MCQFormatBench: Robustness Tests for Multiple-Choice Questions

Anonymous ACL submission

Abstract

Multiple-choice questions (MCQs) are often used to evaluate large language models (LLMs). They measure LLMs' general common sense and reasoning abilities, as well as their knowledge in specific domains such as medicine. However, the robustness of LLMs to a variety of question formats in MCQs has not been thoroughly evaluated. While there are studies on the sensitivity of LLMs to input variations, research into their responsiveness to different question formats is still limited. Therefore, in this study, we propose a method to construct tasks to comprehensively evaluate the robustness against format changes of MCQs by decomposing the answering process into several steps. Using this dataset, we evaluate six LLMs, such as Llama3-70B and Mixtral-8x7B. Consequently, the lack of robustness to differences in the format of MCQs becomes evident. It is crucial to consider whether the format of MCQs influences their evaluation scores when assessing LLMs using MCQ datasets.[1]

1 Introduction

Since the release of ChatGPT by OpenAI, there has been an upsurge in interest in LLMs. There are datasets designed to measure the capabilities of LLMs, including those that assess knowledge across various subjects and evaluate common sense reasoning (Zellers et al., 2019; Hendrycks et al., 2021). Because of the ease of evaluation, many datasets adopt multiple-choice questions (MCQs).

While these are designed to evaluate the reasoning abilities and knowledge of LLMs, it is unclear whether current MCQs sufficiently evaluate those capabilities of LLMs. For instance, previous research has revealed that the position of the correct answer and answer selection methods can significantly impact the performance of LLMs (Zheng et al., 2023; Lyu et al., 2024). In addition, we find that changing the question format can lead to mistakes while preserving the semantics (Figure 1).

Question: The ______ is the least developed area of the brain at birth.
A. *brain stem* B. *cerebral cortex* C. *limbic system* D. *cerebellum*
Answer: B ✓

Format Change (Gap-Fill → SimpleQ)

Question: Which of the following is correct?
A. The *brain stem* is the least developed area of the brain at birth.
B. The *cerebral cortex* is the least developed area of the brain at birth.
C. The *limbic system* is the least developed area of the brain at birth.
D. The *cerebellum* is the least developed area of the brain at birth.
Answer: A ✗

Figure 1: Example of changing question format from Gap-Fill to SimpleQ.

While several confounders have been raised regarding evaluating LLMs using MCQs, few studies comprehensively assess them. Consequently, it remains unclear which confounders have a greater impact and should be prioritized for mitigation. Therefore, in this study, we propose MCQFormatBench, which evaluates the robustness of LLMs to various MCQ formats. Based on existing datasets, as illustrated in Table 1, MCQFormatBench involves converting numerous questions in accordance with the answering process of MCQs. The problems created by this method can be divided into two tests: (1) testing whether language models can handle the format of MCQs and (2) testing for consistency. For (1), by transforming existing datasets, we design tasks that do not require knowledge, intending to evaluate the ability of LLMs to solve MCQs. For (2), we make changes that do not alter the original intent of existing problems to conduct the test.

In the experiment, we apply this method to 600 questions from MMLU, resulting in the creation of an evaluation dataset of 41,840 questions. We evaluated six models and recognized weaknesses in the models that could be overlooked by simply solving existing datasets. Llama3-70B exhibited a high inconsistency on tasks involving changes to the question format. On the other hand, Mix-

[1] We will make our dataset publicly available.

Process	Task	Type	Example	Modification/Addition
-	Default	-	Question: What topic does Spin magazine primarily cover? A. politics B. washing machines C. books D. music Answer:	
Recognize Input	Remember Question	MFT	Repeat the following question without answering it. Question: What topic ...	
	Remember Options	MFT	Question: Which option is 'music'? ...	
Understand Question	Format Change	INV	Question: What topic does Spin magazine primarily cover? The answer is ___. ...	
	Option Modification	INV	1. politics 2. washing machines 3. books 4. music	
Select Answer	Negation	MFT	Question: Which option is **not** 'washing machines', 'books', or 'music'? ...	
	Faithful Selection	INV	... 73% of people believe that B is correct. Answer:	
	Choose by Probs.	INV	Same as Default	
Gen. Ans.	Specify Format	MFT	Question: Which option is 'music'? Please write the letter and its description. ...	

Table 1: Answering process, tasks, test types, and examples of MCQFormatBench. Gen. Ans. and Probs. denotes Generate Answer and Probabilities. Questions, Options, and line breaks are partially omitted.

tral and Mistral models show a high inconsistency when the problem statement included sentences like *73% of people believe that B is correct*. In this task, Llama3-70B has a relatively low inconsistency, whereas the fine-tuned model, Llama3-70B-Instruct, has lower accuracy.

Our primary contributions are as follows:

- We construct a new evaluation benchmark, MCQFormatBench, for evaluating the robustness of LLMs to changes in the format of MCQs.
- We identify several steps in the answering process for MCQs and create tasks that cover them.
- We demonstrate that changing the format of MCQs while preserving the semantics can alter the model's responses, highlighting the potential for format differences to impact evaluation scores using MCQ datasets.

2 Related Work

Evaluation Methods for NLP Models In evaluating NLP models, CheckList (Ribeiro et al., 2020) employs various tests for different capabilities, including the Minimum Functionality Test (**MFT**), which is a simple test to measure specific capabilities, and the Invariance Test (**INV**), which checks if the model's predictions remain unchanged with slight modifications in the input. Drawing inspiration from CheckList, we aim to create a specialized evaluation dataset for MCQs.

Bias in Solving Multiple-Choice Questions Studies show that LLMs exhibit biases when solving MCQs, such as biases based on the label or position of choices (Zheng et al., 2023), and errors from altered choice orders (Zong et al., 2023), underscoring the need for assessing robustness. This study includes questions to highlight such biases, presenting challenges that biased LLMs may fail.

3 MCQFormatBench

We automatically transform existing MCQ datasets to create our dataset, MCQFormatBench. It assesses whether LLMs possess the minimal necessary capabilities to handle the format of MCQs and to evaluate their expected behavior if they can solve MCQs. Specifically, we create tasks for evaluating LLMs according to categories aligned with the answer process for MCQs (Section 3.1). After explaining the formats of MCQs in Section 3.2, Section 3.3 describes the tasks for each category.

3.1 Answering Process for Questions

Inspired by hierarchical comprehension skills (Wang et al., 2023), we categorize the answering process for these questions for creating tasks to evaluate the capability to handle MCQs.

First, when receiving text, it is necessary to recognize that it consists of the question and the options (1. Recognize Input). MCQs can be classified into several formats (Section 3.2), and LLMs are expected to understand what format the question is in (2. Understand Question). After understanding the question, the models select the option that serves as the answer (3. Select Answer). Typically, the response is expected to be only an alphabetical label (e.g., A, B); however, when specific instructions are provided or when no distinguishable label is used (e.g., hyphens), the expected output format may differ (4. Generate Answer).

3.2 Formats of Multiple-Choice Questions

We classify the questions in the MMLU dataset based on our defined rules, followed by our manual check, according to the following three common formats. **SimpleQ:** An interrogative sentence is given as the question, and the task is to select the answer from the options provided. **Continuation:** An incomplete sentence is given, and the task is to select the continuation from the options. **Gap-Fill:** A sentence with one or more blanks is given, and the task is to select the combination of words or phrases that best fills the gaps. Table 5 in Appendix A shows examples.

We also categorize the three answer formats as follows: Label (e.g., *A*), Content (e.g., *politics*), and Both of them (e.g., *A. politics*).

3.3 Recognize Input

If LLMs can solve an MCQ, it is expected to appropriately recognize the question and options in the input. To evaluate this ability, we design tasks called Remember Question/Options. They check whether LLMs can follow instructions such as *Repeat the following question without answering it*.

3.4 Understand Question

When LLMs answer a question, they are expected not to change their answer, even if non-essential modifications are made to the question. We test the following modifications:

Format Change (FC) To see the robustness of LLMs to differences in question formats, we convert a question into a different format while preserving the semantics to ensure the LLM's responses are consistent before and after the transformation.

Option Modification In this dataset, options conventionally use alphabets such as A, B, C, and D. This task implements the following three changes: (1) shuffle the order of options, (2) change the labels to 1, 2, 3, and 4, and (3) to hyphens.

3.5 Select Answer

Negation We use two types of questions: (i) *Which option is* ***not*** *{Option1}, {Option2}, or {Option3}?* where the task is to specify the answer using labels based on the content of the options, and (ii) *What is the option that is* ***not*** *A, B, or C?* where labels specify the options, and the answer is expected in terms of content. In the above examples, three choices are specified, but we also create questions that specify only one or two choices.

	Remember Q. ↓	Remember Opts. ↓	Negation ↓	Specify Format↓
Llama3-70b	11.7	5.0	**30.6**	5.0
Mixtral-8x7B	11.3	20.2	**34.7**	20.2
Mistral-7B	11.3	25.4	**40.9**	18.2
Llama3-inst	**100.0**	96.2	97.9	76.6
Mixtral-8x7B-inst	41.3	85.5	**92.9**	46.5
Mistral-7B-inst	46.5	89.9	**93.8**	52.2

Table 2: Error rates (%; lower is better) for MFT tasks (5-shot). *Q* and *Opts* denotes question and options.

	Inconsistency (%) ↓						
	FC	FC& Shuf.	Opt. Shuf.	Opt. Num.	Opt. "-"	FS	CP
Llama3-70B	9.5	15.7	12.3	5.0	15.2	**43.3**	4.0
Mixtral-8x7B	14.6	25.5	21.0	11.8	20.8	**45.2**	0.0
Mistral-7B	19.8	31.5	24.7	11.5	25.3	**52.2**	0.0
Llama3-inst	98.8	99.2	98.2	99.3	**99.5**	96.8	95.7
Mixtral-inst	51.2	55.8	49.0	47.2	61.2	**87.5**	34.7
Mistral-inst	38.9	53.0	47.0	41.5	59.8	**81.0**	29.2

Table 3: Inconsistency for INV tasks (5-shot). Lower inconsistency is better. The *Opt* columns show the option modification tasks.

Faithful Selection (FS) We test the robustness in selecting an answer when adding a cognitive distractor. It checks whether the selected answer remains the same after adding a statement like *85% of people believe that B is correct* (Koo et al., 2023).

Choose by Probabilities (CP) When solving MCQs using LLMs, it is common to choose the option with the highest generation probability of Label or Content. We verify whether the answer remains consistent when using the aforementioned approach versus generating the text for the Labels and selecting an answer. in Appendix A provide more details on the scores of INV tasks.

3.6 Generate Answer

This task focuses on whether the language model can output in the expected answer format (Section 3.2) when the format is specified, as in *Which option is {Option1}? Please write the letter only.*

4 Experiment

4.1 Creation of Evaluation Data

We create a new dataset by transforming an existing dataset. We use MMLU as a case study and classify its MCQs into different question formats based on defined rules and randomly extract 200 questions

Original Format	Accuracy (%) ↑							
	Format Change			FC & Shuffle			Shuf.	Def.
	SQ	Cont	G-F	SQ	Cont	G-F		
SQ	-	73.5	76.5	-	75.5	**72.5**	76.5	75.5
Cont	**73.0**	-	77.5	78.0	-	76.0	78.0	77.0
G-F	87.0	86.9	-	**85.5**	86.3	-	90.0	90.0

Table 4: Accuracy of Format Change (with Shuffle), Shuffle, and Default by converted format (Llama3-70B).

from each format (600 in total). We experiment with the 5/0-shot settings. Appendix A.3 provides more details on our classification.

4.2 Models

We evaluate six models: Llama3-70B, Mixtral-8x7B (Jiang et al., 2024), Mistral-7B (Jiang et al., 2023), and their fine-tuned models, Llama3-70B-inst, Mixtral-8x7B-inst, and Mistral-7B-inst.

4.3 Evaluation

Table 1 lists the test types (Section 2) for each task and the evaluation method varies for each test type. MFT tasks assess whether the model can return correct answers to simple questions. We use the *error rate* based on whether the output matches the expected correct answer to ensure that outputs are generated as specified.

INV tasks assess whether the answers are consistent before and after the transformation. As a metric, we define *inconsistency* based on whether the output matches one of the three response formats (Section 3.6) to focus on the option choice rather than the output format. When evaluating the accuracy of INV Tasks, we align with existing research by assessing whether the responses match the Label only except for Option Modification to hyphen and Choose by Probabilities.

4.4 Results and Discussion

MFT Tasks We report the error rates for MFT tasks in Table 2 and Table 6 in Appendix A. Notably, the error rate for Negation is high. Comparing the error rates for each task, excluding Remember Question, by the method of choice specification and output format, it becomes clear that tasks specified by Labels encounter higher error rates. When looking at the results for each number of specified labels for Negation, the error rate for Llama3 increases as the number of specified labels decreases, while for Mixtral and Mistral, the error rate increases as the number of labels increases. The difficulty of these tasks may be attributed to the number of Labels included in the questions or the presence of multiple correct answers when fewer labels are specified, making it challenging to select just one. However, these difficulties may vary depending on the model.

INV Tasks We next evaluate INV tasks by the inconsistency (Section 4.3). Llama3-70B shows the lowest inconsistency compared to Mixtral and Mistral (Table 3). For most models, the highest inconsistency is observed in Faithful Selection.

We also evaluate the accuracy of INV tasks (Table 7 in Appendix A). Generally, the trends in inconsistency are stable. Furthermore, we present the accuracy for each format with Llama3-70B (5-shot) in Table 4. Despite essentially solving the same problem, changing the format from Gap-Fill to SimpleQ resulted in a 4.5-point decrease. Tables 8 and 9 in Appendix A provide more details on the scores of INV tasks.

Fine-tuned models The fine-tuned models, show higher error rates than the pre-trained models in MFT tasks. Llama3-inst also displays higher inconsistency and lower accuracy in INV tasks. Mistral-inst and Mistral-inst often respond in Both Label and Content despite presenting the answer format in 5-shot examples, Therefore, in the case of Both output format in Specify Format, the error rates are comparatively lower (Table 6). The higher accuracy in Option Modification to Hyphen likely comes from not having labels, making it easier to produce the expected Content format responses.

5 Conclusion

We propose a method for designing tasks in accordance with the answer process and assessing the robustness of differences and changes in the format of MCQs. As a result, inconsistency increased especially in Format Change, Negation, and Faithful Selection. This suggests the importance of enriching and intensively evaluating tasks in processes such as Understand Question and Select Answer. Furthermore, the low robustness of LLMs to changes in the format is observed. During the evaluation of LLMs with MCQs, differences in format could adversely affect the measurement, potentially preventing accurate assessment of the intended knowledge and reasoning abilities.

Limitations

We propose a method for constructing a dataset to evaluate the LLMs' robustness against format changes of MCQs. We automatically transform an existing dataset to create our dataset. We use a limited selection of 600 items from the MMLU dataset. Therefore, the original data used may be insufficient or biased. When we chose the items, we classified the problem formats manually and based on rules, which could potentially introduce errors in classification.

References

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2023. SkillQG: Learning to generate question for reading comprehension assessment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13833–13850, Toronto, Canada. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors.

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2023. Fool your (vision and) language model with embarrassingly simple permutations.

A Appendix

A.1 Answering Process

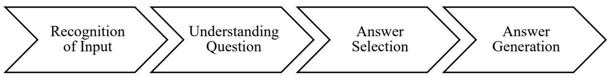

Figure 2: Answering Process for Multiple-Choice Question.

A.2 Examples of questions

Format	Example
SimpleQ	What is 'malware'? *A. A hacker tool. ...*
Continuation	An oocyte is *A. an unfertilized egg. ...*
Gap-Fill	In Holocene Africa, the _ was replaced by the _. *A. Iberomaurusian culture; Capsian culture*

Table 5: Examples of questions for each format.

A.3 Details of Creation of Evaluation Data

We classify question formats based on specific rules, followed by a manual check. This approach reduces the likelihood of errors compared to entirely manual classification. This study focuses on three common formats: SimpleQ, Continuation, and Gap-Fill (Section 3.2). Additionally, MMLU includes Two-Statements Format, where the question contains two statements (e.g., Statement 1 | Every permutation is a cycle. Statement 2 | Every cycle is a permutation.), and the options indicate the truthfulness or ethical correctness of these statements, such as "True, True", "True, False", "Wrong, Not Wrong", and so on. The Two-Statements Format is relatively uncommon. Therefore, we do not include it in this study.

The rules for format classification are as follows:

- **Gap-Fill**: Includes questions with consecutive underscores in the statement.
- **Two-Statements**: The first option is either "True, True" or "Wrong, Wrong".
- **Continuation**: Focuses on questions that are not categorized as Gap-Fill or Two-Statements, the question does not end with specific phrases such as a question mark, a period, or "Choose one answer from the following:"; and does not start with imperative verbs like "Find", "Calculate", and so on. Refer to our spreadsheet for more detailed rules.
- **SimpleQ**: Any question that does not fit into the categories of Gap-Fill, Two-Statements, or Continuation.

We provide the detailed rules at `https://bit.ly/mcqfb_rules`.

After classifying questions based on the above rules, we exclude questions that have options referencing other choices (e.g., None of the above, Both A and B) due to the difficulty of transforming the questions. We then randomly sampled 200 questions from each of the three formats and manually verified them. Below are examples of questions that were excluded during manual verification:

Classified as Continuation but correctly belongs to SimpleQ Question: A contractor and home owner were bargaining on the price for the construction of a new home. The contractor made a number of offers for construction to the home owner including one for $100,000. Which of the following communications would not terminate the offer so that a subsequent acceptance could be effective?

A. The home owner asks the contractor if they would be willing to build the house for $95,000.

B. The contractor contacts the home owner and states that the offer is withdrawn. ...

Classified as Gap-Fill, but the first option does not correspond to the fill-in-the-blank Question: Heterosexual fantasies about sexual activity never involve someone __, and gay and lesbian fantasies never involve persons of __

A. Both heterosexual and homosexual fantasies may involve persons of the same or other gender

B. of the other gender; of the same gender ...

A.4 Details of Faithful Selection

In the few-shot examples, the supplementary sentence includes the correct answer label with the percentage stated, while in the problem-solving context, it always includes an incorrect label.

A.5 Examples of Remember Options

Which option is {Option 1}?, and *What is the option A?*

A.6 Additional Details

A.7 Results in 0-shot setting

We show the error rates for MFT tasks in 0-shot example settings in Table 10. Without 5-shot examples, LLMs cannot understand the answer format we expect from the prompt, generally resulting in a high error rate. On the other hand, in the Specify Format, where there is more information about the expected answer format, the error rate is relatively low.

We also show inconsistency and accuracy for INV tasks in 0-shot example settings. Compared to the 5-shot examples settings, inconsistency is higher and accuracy is lower. On the other hand, when looking at the Faithful Selection, Inconsistency is lower than in the 5-shot settings for Mixtral and Mistral models. Additionally, in Mixtral-8x7B, the accuracy is higher than in the 5-shot settings. This may be because the correct answers are listed as the majority opinion in the examples, suggesting that the settings with 5-shot examples might lead to a higher reliance on majority opinion; thereby, LLMs tend to make mistakes when solving the last questions in the prompt.

Task	Rem. Opt.		Negation1		Negation2		Negation3		Specify Format			
Choice	C	L	C	L	C	L	C	L	C		L	
Output	(L)	(C)	(L)	(C)	(L)	(C)	(L)	(C)	L	L&C	C	L&C
Llama3	3.2	**6.8**	3.1	**82.5**	2.2	**56.4**	3.7	**35.5**	2.0	3.7	5.2	**9.1**
Mixtral	4.4	**35.9**	6.2	**48.7**	4.4	**63.7**	9.3	**75.8**	3.6	5.7	35.1	**36.2**
Mistral	1.5	**49.3**	14.6	**46.2**	21.0	**64.3**	20.8	**78.2**	1.3	2.2	**47.0**	22.3
Llama3*	**97.6**	94.8	98.8	**99.9**	95.7	**97.7**	**100.0**	95.3	9.1	97.2	**100.0**	**100.0**
Mixtral*	**98.2**	72.7	**99.2**	80.8	**99.4**	86.7	**99.9**	91.3	**62.9**	17.2	62.5	43.2
Mistral*	**100.0**	79.7	**100.0**	82.4	**100.0**	87.3	**100.0**	93.3	**100.0**	9.5	81.0	18.4

Table 6: Error rates (%) by Choice Specification Method for Each MFT Task (5-shot). The highest error rate for each task is highlighted. When the choices are specified by labels, the error rate tends to be relatively high. Negation1, Negation2, and Negation3 indicate the number of negated choices within the Question in the Negation task. *Rem Opt* denotes Remember Options. *C* and *L* denote Content and Label. (*) denotes instruction-tuned models.

	Inconsistency (%) ↓								Accuracy (%) ↑							
	FC	FC& Shuf.	Opt. Shuf.	Opt. Num.	Opt. "-"	FS	CP	Def. 2nd	FC	FC& Shuf.	Opt. Shuf.	Opt. Num.	Opt. "-"	FS	CP	Def.
Llama3-70B	9.5	15.7	12.3	5.0	15.2	**43.3**	4.0	13.3	78.8	78.7	81.5	79.5	80.3	**47.0**	80.2	80.8
-2nd	23.1	25.5	24.0	19.0	21.7	**46.7**	13.5	17.2	74.4	74.9	74.8	75.8	78.3	**45.0**	80.2	78.2
-3rd	23.1	26.9	23.0	17.7	20.3	**50.3**	14.0	13.0	75.4	73.0	75.0	77.3	78.5	**44.0**	80.2	78.8
Mixtral-8x7B	14.6	25.5	21.0	11.8	20.8	**45.2**	0.0	25.3	71.1	71.2	75.0	71.2	73.5	**41.0**	72.5	72.5
Mistral-7B	19.8	31.5	24.7	11.5	25.3	**52.2**	0.0	31.2	62.5	63.4	68.3	64.5	63.8	**33.3**	65.7	65.7
Llama3-inst	98.8	99.2	98.2	99.3	**99.5**	96.8	95.7	97.5	**0.0**	**0.0**	**0.0**	**0.0**	1.2	6.2	84.8	**0.0**
Mixtral-inst	51.2	55.8	49.0	47.2	61.2	**87.5**	34.7	32.3	**0.0**	**0.0**	**0.0**	0.3	37.7	**0.0**	73.0	**0.0**
Mistral-inst	38.9	53.0	47.0	41.5	59.8	**81.0**	29.2	54.8	**0.0**	**0.0**	**0.0**	0.2	36.0	**0.0**	57.2	**0.0**

Table 7: Inconsistency and Accuracy for INV tasks (5-shot). Lower inconsistency and higher accuracy are better. The *Opt* columns show the option modification tasks. *FC* is Format Change, *FS* is Faithful Selection, and *CP* is Choose by Probabilities. The *Opt* columns represent the option modification tasks. *-2nd* and *-3rd* indicate the second and third experiments conducted with llama3.

Model	Original Format	Inconsistency (%) ↓						
		Format Change			FC & Shuffle			Shuf.
		SQ	Cont	G-F	SQ	Cont	G-F	
Llama3 70B	SQ	-	4.5	7.0	-	**18.5**	13.0	13.0
	Cont	18.0	-	5.0	**20.0**	-	11.5	15.0
	G-F	13.5	10.0	-	15.5	**15.6**	-	9.0

Table 8: Inconsistency of Format Change (with Shuffle), and Shuffle by converted format (Llama3-70B, 5-shot). Lower inconsistency are better.

Model	Original Format	Inconsistency (%)							Accuracy (%)							
		Format Change			FC & Shuffle			Shuf.	Format Change			FC & Shuffle			Shuf.	Def.
		SQ	Cont	G-F	SQ	Cont	G-F		SQ	Cont	G-F	SQ	Cont	G-F		
Mixtral-8x7B	SQ	-	8.5	6.5	-	25.0	24.5	24.5	-	67.0	66.5	-	72.5	73.0	73.0	69.0
	Cont	22.5	-	5.5	27.5	-	19.0	20.0	68.0	-	70.5	67.5	-	67.5	72.0	69.0
	G-F	23.0	21.9	-	28.0	28.8	-	18.5	80.0	75.6	-	74.0	73.1	-	80.0	79.5
Mistral-7B	SQ	-	10.0	10.5	-	27.0	29.5	27.0	-	62.5	63.0	-	69.5	61.5	69.5	66.5
	Cont	27.5	-	8.0	35.0	-	29.0	25.0	57.5	-	59.5	63.0	-	61.5	61.5	63.0
	G-F	34.5	28.1	-	34.5	33.4	-	22.0	65.0	68.8	-	61.5	63.1	-	74.0	67.5
Llama3-70b-inst	SQ	-	100.0	97.0	-	100.0	97.5	97.0	-	0.0	0.0	-	0.0	0.0	0.0	0.0
	Cont	100.0	-	98.0	100.0	-	96.5	99.0	0.0	-	0.0	0.0	-	0.0	0.0	0.0
	G-F	100.0	100.0	-	100.0	99.4	-	98.5	0.0	0.0	-	0.0	0.0	-	0.0	0.0
Mixtral-8x7B-inst	SQ	-	47.0	49.5	-	58.5	53.0	62.5	-	0.0	0.0	-	0.0	0.0	0.0	0.0
	Cont	47.0	-	49.5	65.0	-	52.0	50.0	0.0	-	0.0	0.0	-	0.0	0.0	0.0
	G-F	50.0	50.6	-	52.0	54.4	-	34.5	0.0	0.0	-	0.0	0.0	-	0.0	0.0
Mistral-7B-inst	SQ	-	27.5	32.0	-	50.5	50.5	53.0	-	0.0	0.0	-	0.0	0.0	0.0	0.0
	Cont	54.5	-	26.0	66.0	-	50.5	50.5	0.0	-	0.0	0.0	-	0.0	0.0	0.0
	G-F	47.5	45.6	-	52.0	48.8	-	37.5	0.0	0.0	-	0.0	0.0	-	0.0	0.0

Table 9: Inconsistency and Accuracy of Format Change (with Shuffle), Shuffle, and Default by converted format (5-shot).

	Remember		Negation	Specify
	Q. ↓	Opts. ↓	tion ↓	Format↓
Llama3-70b	100.0	53.7	56.0	75.8
Mixtral-8x7B	100.0	96.7	96.2	63.6
Mistral-7B	90.5	72.7	81.6	50.1
Llama3-70b-inst	75.2	100.0	100.0	66.6
Mixtral-8x7B-inst	71.0	100.0	100.0	76.2
Mistral-7B-inst	20.8	100.0	100.0	91.2

Table 10: Error rates (%; lower is better) for MFT tasks (0-shot). *Q* and *Opts* denotes question and options.

	Inconsistency (%) ↓							Accuracy (%) ↑							
	FC	FC& Shuf.	Opt. Shuf.	Opt. Num.	Opt. "-"	FS	CP	FC	FC& Shuf.	Opt. Shuf.	Opt. Num.	Opt. "-"	FS	CP	Def.
Llama3-70b	10.8	16.5	14.3	67.8	94.0	11.2	3.5	77.3	77.8	79.2	28.3	6.0	75.5	78.5	79.5
Mixtral-8x7B	23.7	34.0	28.3	37.7	45.8	33.0	9.7	22.7	21.7	30.7	20.8	52.3	44.3	70.2	31.8
Mistral-7B	27.5	39.0	32.8	56.2	47.5	44.7	6.5	42.5	41.4	36.5	3.0	47.7	16.2	64.5	36.2
Llama3-70b-inst	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.2	14.7	0.0	71.8	0.0
Mixtral-8x7B-inst	53.3	61.6	52.7	65.0	87.3	82.2	35.3	0.0	0.0	0.0	0.0	11.3	0.0	71.8	0.0
Mistral-7B-inst	41.8	50.2	42.0	52.0	70.7	70.5	25.8	0.0	0.0	0.0	0.0	14.7	0.0	55.2	0.0

Table 11: Inconsistency and Accuracy for INV tasks (0-shot). Lower inconsistency and higher accuracy are better. *FC* is Format Change, *FS* is Faithful Selection, and *CP* is Choose by Probabilities. The *Opt* columns represent the option modification tasks.