as the metric, as defined by:

$$\mathrm{CA}_{\mathcal{V}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{C}_0^{(\mathrm{i})} \prod_{\mathrm{v} \in \mathcal{V}} \mathrm{C}_{\mathrm{v}}^{(\mathrm{i})}, \tag{1}$$

where $\mathcal{V} \subseteq [1, 2, 3]$ denoting the use of which proposed question variants. $C_0^i \in [0, 1]$ indicates whether the model answer correct in the original i-th question, which $C_v^{(i)}$ denotes the result of v-th question variant. When $\mathcal{V} = \varnothing$, $\mathrm{CA}_{\mathcal{V}}$ denotes the standard accuracy. In contrast, $\mathcal{V} = [1, 2, 3]$ denotes the most strict accuracy that only the model gives correct answers in all question variants, it is considered to own the knowledge truly.

4. Experimental results

While the issue of self-consistency has been explored in previous research, the specific factors influencing this for LLMs remain inadequately understood. We suppose that the model's parameter size and the type of model training—be it a pretrained language model (PLM), supervised finetuning model (SFT), or reinforcement learning from human feedback (RLHF)—may be linked to this issue.

4.1. Setup

To validate it, we run experiments using two popular benchmarks for assessing world knowledge: MMLU for English and C-Eval for Chinese. Our study includes a range of widely recognized open-source LLM models encompassing various sizes and types. These models are LLaMA(Touvron et al.), LLaMA-2-chat(Touvron et al., 2023), Platypus2(Lee et al., 2023; Hu et al., 2022), Orca-mini-v3(Mukherjee et al., 2023), Vicuna(Zheng et al., 2023), ChatGLM2(Zeng et al., 2022), WizardLM(Xu et al., 2023), Qwen(Bai et al., 2023), and Baichuan(Baichuan, 2023).

For inference, our approach aligns with the methodology presented in (Liang et al., 2022), using a 5-shot in-context learning setting. The models perform a greedy search across the entire vocabulary, limiting the max-new-tokens parameter to 1. We utilize vLLM for more efficient inference (Kwon et al., 2023).

For option position shuffle, there's no need to adjust the few-shot prompt; just rearrange the options for the question being tested. For option label replacement, we need to synchronize the update of option symbols [A, B, C, D] in every shot within the few-shot prompt to emojis or other symbols. For question format change, each shot in the few-shot prompt needs to be formatted together, with correct or incorrect options randomly selected and labeled correct/incorrect accordingly, then we select the correct option of the question being tested to ask the model whether it is correct or incorrect. The prompt examples we used in experiments can be seen in Appendices.

4.2. Results and analysis

Table 1 presents the accuracy and consistent accuracy results for various knowledge-equivalent questions.

4.2.1. Results of accuracy

Primarily, the accuracy for most models tends to decrease when altering the question format (as seen in ACC_1 and ACC_2) compared to the original multi-choice questions. The decline in accuracy is more pronounced for option label replacement than for position shuffle.

Specifically, Baichuan2 13b on the C-Eval task records ACC and ACC_1 values of 59.1 and 57.7, respectively. However, ACC_2 drops sharply to 41.8, marking a 15.9 point difference from ACC_1. On the other hand, ACC_3 for binary questions varies considerably and doesn't align closely with ACC.

These findings underscore the sensitivity of LLMs to input variations, suggesting that accuracy derived from a single-question format may not comprehensively reflect the model's knowledge capacity.

4.2.2. Results of consistent accuracy

Consider $\mathrm{CA}_{[1,2]}$ as an example in terms of accuracy consistency. Two distinct patterns emerge:

Firstly, consistency's markedly improved as parameter size increases. For context, the $\mathrm{CA}_{[1,2]}$ for LLaMA2 models with 7B, 13B, and 70B parameters are 19.4, 30.3, and 54.2, respectively. This behavior is echoed in SFT models (e.g., platypus2 and orca-mini-v3) and RLHF models like llama2-chat. Additionally, on average, SFT and RLHF mod-

Type	Model	Size	ACC	ACC_1	$CA_{[1]}$	ACC_2	$CA_{[2]}$	ACC_3	$CA_{[3]}$	$CA_{[1,2]}$	$CA_{[1,2,3]}$
				Experimental results on MMLU task							
PLM	MPT	7b	30.7	29.3	10.8	29.7	13.7	null	null	4.9	null
		30b	47.5	46.6	33.6	37.7	26.6	null	null	21.2	null
	Falcon	7b	26.4	25.0	7	25.1	8.1	10.4	11.3	2.0	0.2
		40b	50.0	48.1	36.1	35.7	22.1	24.1	16.8	17.1	7.8
	Llama	7b	35.2	33.5	15.8	27.4	10.5	34.6	15.0	5.2	2.9
		13b	46.9	46.3	31.8	34.8	21.6	40.9	23.5	16.7	10.4
		30b	58.5	56.2	45.3	49.4	39.2	76.7	48.6	32.8	29.1
		65b	63.6	61.2	51.5	58.6	51.2	65.4	48.0	43.9	35.3
	Llama2	7b	45.8	44.5	30.6	38.4	25.5	56.1	30.3	19.4	13.8
		13b	55.7	53.1	41.7	46.5	36.6	33.8	25.2	30.3	17.9
		70b	69.1	67.9	59.9	67.0	60.1	82.8	61.9	54.2	50.1
SFT	ChatGLM2	6b	45.8	45.4	34.1	42.6	33.6	34.3	19.4	27.4	13.3
	Platypus2	7b	50.1	48.5	36.8	44.1	35.0	63.4	36.0	27.7	21.6
		13b	57.1	55.9	46.6	52.1	42.7	60.7	40.6	37.6	29.4
		70b	71.2	70.3	63.8	70.0	65.0	85.5	66.1	60.3	57.4
	Orca-mini-v3	7b	51.6	50.2	39.1	46.3	37.6	54.3	35.3	31.0	24.6
		13b	55.9	54.7	45.1	52.9	45.8	49.4	36.2	39.0	28.8
		70b	70.1	69.1	61.6	69.5	63.4	77.2	61.0	57.7	52.5
	Vicuna-v1.5	7b	49.8	49.7	38.1	45.4	36.9	54.5	35.0	30.0	23.1
		13b	55.7	55.1	45.5	53.0	43.8	28.3	22.8	38.1	19.1
RLHF	Llama2-chat	7b	45.8	46.0	31.5	41.5	33.4	52.6	29.1	24.4	17.9
		13b	53.5	52.5	40.9	50.7	43.4	39.4	28.3	35.2	21.9
		70b	63.0	61.3	51.7	61.8	54.8	59.7	45.9	47.0	38.1
				Experiment results on C-eval task							
PLM	Qwen	7b	59.9	56.8	47.0	24.3	21.8	null	null	20.4	null
		14b	68.1	66.0	56.6	40.4	38.4	null	null	35.3	null
	Baichuan2	7b	54.3	54.9	43.0	25.9	21.9	88.5	28.9	19.7	18.1
		13b	59.1	57.7	47.3	41.8	28.9	56.3	38.2	25.8	18.5
RLHF	Baichuan2-chat	7b	52.7	53.1	41.4	26.2	23.4	52.1	30.3	20.7	12.9
		13b	57.2	56.1	45.9	50.4	41.5	61.1	40.0	36.3	26.9

Table 1: consistent accuracy results for MMLU and C-Eval across various knowledge-equivalent question formats. ACC represents percentage accuracy for the original multi-choice questions, whereas ACC1, ACC2, and ACC$_3$ indicate accuracy for the three specific proposed variants. Bar lengths show percentages compared to the original accuracy. The blue, brown, and orange shades correspond to PLM, SFT, and RLHF model types, respectively.

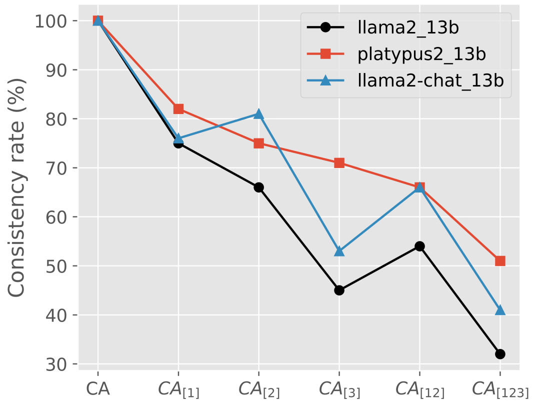


Figure 3: Comparing the consistent accuracy across various model types at a 13b parameter size.

els outperform PLM models in consistency. As evidence, while the average $CA_{[1,2]}$ of platypus2 and LLaMA2-chat are 41.9 and 35.5 respectively, LLaMA2 registers a mere 34.6. A detailed comparison under the 13b model size is illustrated in Figure 3.

These findings underscore potential evaluation pitfalls when using multi-choice questions for LLMs, particularly for smaller PLM models. Furthermore, it's worth noting that certain models, specifically MPT on MMLU and Qwen on C-Eval, exhibit difficulties within the few-shot setting, yielding less meaningful outputs when shifted to a True/False question format. This inability to adapt might suggest that these LLMs, despite having the same parameter sizes, may be less proficient than their counterparts in certain scenarios.

4.2.3. Results of model ranking

Our previous findings show that model accuracy scores fluctuate significantly when subjected to diverse question variants. This fluctuation calls into question the reliability of current leaderboards that rely on multi-choice questioning.

As depicted in Figure 4, we illustrate these ranking variations across various model sizes on MMLU and C-Eval, using metrics like ACC_1, ACC_2, and $CA_{[1,2]}$. Specifically, the 7b and 13b SFT mod-

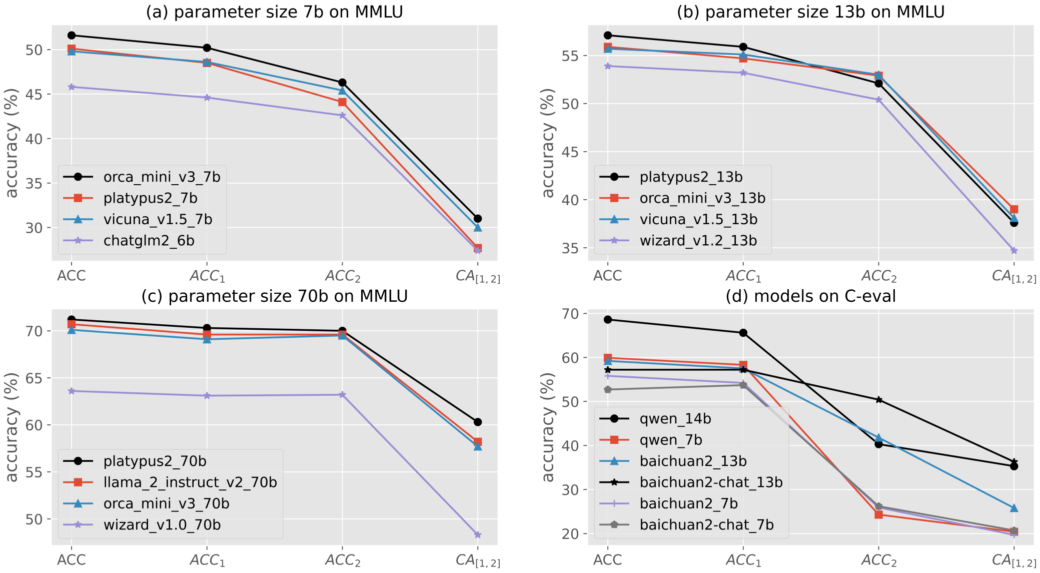


Figure 4: Visual representation of ranking shifts among models under varied metrics. Intersecting line segments signify alterations in rankings.

els are particularly sensitive to minor changes in option positions or labels. For example, the initially top-ranking platypus2-13b in the 13b category slipped two positions due to label alterations or when assessed with stricter metrics—a trend more pronounced in C-Eval. In contrast, the 70b models remain relatively stable in their rankings, with closer scores among them.

Based on these findings, we advise caution when selecting models based on evaluation leaderboards, especially for models with parameters less than or equal to 13B. This underscores the importance of using consistency assessment criteria.

5. Conclusion

In this study, we explore the issue of self-consistency in the multi-choice question format used for evaluating large language models. We introduce three knowledge-equivalent question variants: option position shuffle, option label replacement, and conversion to a True/False question format from a given multi-choice question. Our experiments span model parameter sizes from 7B to 70B and encompass various training types, including PLM, SFT, and RLHF. Findings from the MMLU and C-Eval benchmarks reveal that the input question format significantly impacts accuracy in multi-choice question assessments. This influence is particularly noticeable for models with smaller sizes (less than 30B) and those that are not trained by supervised fine-tuning. To ensure a more robust evaluation of model performance, we recommend adopting consistent accuracy for multi-choice questions.

6. Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62261160648, 62276179).

7. Bibliographical References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt. *arXiv preprint arXiv:2303.06273*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Evaluating task understanding through multilingual consistency: A chatgpt case study. *arXiv preprint arXiv:2305.11662*.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: open and efficient foundation language models, 2023. *URL https://arxiv. org/abs/2302.13971*.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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.

8. Appendices

A. Appendix: Llama Inconsistency Examples with different methods

Original format and question
[few-shot(5-shot) prompt]
The following are multiple choice questions (with answers) about astronomy.
Question1: You are pushing a truck along a road. Would it be easier to accelerate this truck on Mars? Why? (Assume there is no friction)
A. It would be harder since the truck is heavier on Mars.   B. It would be easier since the truck is lighter on Mars.
C. It would be harder since the truck is lighter on Mars.   D. It would be the same no matter where you are.
Answer: D

......(Question2、Question3、Question4、Question5)

The MCQ you want to ask: Why is the sky blue?
A. Because the molecules that compose the Earth′ s atmosphere have a blue-ish color.
B. Because the sky reflects the color of the Earth′ s oceans.
C. Because the atmosphere preferentially scatters short wavelengths.
D. Because the Earth′ s atmosphere preferentially absorbs all other colors.Answer:
Reference Answer: C   Model Response: C

After location shuffle
[few-shot(5-shot) prompt]
The following are multiple choice questions (with answers) about astronomy.
......(Question1、Question2、Question3、Question4 and Question5 remain no change)

The MCQ you want to ask: Why is the sky blue?
A. Because the Earth′ s atmosphere preferentially absorbs all other colors.
B. Because the molecules that compose the Earth′ s atmosphere have a blue-ish color.
C. Because the sky reflects the color of the Earth′ s oceans.
D. Because the atmosphere preferentially scatters short wavelengths.
Answer:
Reference Answer: D   Model Response: A

Figure 5: Example of Option Position Inconsistency on MMLU task

Original format and question
[few-shot(5-shot) prompt]
The following are multiple choice questions (with answers) about clinical_knowledge.
Question1: What is the difference between a male and a female catheter?
A. Male and female catheters are different colours.   B.Male catheters are longer than female catheters.
C. Male catheters are bigger than female catheters.   D.Female catheters are longer than male catheters.
Answer: B

......(Question2、Question3、Question4、Question5)

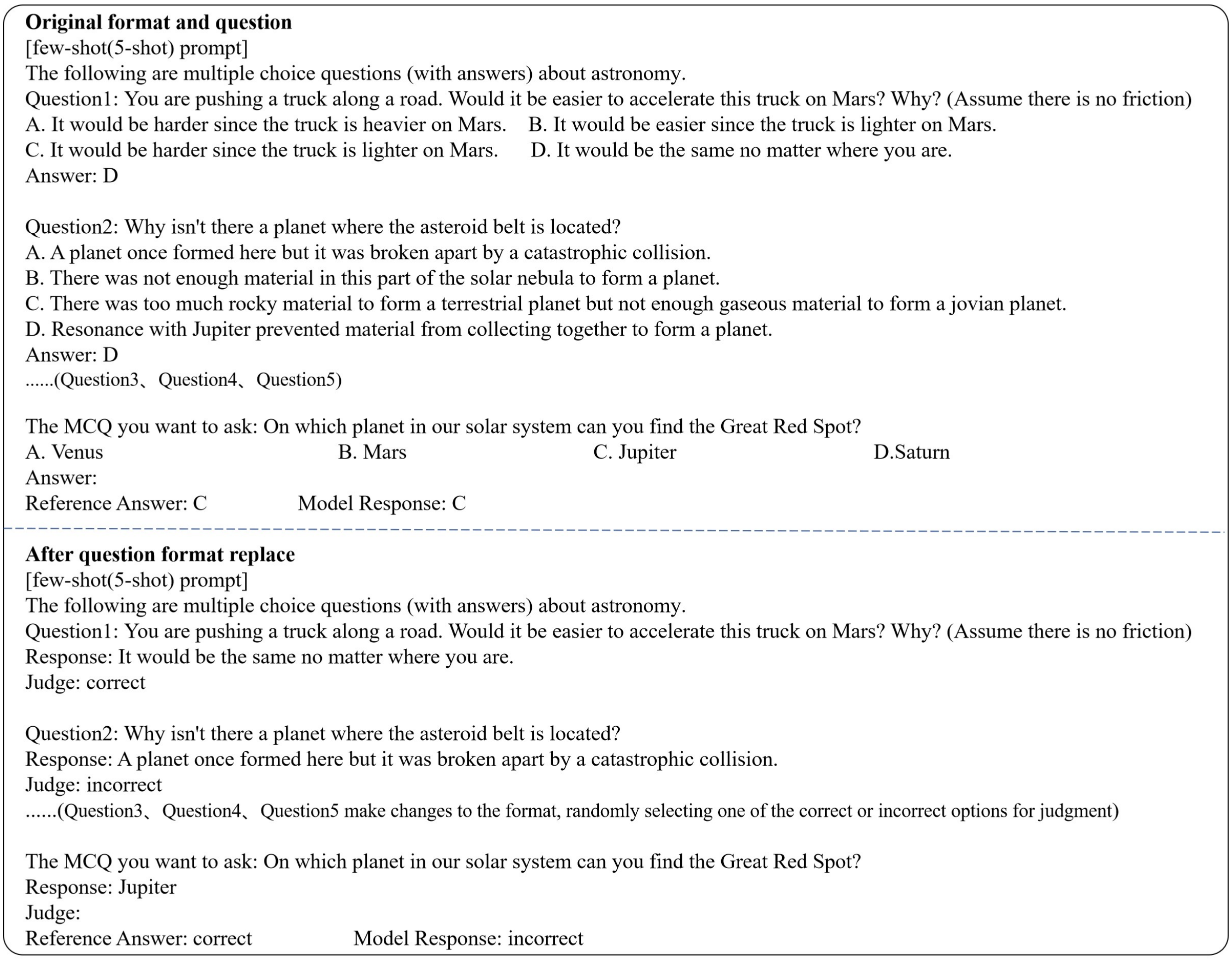
The MCQ you want to ask: The key attribute in successful marathon running is:
A. strength.   B. power.   C. stride length.   D.stamina.
Answer:
Reference Answer: D   Model Response: D

After label replace
[few-shot(5-shot) prompt]
The following are multiple choice questions (with answers) about clinical_knowledge.
Question1: What is the difference between a male and a female catheter?
😩 . Male and female catheters are different colours.   😰 .Male catheters are longer than female catheters.
😶 . Male catheters are bigger than female catheters.   😲 .Female catheters are longer than male catheters.
Answer:

......(Question2、Question3、Question4、Question5 make the same changes to the option labels)

The MCQ you want to ask: The key attribute in successful marathon running is:
😩 . strength.   😰 . power.   😶. stride length.   😲 .stamina.
Answer:
Reference Answer: 😲   Model Response: 😶

Figure 6: Example of Option Symbol Inconsistency on MMLU task



Original format and question

[few-shot(5-shot) prompt]

The following are multiple choice questions (with answers) about astronomy.

Question1: You are pushing a truck along a road. Would it be easier to accelerate this truck on Mars? Why? (Assume there is no friction)

A. It would be harder since the truck is heavier on Mars. B. It would be easier since the truck is lighter on Mars.

C. It would be harder since the truck is lighter on Mars. D. It would be the same no matter where you are.

Answer: D

Question2: Why isn't there a planet where the asteroid belt is located?

A. A planet once formed here but it was broken apart by a catastrophic collision.

B. There was not enough material in this part of the solar nebula to form a planet.

C. There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.

D. Resonance with Jupiter prevented material from collecting together to form a planet.

Answer: D

......(Question3、Question4、Question5)

The MCQ you want to ask: On which planet in our solar system can you find the Great Red Spot?

A. Venus B. Mars C. Jupiter D.Saturn

Answer:

Reference Answer: C Model Response: C

After question format replace

[few-shot(5-shot) prompt]

The following are multiple choice questions (with answers) about astronomy.

Question1: You are pushing a truck along a road. Would it be easier to accelerate this truck on Mars? Why? (Assume there is no friction)

Response: It would be the same no matter where you are.

Judge: correct

Question2: Why isn't there a planet where the asteroid belt is located?

Response: A planet once formed here but it was broken apart by a catastrophic collision.

Judge: incorrect

......(Question3、Question4、Question5 make changes to the format, randomly selecting one of the correct or incorrect options for judgment)

The MCQ you want to ask: On which planet in our solar system can you find the Great Red Spot?

Response: Jupiter

Judge:

Reference Answer: correct Model Response: incorrect

Figure 7: Example of Question Format Inconsistency on MMLU task