Beyond the Surface: Rewording as a Lens for Robust LLM Evaluation

Hillel Darshan **Alon Blyachman** **Omer Mushlion**

{hillel.darshan, alon.blyachman, omer.mushlion} @mail.huji.ac.il

All authors contributed equally to this work.

Abstract

The rapid advancements in large language models (LLMs) have challenged their accurate evaluation, with existing benchmarks often failing to capture their true capabilities. This study investigates the impact of rephrasing questions and answers on LLM performance in multi-choice question answering tasks. We hypothesize that high-performing models should exhibit consistency regardless of phrasing.

Using the MMLU dataset, we selected five verbose categories and extracted a balanced set of 100 questions. Two models, Llama-3-8B-Instruct and Claude-Sonnet-3.5, were employed to rephrase the questions and answers, generating 580 samples. The rephrased datasets were evaluated using the Llama model.

Results show that rephrasing significantly improves accuracy, with the best performance when both questions and answers are rephrased. However, an inverse correlation was found between success and consistency, with more creative rephrasing leading to better performance but lower consistency. We argue that the exact phrasing of benchmarks matters and that LLM evaluation should involve multiple rephrases to assess context understanding rather than specific wording.

This study highlights the importance of rephrasing in enhancing LLM evaluation and provides insights for developing more robust benchmarks. Incorporating rephrasing can yield more accurate performance scores representing a model's true understanding and real-life performance.

1 Introduction

The rapid advancements in the capabilities of large language models (LLMs) have posed significant challenges for their accurate evaluation (Chang et al., 2023). Despite the establishment of numerous benchmarks within a short time frame, the scores obtained on these benchmarks do not consistently reflect the models' performance in real-world applications. Ideally, an intelligent LLM should display resilience to modifications, and would consistently select the same option in a MCQA (multiple choice question answer) tasks, irrespective of the specific phrasing employed , since the knowledge required by the question remains unchanged (Wang et al., 2024). Recent studies have revealed that variations in the formatting of prompts, even minor alterations such as the placement of punctuation marks, can substantially influence a model's performance (Mizrahi et al., 2024). Motivated by these findings, we endeavor to investigate the impact of rephrasing questions and their corresponding answers on a model's ability to generate accurate responses. If we operate under the assumption that a model exhibits a high degree of accuracy, the rephrasing of questions and answers while preserving their semantic content should yield consistent results.

Furthermore, Zhu et al. (Zhu et al., 2024) suggest that rephrasing question-answering benchmark datasets can significantly enhance their resilience against data leakage. When a dataset is compromised, rephrasing can help detect overfitting, where a model's performance deteriorates due to reliance on the original phrasing rather than genuine comprehension. This approach aligns with emerging strategies in LLM evaluation, such as those explored in recent works (Yang et al., 2024; Deng et al., 2024), where the emphasis is on improving model robustness and generalization through diverse and context-aware testing methods. By incorporating rephrasing, these evaluations become more reflective of real-world applications, providing a clearer picture of a model's true capabilities.

The primary objective of this study is to investigate the impact of rephrasing multi-choice questions on model performance. Current evaluation practices are limited by their reliance on fixed benchmarks, which may not adequately capture a model's true capabilities. Our novel approach in-

troduces the rephrasing of questions and answers as a means to assess a model's robustness and generalization ability. The successful implementation of this methodology has the potential to affect the development of better evaluation techniques of LLMs, leading to more reliable and meaningful assessments of their performance in real-world applications.

2 Data

The data utilized in our experiments was derived from the MMLU dataset, a comprehensive benchmark designed to rigorously assess the performance of language models across a wide spectrum of tasks and domains. The MMLU dataset, meticulously curated from diverse real-world sources, ensures its relevance and efficacy in evaluating model performance in practical applications (Hendrycks et al., 2021).

Our focus was directed towards five categories within the MMLU dataset that exhibit a high degree of verbosity, characterized by questions and answers comprising sentences with words amenable to rephrasing. The selected categories encompassed Jurisprudence, Human aging, Professional law, College biology, and Marketing. To ensure a representative sample, we extracted a total of 500 questions, with 100 questions from each of the five chosen subjects. These questions were subsequently presented to the Meta-Llama-3-8B-Instruct model for resolution.

From the model's responses, we meticulously selected 20 questions per subject, strategically composed of 16 questions on which the model failed to provide accurate answers and 4 questions on which the model succeeded. This carefully curated selection process was designed to facilitate the evaluation of the model's performance post-rephrasing, not only on samples that proved challenging for the model but also on those it could successfully address, as illustrated in Table 1.

The rationale behind this selection methodology was to create a balanced dataset that would enable a comprehensive assessment of the impact of rephrasing on the model's performance. This assessment encompassed both the model's ability to rectify previously incorrect responses and maintain accuracy on questions it had already answered correctly. By diligently curating a diverse set of questions from the MMLU dataset, we aimed to ensure the robustness and generalizability of our findings.

3 Methods

Due to computational resource constraints, we selected the Meta-Llama-3-8B-Instruct model for evaluation in our experiments. The evaluation was conducted on multi-choice questions presented to the model, wherein the model was tasked with selecting the correct answer, as illustrated in 8.

To investigate the impact of rephrasing on the model's performance, we employed two models to rephrase the questions and answers of our dataset. The first model, Llama-3-8B-Instruct, was utilized to assess the performance of the model on its own generated rephrasings. The second model, Claude-Sonnet-3.5, a state-of-the-art LLM optimized for creative writing and conversational tasks, was used to evaluate the Llama model's performance on rephrasings generated by a different LLM. This approach allowed us to examine the model's robustness and adaptability to variations in question and answer formulation.

For each question in the dataset, we generated a rephrasing using Claude 3.5 Sonnet, as exemplified in 8. To ensure the integrity of the rephrasing process, we provided additional guidelines (8), specifying that while self-hinting by the model is acceptable, external models should not provide hints when rephrasing the samples. All Claude rephrases where manually approved to make sure no hints where inserted into the rephrases. Due to computational resource limitations, we applied the Meta-Llama-3-8B-Instruct model for rephrasing only on three subjects: Jurisprudence, College biology, and Marketing, resulting in a total of 60 rephrased questions.

Each model was prompted to rephrase both the questions and answers of the samples. From the generated outputs, we created three additional variations: (1) the generated question paired with the generated answers, (2) the original question paired with the generated answers, and (3) the generated question paired with the original answers. This process resulted in an expanded dataset of 580 samples, as detailed in Table 2.

To assess the consistency of the model's performance across the dataset, we evaluated all questions twice using the Llama-3-8B-Instruct model. During inference, a low temperature parameter of 0.01 was applied to ensure a more accurate comparison between the model's responses.

4 Results

The experimental results, as illustrated in Fig. 1 and Table 3, demonstrate that rephrasing significantly improved the model's ability to generate accurate answers. The best overall performance was observed when both questions and answers were rephrased, followed by rephrasing answers only, questions only, and the original data, in descending order of effectiveness. Notably, the rephrasing performed by Claude yielded superior results compared to Llama-3's rephrasing. However, Llama's full rephrasing surpassed the performance of Claude-rephrased questions paired with original answers.

Delving deeper into the data, we investigated not only the model's performance on the MMLU questions but also the consistency of its answers, aiming to assess its robustness to rephrasing. We analyzed the changes within the answers and classified them by dataset. On the second run of Llama-3-8B-instruct on the original dataset, only 2 out of 100 samples exhibited inconsistency, with both initially receiving a wrong answer and subsequently a correct answer on the second run. This finding suggests that, despite potential ambiguity in the original dataset, the model's responses are highly consistent. The use of a low temperature parameter during inference ensured stability and consistency, implying that significant changes in the model's responses should be attributed to the rephrasing of the input rather than the model itself. Across all 580 questions, the overall consistency rate between the first and second runs was 97.24%.

The analysis, presented in Fig. 2 and Table 4, reveals an inverse correlation between overall success on the dataset and the consistency of Llama-3-8B-instruct's performance on the rephrased dataset. Claude's rephrasing of both questions and answers resulted in the model changing its answer in no less than 70% of the samples, more than double the 33% changes observed when using Llama-rephrased questions with original answers. While some of these changes can be explained by the model selecting the correct answer instead of the wrong one (52% of the samples incorrectly answered in the original phrasing were answered correctly with Claude's rephrasing), a substantial 29% of the samples originally answered incorrectly by the model had their answers changed to another incorrect option due to Claude's rephrasing.

5 Analysis

The results indicate that more conservative rephrasing leads to higher model consistency, with the overall score remaining closer to the initial performance. Conversely, as the rephrasing becomes more creative, the overall results improve, but the model's inconsistency increases, not only in terms of the "good" inconsistency of choosing correct answers instead of incorrect ones.

Overall, the findings underscore the potential of rephrasing to enhance the model's performance, with the extent of improvement proportional to the amount of text rephrased. Rephrasing questions alone yields far less benefit compared to rephrasing answers, which can be attributed to the fact that the combined length of the answers typically exceeds that of the question.

We postulate that the negative correlation between rephrasing and consistency may stem from several factors. Some changes could be explained by the assumption that certain samples are ambiguous, leading the model to provide different answers across multiple runs. Another hypothesis is that the rephrasing may not have altered the context of the questions or answers but rather introduced a shift in noise, influencing the model's choices. However, the model's high consistency on the original dataset reduces the likelihood of these hypotheses.

6 Conclusion

Consequently, we assert that rephrasing is the primary driver of the observed differences and that the exact phrasing of the benchmark holds significant importance. To accurately evaluate a language model's ability to understand context rather than merely a specific question wording, testing should involve multiple rephrased versions of the benchmark. Furthermore, having the language model rephrase the input itself before generating a response may yield improved outputs and more accurate performance scores, better representing the model's genuine understanding of the concept and its performance in real-life scenarios. It is important to acknowledge the computational cost associated with this approach, which should be carefully considered. However, it may be possible to prompt a request to rephrase and reply in the same run, such an approach might be less costly, and is left for future research.

7 Limitations and Ethical Concerns

While our study offers valuable insights into the impact of rephrasing on language model performance and consistency, it has limitations that future research should address. The relatively small dataset used in our experiments, with a disproportionate number of initially incorrectly answered samples, may limit the generalizability of our findings. Future studies should consider using larger, differently balanced datasets spanning a wider range of subjects and question types, including those involving mathematical reasoning or domain-specific knowledge. Additionally, normalizing rephrased questions and answers based on length and evaluating the impact of rephrasing across multiple language models could provide a more comprehensive understanding of the relationship between phrasing and model behavior. Investigating the effects of iterative rephrasing and exploring alternative prompts could further elucidate the dynamics between rephrasing and model consistency.

From an ethical standpoint, it is crucial to consider the potential risks associated with rephrasing in real-world applications. Biased language models may amplify or introduce biases during the rephrasing process, leading to unintended consequences. Future research should investigate bias propagation through rephrasing and develop mitigation strategies. Moreover, the computational resources required for large-scale rephrasing and model evaluation raise concerns about energy consumption and carbon footprint. As NLP researchers, we have a responsibility to consider the environmental impact of our work and explore more sustainable approaches to model development and evaluation(Schwartz et al., 2019).

8 Git - Code

https://github.cs.huji.ac.il/alonzo/NLP-project-2024.git

Answer multi-choice question prompt

Answer the following question by selecting the most appropriate choice.
Question: {question}
Choices:
0. {choice 0}
1. {choice 1}
2. {choice 2}
3. {choice 3}
Answer (return only the corresponding number):

Rephrase Prompt example

You are given the following multi-choice question:
Which statement best explains the purpose of Hart's distinction between 'being obliged' and 'having an obligation'?
with the following answers:
0.It demonstrates the difference between the internal and the external aspect of a rule.
1.It refutes the natural lawyer' view of the role of morality in law.
2.It explains the nature of power-conferring rules.
3.It illuminates the concept of a rule.

Please rephrase the question to 50-220 chars.
Please rephrase answer 0 to 50-220 chars.
Please rephrase answer 1 to 50-220 chars.
Please rephrase answer 2 to 50-220 chars.

Claude additions

Don't use the web to find the answer.
Don't hint the answer when rephrasing the question and answers.

Table 1: Data statistics for the selected questions from the MMLU dataset.

Subject	**Model Failed**	**Model Succeeded**	**Selected Questions**
Jurisprudence	16	4	20
Human aging	16	4	20
Professional law	16	4	20
College biology	16	4	20
Marketing	16	4	20
Total	**80**	**20**	**100**

Table 2: Data Set

Original - 100	
Jurisprudence	20
College biology	20
Marketing	20
Human Aging	20
Professional Law	20
Claude Q&A - 100	
Jurisprudence	20
College biology	20
Marketing	20
Human Aging	20
Professional Law	20
Claude Q, Original A - 100	
Jurisprudence	20
College biology	20
Marketing	20
Human Aging	20
Professional Law	20
Original Q, Claude A - 100	
Jurisprudence	20
College biology	20
Marketing	20
Human Aging	20
Professional Law	20
Llama Q&A - 60	
Jurisprudence	20
College biology	20
Marketing	20
Llama Q & Original A - 60	
Jurisprudence	20
College biology	20
Marketing	20
Original Q& Llama A - 60	
Jurisprudence	20
College biology	20
Marketing	20
Total	580

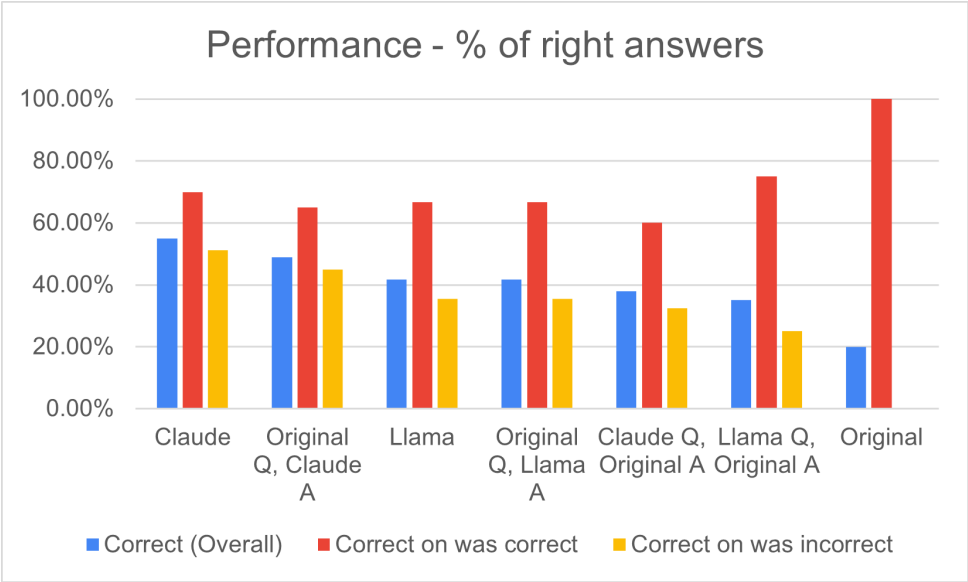

Figure 1: Performance - % of right answers

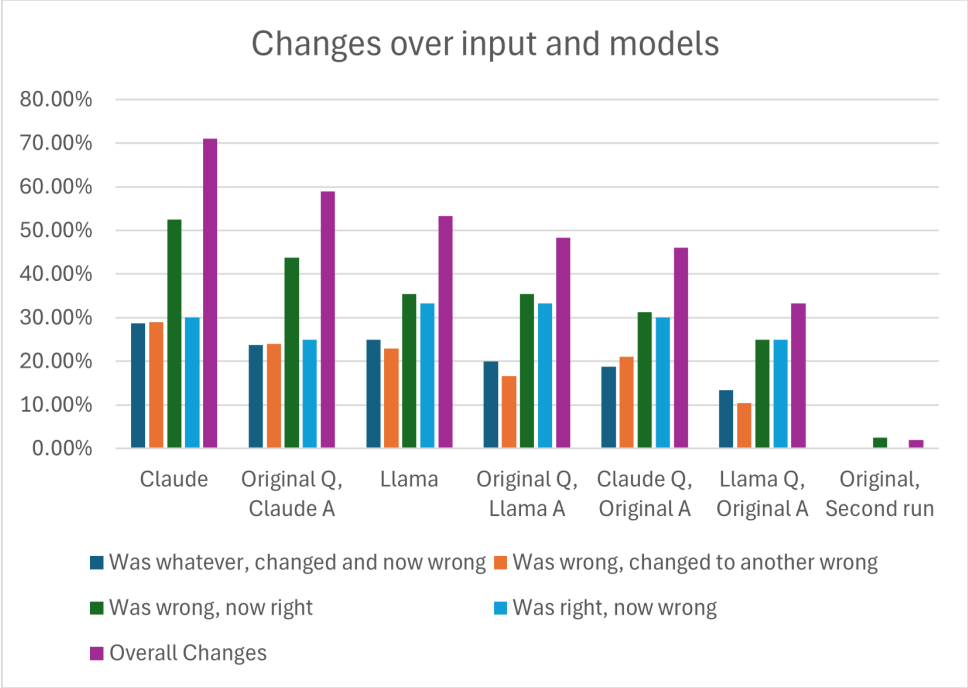

Figure 2: Changes - % of changes in answers

Data Set	Correct (Overall)	Correct on was correct	Correct on was incorrect
Claude	55.00%	70.00%	51.25%
Original Q, Claude A	49.00%	65.00%	45.00%
Llama	41.67%	66.67%	35.42%
Original Q, Llama A	41.67%	66.67%	35.42%
Claude Q, Original A	38.00%	60.00%	32.50%
Llama Q, Original A	35.00%	75.00%	25.00%
Original	20.00%	100.00%	0.00%

Table 3: Performance comparison across different datasets

Model	Was whatever, changed and now wrong	Was wrong, changed to another wrong	Was wrong, now right	Was right, now wrong	Overall Changes
Claude	28.75%	29.00%	52.50%	30.00%	71.00%
Original Q, Claude A	23.75%	24.00%	43.75%	25.00%	59.00%
Llama	25.00%	22.92%	35.42%	33.33%	53.33%
Original Q, Llama A	20.00%	16.67%	35.42%	33.33%	48.33%
Claude Q, Original A	18.75%	21.00%	31.25%	30.00%	46.00%
Llama Q, Original A	13.33%	10.42%	25.00%	25.00%	33.33%

Table 4: Comparison of different models and question-answer combinations across various change categories.

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2024. Rephrase and respond: Let large language models ask better questions for themselves.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models.

Adam Yang, Chen Chen, and Konstantinos Pitas. 2024. Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries.

Qin Zhu, Qingyuan Cheng, Runyu Peng, Xiaonan Li, Tengxiao Liu, Ru Peng, Xipeng Qiu, and Xuanjing Huang. 2024. Inference-time decontamination: Reusing leaked benchmarks for large language model evaluation.