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 law and medicine. However, the robustness of LLMs to various 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 eight LLMs, such as Llama3-70B and Mixtral-8x7B. We find the lack of robustness to differences in the format of MCQs. 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, large language models (LLMs) have drawn widespread interest. In advancing LLM research and development, there is a critical need to quantitatively evaluate the various capabilities of these models, such as knowledge across various subjects and common sense reasoning (Clark et al., 2018; Dua et al., 2019; Zellers et al., 2019; Geva et al., 2021; Hendrycks et al., 2021a; Sakaguchi et al., 2021; Rein et al., 2023). For such quantitative evaluation, multiple-choice questions, which expect discriminative answers, are widely adopted across many datasets.

While these datasets are designed to evaluate LLMs' reasoning abilities and knowledge, it remains unclear whether current MCQs sufficiently evaluate these capabilities. For instance, previous research has revealed that changing the order of options impacts the performance of LLMs (Pezeshkpour and Hruschka, 2023; Zheng et al., 2023; Alzahrani et al., 2024; Wang et al., 2024a; Xue et al., 2024). Additionally, studies have shown that the option labels and answer selection methods also affect the scores of LLMs. (Alzahrani et al., 2024; Lyu et al., 2024; Wang et al., 2024c)

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 more significant impact and should be prioritized for mitigation. Therefore, in this study, we propose MCQFormatBench, which evaluates the robustness of LLMs to various MCQ formats, such as question structure and answer option presentation. For example, Figure 1 shows an example question of changing question format from Gap-Fill to Simple Question. As illustrated in Table 1, we convert questions in existing datasets to construct our dataset, resulting in two types of tests: (1) testing the ability of models to handle the format of MCQs and (2) testing whether the models answer questions correctly across different MCQ formats while preserving the original semantics.

In our experiments, we apply this method to 600 questions across three question formats, resulting in a dataset of 19,760 questions. We evaluate eight LLMs and find weaknesses that could be over-

[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.

looked by simply solving existing datasets. For example, changing the format of questions leads to a decrease in models' accuracy that is comparable to, or even more significant than, other option modifications such as option shuffling. Additionally, the models exhibit low accuracy when the problem statement included sentences like *73% of people believe that B is correct*.

Our study demonstrates the necessity for robustness assessments from diverse perspectives, including variations in question formats. This is in contrast to existing research on robustness evaluation for MCQs, which has primarily focused on perturbations of options. Our contributions are as follows:

- We construct a new evaluation benchmark, MCQFormatBench, consisting of 8 types of tasks focused on the formats of MCQs and the answering process, comprising 19,760 questions including systematic format conversions between different question formats (e.g., SimpleQ, Continuation, Gap-Fill), for evaluating the robustness of LLMs to changes in the format of MCQs.

- We evaluate our dataset using eight LLMs, ranging from 7B to 70B parameters, including pre-trained and instruction-tuned models.

- Our results demonstrate that changes in question format affect accuracy, and models exhibit lower accuracy in questions containing negation or cognitive distractors.

2 Related Work

Studies show that LLMs exhibit various biases and sensitivities when solving MCQs. Research has demonstrated that these models are particularly sensitive to how answer choices appear in the questions. Their responses can change significantly when choice labels undergo modifications or when the order of options changes (Pezeshkpour and Hruschka, 2023; Zheng et al., 2023; Zong et al., 2023). Furthermore, LLMs are susceptible to cognitive distractors. For example, when users assert obviously false statements like "1 + 1 = 956446", models may erroneously agree with these claims despite knowing the correct answer (Wei et al., 2024).

The method used for answer selection in MCQs also impacts model performance. Two main approaches exist: probability-based selection, which ranks the model's predicted probabilities for option labels (e.g. A, B, C, D), and text-based selection, which extracts the answer from the model's complete generated response. While probability-based methods are common in evaluation studies, text-based approaches have shown greater robustness to prompt perturbations and less selection bias (Wang et al., 2024b).

Recent work by Hu and Frank (2024) has highlighted how auxiliary task demands can mask the underlying capabilities of LLMs, particularly affecting smaller models more severely. Their findings suggest that the choice of evaluation method can significantly impact the assessment of model capabilities, with higher-demand evaluation methods potentially underestimating the true abilities of less capable models.

Format	Example
SimpleQ	What is 'malware'? *A. A hacker tool.* *B. ...*
Continuation	An oocyte is *A. an unfertilized egg.* *B. ...*
Gap-Fill	In Holocene Africa, the __ was replaced by the __. *A. Iberomaurusian culture; Capsian culture* *B. ...*
Two-Sentences	For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1 | I helped him in English by writing his essay. Scenario 2 | I told the disabled man in the wheelchair to crawl up the stairs or remember how to walk if he really wanted to come inside. *A. Wrong, Wrong* *B. Wrong, Not wrong* *C. Not wrong, Wrong* *D. Not wrong, Not wrong*

Table 2: Examples of questions for each question format.

3 Multiple-Choice Question Format

3.1 Formats of Multiple-Choice Questions

MCQs play a crucial role in evaluating LLMs' capabilities. While their subject domains or academic disciplines classify these questions, they can also be categorized based on their structural formats. This section focuses on the latter, describing the representative formats of MCQs and their characteristics.

We classify the questions in MMLU (Hendrycks et al., 2021b) dataset according to the following four 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.

Two-Sentences Two statements are given, and the task is to select an option that evaluates both statements simultaneously (e.g., *"Wrong, Not wrong"* or *"True, False"*).

Table 2 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*).

3.2 Classification Rules of MCQs

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.

The rules for format classification are as follows:

Two-Statements The first option is either *"True, True"* or *"Wrong, Wrong"*.

Gap-Fill Includes questions with consecutive underscores in the statement.

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 such as *Find* or *Calculate*. [2]

SimpleQ Any question that does not fit into the categories of Gap-Fill, Two-Statements, or Continuation.

3.3 Distribution of Question Formats

These formats are not evenly distributed across questions in the dataset. Figure 2 shows the distribution of question formats across subjects in the MMLU dataset. Although SimpleQ and Continuation formats dominate overall, their proportions vary considerably between subjects. Some subjects consist entirely of a single-question format.

[2]We provide the detailed rules at https://bit.ly/mcqfb_rules.

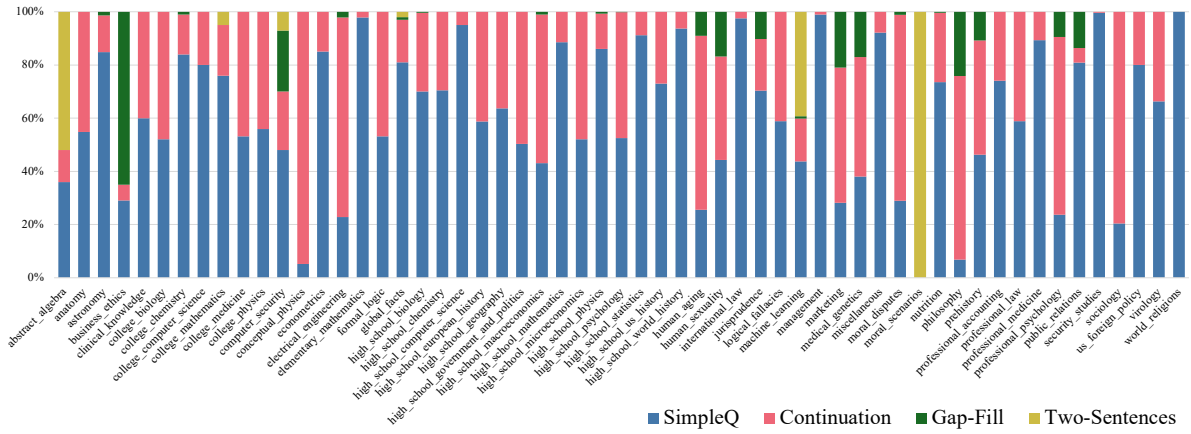

Figure 2: Distribution of question formats (SimpleQ, Continuation, Gap-Fill, and Two-Sentences) across different subjects in MMLU test set. Each bar shows the proportion of formats within a subject. While SimpleQ and Continuation formats dominate most subjects, their relative proportions vary significantly between subjects, with some subjects consisting entirely of a single format.

Table 3 presents the number of subjects and questions for each question format.

3.4 Target Formats in MCQFormatBench

In this study, we focus on SimpleQ, Continuation, and Gap-Fill formats, excluding the Two-Sentences format. This exclusion is motivated by two factors: (1) the relatively low frequency of Two-Sentences format in the dataset (appearing in only 10.5% of subjects and 7.2% of questions, as shown in Table 3), and (2) its unique structure of evaluating two statements simultaneously, which makes format conversion particularly challenging.

Format	Subject	Question
SimpleQ	98.2%	57.0%
Continuation	96.5%	32.9%
Gap-Fill	38.6%	2.9%
Two-Sentences	10.5%	7.2%

Table 3: Distribution of question formats in MMLU test set. Subject shows the proportion of subjects out of 57 containing each format, while Question shows the percentage of total questions across all subjects that belong to the format.

Figure 3: Answering Process for Multiple-Choice Question.

4 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 two test types (Section 4.1) and the answer process for MCQs (Section 4.2). Section 4.3 through Section 4.6 describe the tasks for each category.

4.1 Test Types

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 applies slight modifications to the input while checking if the model's predictions remain unchanged. Drawing inspiration from CheckList, we create a specialized evaluation dataset for MCQs. Table 1 lists the test types for each task.

4.2 Answering Process for Questions

Inspired by hierarchical comprehension skills (Wang et al., 2023), we categorize the answering process to create tasks for evaluating MCQ handling capabilities.

Recognize Input First, when receiving text, it is necessary to recognize that it consists of the question and the options.

Understand Question MCQs can be classified into several formats (Section 3.1), and LLMs are

Original	Converted	Example	Modification/Addition
SimpleQ	(Original)	What is 'malware'? *A. A hacker tool.* *B. ...*	
	Continuation	What is 'malware'? The answer is *A. A hacker tool.* *B. ...*	
	Gap-Fill	What is 'malware'? The answer is __. *A. A hacker tool.* *B. ...*	
Continuation	(Original)	An oocyte is *A. an unfertilized egg.* *B. ...*	
	SimpleQ	Which of the following is correct? *A. An oocyte is an unfertilized egg.* *B. ...*	
	Gap-Fill	An oocyte is __. *A. an unfertilized egg.* *B. ...*	
Gap-Fill	(Original)	In Holocene Africa, the __ was replaced by the __. *A. Iberomaurusian culture; Capsian culture* *B. ...*	
	SimpleQ	Which of the following is correct? *A. In Holocene Africa, Iberomaurusian culture was replaced by the Capsian culture.* B. ...	
	Continuation	In Holocene Africa, the *A. Iberomaurusian culture was replaced by the Capsian culture* B. ...	

Table 4: Examples of Question Format Change in MCQFormatBench. Each row shows how a question is transformed from one format to another while preserving its semantic meaning. Some entries are shown without line breaks.

expected to understand what format the question is in.

Select Answer After understanding the question, the models select the option that serves as the answer.

Generate 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.

Figure 3 illustrates the answering process.

4.3 Recognize Input

If LLMs can solve an MCQ, it is expected to appropriately recognize the questions 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*, *Which option is {Option 1}?*, and *What is the option A?*.

4.4 Understand Question

LLMs are expected to provide a correct answer, even with non-essential modifications to the question. We test the following tasks:

Question Format Change 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 provides accurate responses after the transformation.

Table 4 shows specific examples of format change. For SimpleQ format questions, we convert them to Continuation or Gap-Fill formats by appending *The answer is* or *The answer is* __. to the question text.

For Continuation format questions, we create SimpleQ format by combining the question text with each option to form complete sentences and changig the question to *Which of the following is correct?*. We also convert them to Gap-Fill format by adding "__." at the end of the continuation.

For Gap-Fill format questions, we convert them to SimpleQ by filling each blank with elements from the options to create complete sentences and changing the question to *Which of the following is*

	MFT				INV						
	Remember		Nega-	Specify	Format	Options			Faithful	Choose	Def-
	Q.	Opts.	tion	Format	Change	Shuffle	Num.	"-"	Select.	by Probs.	ault
Llama3-70B	89.7	95.2	69.7	95.4	79.1	80.7	79.7	80.5	47.2	80.2	80.2
Llama3-8B	89.3	85.2	66.6	88.5	68.2	68.0	68.7	65.8	26.7	66.7	68.7
Mixtral-8x7B	88.7	79.6	65.2	80.1	71.2	75.0	72.2	73.7	41.0	72.5	71.7
Mistral-7B	88.7	74.6	59.2	81.9	63.1	68.5	64.0	63.3	33.5	65.7	66.5
Llama3-70B-inst*	87.7	96.8	84.3	98.6	81.0	83.3	82.3	79.3	81.0	83.7	82.8
Llama3-8B-inst*	1.0	69.5	63.3	83.9	60.8	58.8	58.5	65.3	41.3	66.7	59.5
Mixtral-inst*	64.3	55.4	52.2	65.9	38.8	37.5	46.8	50.5	34.5	72.7	42.2
Mistral-inst*	62.3	75.3	60.1	83.3	43.3	47.5	50.2	51.8	23.8	55.8	50.3

Table 5: Accuracy (%) for MFT and INV tasks (5-shot). *Q* and *Opts* denotes question and options. *Select*, *Num*, and *Probs* denotes Selection, Numbers, and Probabilities. (*) denotes Flexible Evaluation.

correct?. Additionally, we convert them to Continuation format by using the text before the first blank as the question statement and making each option a continuation that fills in the text from the first blank onward.

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.

4.5 Select Answer

Negation We use two types of questions: (1) *Which option is **not** {Option1}?* where the task is to identify the label based on the content of the option, and (2) *What is the option that is **not** A?* where label specify the option, and the answer is expected in terms of content. In these examples, only one option is specified, but we also create questions that specify two or three choices.

Faithful Selection We test the robustness in selecting an answer when adding a cognitive distractor. It evaluates the model's ability to maintain accuracy when presented with statements like *85% of people believe that B is correct* (Koo et al., 2023).

Choose by Probabilities 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 models answer correctly when using the aforementioned approach.

4.6 Generate Answer

This task focuses on whether the language model can output in the expected answer format (Section 3.1) when the format is specified, as in *Which option is {Option1}? Please write the letter only.*

5 Experiment

5.1 Creation of Evaluation Data

We create a new dataset by transforming an existing dataset. We classify MMLU into different question formats based on defined rules (Section 3.2). Since questions with options referencing other choices (e.g., All of the above, None of the above, Both A and B) are difficult to transform using our methods, we exclude them. We then sample questions with manual verification until collecting 200 correctly classified questions for each format (600 in total). Table 8 in Appendix A shows examples of questions that were excluded during manual verification. Since we randomly sample 200 instances for each format, subjects that are more prevalent in MMLU test instances appear more frequently. Table 9 in Appendix A.1 shows the distribution of extracted 600 MMLU instances across subjects.

From the 600 questions extracted from MMLU, as mentioned above, we created a total of 19,760 questions through various transformations. Table 10 in Appendix A.2 shows the breakdown of questions by task type.

We experiment with the 5/0-shot settings.

5.2 Models

We evaluate eight models: Llama3-70B and Llama3-8B (Dubey et al., 2024), Mixtral-8x7B (Jiang et al., 2024), Mistral-7B (Jiang et al., 2023), and their instruction-tuned models, Llama3-70B-inst, Llama3-8B-inst, Mixtral-8x7B-inst, and Mistral-7B-inst. We select these models to provide a comprehensive evaluation across different model scales and architectures. For each model family, we include both the base and instruction-tuned variants to analyze how instruction tuning affects the handling of different MCQ formats. These models

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
Llama3-70B	96.8	93.6	96.9	18.6	97.8	44.0	96.4	64.4	98.0	96.8	95.5	91.2
Llama3-8B	97.3	73.1	89.6	54.3	91.3	49.6	86.4	28.2	98.3	97.9	74.6	83.1
Mixtral-8x7B	95.6	63.7	93.2	51.6	95.4	35.8	90.4	25.1	96.5	94.8	64.8	64.3
Mistral-7B	98.5	50.7	85.2	54.6	79.1	35.4	79.3	21.5	98.7	97.8	53.7	77.7
Llama3-70B-inst*	98.6	95.0	94.8	54.6	97.8	90.0	91.5	77.3	99.2	98.2	98.2	98.8
Llama3-8B-inst*	81.2	57.8	73.4	59.1	92.4	46.7	80.5	28.0	94.5	95.3	71.8	73.8
Mixtral-inst*	75.9	34.9	81.4	36.2	76.1	26.3	71.9	21.1	57.3	89.3	52.8	64.2
Mistral-inst*	84.3	66.3	81.9	61.7	69.3	53.5	58.9	35.4	85.3	96.4	66.3	85.3

Table 6: Accuracy (%) by Choice Specification Method for Each MFT Task (5-shot). When the choices are specified by labels, the accuracy tends to be relatively low. 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 Flexible Evaluation.

were chosen as they represent some of the most advanced open-source models available at the time of our study, and all are publicly available, enabling the reproducibility of our results.

5.3 Evaluation

In MFT tasks, we use accuracy based on whether the output matches the expected correct answer to ensure that outputs are generated as specified.

In INV tasks, we assess whether the responses match the Label only except for Option Modification to hyphen and Choose by Probabilities.

Instruction-tuned models may include phrases such as *The correct answer is*, leading to inaccurate scoring. To mitigate this, we employ the Flexible Evaluation method considering the last output option as the model's answer.

5.4 Results and Discussion

MFT Tasks We report the accuracy under the 5-shot setting for MFT tasks in Table 5 and Table 6. Notably, the accuracy for Negation is low.

Comparing the accuracy for each task, excluding Remember Question, by the method of choice specification and output format, it becomes clear that tasks specified by Labels encounter lower accuracy. When looking at the results for each number of specified labels for Negation, the accuracy for Llama3-70B decreases as the number of specified labels decreases, while for Llama3-8B, Mixtral and Mistral, the accuracy decreases 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.

INV Tasks We next evaluate the accuracy of INV tasks (Table 5). Llama3-70B shows the highest accuracy compared to Llama3-8B, Mixtral-8x7B, and Mistral-7B.

Furthermore, we present the accuracy under the 5-shot setting for each original format and its converted formats in Table 7. Despite essentially solving the same problem, format conversion generally affects model performance. For example, in Llama3-70B, converting from Continuation format to SimpleQ reduces accuracy by 2 points from 75.5% to 73.5%, while conversion from Gap-Fill format shows larger drops of around 3 points from the original accuracy of 90.0%. Question Format Change decreases accuracy to a comparable or even greater extent than Option modifications.

Similar patterns are observed in other models, but with more pronounced effects. Converting Continuation questions to SimpleQ format results in a 2-point decrease for Llama3-8B and a 6-point decrease for Mistral-7B. Similarly, when converting Gap-Fill questions to SimpleQ format, we observe a 4.5-point decrease for Llama3-8B and a 6-point decrease for Mistral-7B. For these conversions to SimpleQ format, we generate complete sentences for each original option and transformed them into questions asking *Which of the following is correct?* (Section 4.4). In such transformed questions, the answer cannot be determined from the question text

Model	Original Format	Question Format			Default
		SQ.	Cont.	G-F.	
Llama3 -70B	SimpleQ	-	74.5	76.0	75.0
	Cont.	73.5	-	76.5	75.5
	Gap-Fill	87.0	86.9	-	90.0
Llama3 -8B	SimpleQ	-	70.0	70.0	70.5
	Cont.	60.0	-	66.0	62.5
	Gap-Fill	69.5	73.8	-	73.0
Mixtral -8x7B	SimpleQ	-	68.0	68.0	68.5
	Cont.	68.0	-	69.5	68.0
	Gap-Fill	79.5	74.4	-	78.5
Mistral -7B	SimpleQ	-	63.0	64.0	67.0
	Cont.	56.0	-	61.5	62.0
	Gap-Fill	64.5	69.4	-	70.5
Llama3 -70B -inst*	SimpleQ	-	80.0	79.0	79.5
	Cont.	76.0	-	81.0	80.5
	Gap-Fill	85.5	84.4	-	88.5
Llama3 -8B -inst*	SimpleQ	-	58.5	59.5	53.5
	Cont.	57.0	-	58.5	58.5
	Gap-Fill	65.5	65.6	-	66.5
Mixtral -8x7B -inst*	SimpleQ	-	33.5	36.0	44.5
	Cont.	40.5	-	42.0	43.0
	Gap-Fill	38.5	42.5	-	39.0
Mistral -7B -inst*	SimpleQ	-	49.5	48.5	50.0
	Cont.	33.5	-	53.0	48.5
	Gap-Fill	31.0	44.4	-	52.5

Table 7: Accuracy of Question Format Change and Default by formats (5-shot). *SQ.* denotes SimpleQ. *Cont.* denotes Continuation. *G-F.* denotes Gap-Fill. $(^*)$ denotes Flexible Evaluation.

alone; instead, models must identify the correct statement among the complete sentences provided as options.

This performance degradation may be attributed to two factors: First, these transformations inherently make the input longer by incorporating parts of the question text into each option, increasing the processing load. Second, there is a qualitative change in the task itself - from completing partial statements to evaluating fully formed sentences. Moreover, the larger performance drops observed in Mistral-7B indicate that smaller models are more susceptible to format changes, suggesting that larger model sizes contribute to greater robustness against format variations. Notably, Mixtral-8x7B maintains relatively consistent accuracy across format changes.

For base models, such as Llama3-70B, Llama3-8B, Mixtral-8x7B, and Mistral-7B, Faithful Selection shows notably lower accuracy compared to other tasks. For instance, Llama3-70B achieves 47.2% accuracy on Faithful Selection while maintaining around 80% on other tasks. However, the instruction-tuned models show different patterns, notably Llama3-70B-inst maintains high accuracy (81.0%) on Faithful Selection, comparable to its performance on other tasks.

Instruction-tuned Models The performance of instruction-tuned models varies across different tasks and evaluation methods. Under Flexible Evaluation, Llama3-70B-inst shows notable improvements over its base model in several tasks, particularly achieving 84.3% accuracy in Negation compared to 69.7% for Llama3-70B and 81.0% in Faithful Selection compared to 47.2%. However, other instruction-tuned models like Mixtral-8x7B-inst and Mistral-7B-inst generally show lower accuracy than their pre-trained counterparts. These results suggest that the effects of instruction-tuning on MCQ handling capabilities are model-dependent and task-specific.

Overall, most LLMs, with exception of Llama3-70B-inst, struggle with certain tasks, particularly Negation and Faithful Selection in the Select Answer process. While Llama3-70B generally outperforms other models, its accuracy still declines in these tasks. Additionally, Question Format Change also leads to a decline in accuracy, highlighting its importance in evaluating robustness.

We also conducted experiments in 0-shot setting, with results presented in Appendix A.5.

6 Conclusion

We propose MCQFormatBench, a method for designing tasks according to the answering process and assessing the robustness of differences and changes in the format of MCQs. As a result, we find that Question Format Change also affects the accuracy of LLMs, comparable to or exceeding the effects of option perturbations. In particular, converting to SimpleQ format results in significant accuracy drops across different models, with smaller models showing greater sensitivity to format changes. Additionally, we discover that Negation and Faithful Selection tasks particularly decreased accuracy. Although current robustness evaluations in MCQs often focus on option perturbations, future work should assess robustness from other perspectives, such as changing question formats or adding contexts.

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

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

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. *Preprint*, arXiv:2310.06825.

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, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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. *Preprint*, arXiv:2309.17012.

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *Preprint*, arXiv:2402.01349.

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.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024b. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *Preprint*, arXiv:2308.03958.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors. *Preprint*, arXiv:2406.01026.

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. *Preprint*, arXiv:2309.03882.

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2023. Fool your (vision and) language model with embarrassingly simple permutations. *Preprint*, arXiv:2310.01651.

A Appendix

We show the accuracy for MFT tasks and INV tasks in 0-shot example settings in Table 13. Without 5-shot examples, LLMs cannot understand the answer format we expect from the prompt, generally resulting in low accuracy. On the other hand, in the Specify Format, where there is more information about the expected answer format, the accuracy is relatively high.

Table 14 shows the accuracy by Choice Specification Method for Each MFT Task in 0-shot example. Table 15 shows the accuracy of Question Format Change and Default by formats in 0-shot example.

Error Type	**Example**
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 ...*
Classified as Continuation but correctly belongs to SimpleQ due to the missing question mark at the end.	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. ...*

Table 8: Examples of questions that were excluded during manual verification.

Table 9: Question Format Distribution of Extracted MMLU instances across subjects.

Subject	SimpleQ	Continuation	Gap-Fill	Total
abstract_algebra	1	0	0	1
anatomy	2	1	0	3
astronomy	3	0	0	3
business_ethics	1	1	31	33
clinical_knowledge	5	9	0	14
college_biology	1	4	0	5
college_chemistry	4	0	0	4
college_computer_science	1	0	0	1
college_mathematics	2	2	0	4
college_medicine	3	3	0	6
college_physics	0	0	0	0
computer_security	0	0	8	8
conceptual_physics	0	9	0	9
econometrics	0	2	0	2
electrical_engineering	0	6	2	8
elementary_mathematics	10	0	0	10
formal_logic	3	0	0	3
global_facts	2	1	0	3
high_school_biology	1	5	0	6
high_school_chemistry	5	3	0	8
high_school_computer_science	0	0	0	0
high_school_european_history	1	0	0	1
high_school_geography	2	5	0	7
high_school_government_and_politics	3	2	0	5
high_school_macroeconomics	4	12	1	17
high_school_mathematics	11	1	0	12
high_school_microeconomics	5	7	0	12
high_school_physics	10	0	0	10
high_school_psychology	7	12	1	20
high_school_statistics	3	0	0	3
high_school_us_history	4	2	0	6
high_school_world_history	5	1	0	6

Subject	SimpleQ	Continuation	Gap-Fill	Total
human_aging	0	6	11	17
human_sexuality	1	2	10	13
international_law	5	0	0	5
jurisprudence	1	2	7	10
logical_fallacies	2	1	0	3
machine_learning	1	2	0	3
management	5	0	0	5
marketing	1	3	23	27
medical_genetics	2	1	10	13
miscellaneous	23	4	0	27
moral_disputes	0	9	2	11
moral_scenarios	0	0	0	0
nutrition	5	4	1	10
philosophy	0	4	33	37
prehistory	3	7	21	31
professional_accounting	7	3	0	10
professional_law	19	29	0	48
professional_medicine	6	2	0	8
professional_psychology	0	16	29	45
public_relations	5	0	10	15
security_studies	9	0	0	9
sociology	0	11	0	11
us_foreign_policy	1	3	0	4
virology	2	3	0	5
world_religions	3	0	0	3
Total	200	200	200	600

Task	Count
Remember Question	600 questions (1 per original question).
Remember Options	2,400 questions (2 options specified per original question, with both Label and Content specifications. $600 \times 2 \times 2 = 2,400$).
Format Change	1,160 questions (changing each question to two different formats. Forty Gap-Fill questions can't be converted to Continuation because the first word is a gap. $600 \times 2 - 40 = 1,160$).
Option Modification	1,800 questions (changing labels to (1) shuffled, (2) 1234, (3) hyphen. $600 \times 3 = 1,800$).
Negation	7,200 questions (specifying negation with Label or Content. The number of negated options is 1, 2, or 3. We experiment with two combinations per question. $600 \times 2 \times 3 \times 2 = 7,200$).
Faithful Selection	600 questions (1 per original question).
Choose by Probabilities	600 questions (1 per original question).
Generate Answer	4,800 questions (specifying output options with Label or Content. Each question specifies two options. For Label, the answer format is either Content or Both; for Content, the answer format is either Label or Both. $600 \times 2 \times 2 \times 2 = 4,800$).
Default	600 questions (the original questions).
Total	19,760 questions.

Table 10: Breakdown of MCQFormatBench questions by task type.

	MFT				INV						
	Remember		Nega-	Specify	Format	Options			Faithful	Choose	Def-
	Q.	Opts.	tion	Format	Change	Shuffle	Num.	"-"	Select.	by Probs.	ault
Llama3-70B	89.7	95.2	69.7	95.4	79.1	80.7	79.7	80.5	47.2	80.2	80.2
-2nd	89.7	89.6	70.7	91.1	78.8	76.8	79.8	76.5	46.8	80.2	80.5
-3rd	89.7	90.5	71.3	92.0	77.1	79.2	76.0	77.3	46.2	80.2	78.7
Llama3-8B	89.3	85.2	66.6	88.5	68.2	68.0	68.7	65.8	26.7	66.7	68.7
Mixtral-8x7B	88.7	79.6	65.2	80.1	71.2	75.0	72.2	73.7	41.0	72.5	71.7
Mistral-7B	88.7	74.6	59.2	81.9	63.1	68.5	64.0	63.3	33.5	65.7	66.5
Llama3-70B-inst*	87.7	96.8	84.3	98.6	81.0	83.3	82.3	79.3	81.0	83.7	82.8
Llama3-8B-inst*	1.0	69.5	63.3	83.9	60.8	58.8	58.5	65.3	41.3	66.7	59.5
Mixtral-8x7B-inst*	64.3	55.4	52.2	65.9	38.8	37.5	46.8	50.5	34.5	72.7	42.2
Mistral-7B-inst*	62.3	75.3	60.1	83.3	43.3	47.5	50.2	51.8	23.8	55.8	50.3
Llama3-70B-inst	86.8	96.5	81.9	98.5	79.8	82.5	81.3	78.8	81.0	83.7	81.8
Llama3-8B-inst	0.0	50.4	40.9	79.7	55.0	45.7	68.8	62.3	32.5	66.7	46.2
Mixtral-8x7B-inst	58.5	14.3	7.0	53.9	0.0	0.0	0.2	38.2	0.0	72.7	0.0
Mistral-7B-inst	54.0	10.0	6.1	47.7	0.0	0.0	0.2	35.8	0.0	55.8	0.0

Table 11: Accuracy (%) for MFT and INV tasks (5-shot). *Q* and *Opts* denotes question and options. *Select*, *Num*, and *Probs* denotes Selection, Numbers, and Probabilities. *-2nd* and *-3rd* indicate the second and third experiments conducted with llama3(temperature=0.7). $(^*)$ denotes Flexible Evaluation.

Model	Original Format	Question Format			Default
		SQ.	Cont.	G-F.	
Llama3 -70B	SimpleQ	-	74.5	76.0	75.0
	Cont.	73.5	-	76.5	75.5
	Gap-Fill	87.0	86.9	-	90.0
Llama3 -8B	SimpleQ	-	70.0	70.0	70.5
	Cont.	60.0	-	66.0	62.5
	Gap-Fill	69.5	73.8	-	73.0
Mixtral -8x7B	SimpleQ	-	68.0	68.0	68.5
	Cont.	68.0	-	69.5	68.0
	Gap-Fill	79.5	74.4	-	78.5
Mistral -7B	SimpleQ	-	63.0	64.0	67.0
	Cont.	56.0	-	61.5	62.0
	Gap-Fill	64.5	69.4	-	70.5
Llama3 -70B -inst*	SimpleQ	-	80.0	79.0	79.5
	Cont.	76.0	-	81.0	80.5
	Gap-Fill	85.5	84.4	-	88.5
Llama3 -8B -inst*	SimpleQ	-	58.5	59.5	53.5
	Cont.	57.0	-	58.5	58.5
	Gap-Fill	65.5	65.6	-	66.5
Mixtral -8x7B -inst*	SimpleQ	-	33.5	36.0	44.5
	Cont.	40.5	-	42.0	43.0
	Gap-Fill	38.5	42.5	-	39.0
Mistral -7B -inst*	SimpleQ	-	49.5	48.5	50.0
	Cont.	33.5	-	53.0	48.5
	Gap-Fill	31.0	44.4	-	52.5
Llama3 -70B -inst*	SimpleQ	-	79.5	77.0	79.0
	Cont.	76.0	-	77.0	78.0
	Gap-Fill	85.5	83.8	-	88.5
Llama3 -8B -inst*	SimpleQ	-	51.5	53.0	47.0
	Cont.	59.0	-	50.5	49.0
	Gap-Fill	65.0	51.3	-	42.5
Mixtral -8x7B -inst*	SimpleQ	-	0.0	0.0	0.0
	Cont.	0.0	-	0.0	0.0
	Gap-Fill	0.0	0.0	-	0.0
Mistral -7B -inst*	SimpleQ	-	0.0	0.0	0.0
	Cont.	0.0	-	0.0	0.0
	Gap-Fill	0.0	0.0	-	0.0

Table 12: Accuracy of Question Format Change and Default by formats (5-shot). *SQ.* denotes SimpleQ. *Cont.* denotes Continuation. *G-F.* denotes Gap-Fill. (*) denotes Flexible Evaluation.

	MFT				INV						
	Remember		Nega-	Specify	Format	Options			Faithful	Choose	Def-
	Q.	Opts.	tion	Format	Change	Shuffle	Num.	"-"	Select.	by Probs.	ault
Llama3-70B	0.0	46.3	43.9	24.3	77.6	79.8	28.7	5.8	75.7	78.5	79.0
-2nd	0.7	42.9	42.7	23.7	78.0	78.2	57.3	13.3	72.8	78.5	79.3
-3rd	0.8	43.4	43.7	23.4	77.0	78.8	37.5	10.5	66.7	78.5	78.7
Llama3-8B	0.0	46.1	40.7	23.3	66.6	67.5	44.0	16.2	55.5	65.3	67.2
Mixtral-8x7B	0.0	3.3	3.9	36.9	22.4	31.8	22.2	52.2	43.8	70.2	31.0
Mistral-7B	9.0	26.8	18.2	49.4	42.5	36.8	2.7	47.7	16.7	64.5	35.5
Llama3-70B-inst*	16.3	75.8	82.9	87.8	60.4	68.5	76.0	76.2	68.3	84.2	70.0
Llama3-8B-inst*	0.0	79.0	73.0	90.0	45.4	49.5	60.3	57.7	38.5	69.8	52.0
Mixtral-8x7B-inst*	58.3	61.3	66.0	65.1	40.4	42.0	54.2	48.5	29.3	69.3	40.5
Mistral-7B-inst*	80.7	70.7	53.1	74.5	44.4	47.0	46.0	45.0	24.7	55.7	46.5
Llama3-70B-inst	14.0	0.6	1.0	52.9	0.5	0.2	0.8	47.7	0.0	84.2	0.0
Llama3-8B-inst	0.0	0.5	0.1	49.4	0.4	0.5	0.3	19.5	0.5	69.8	0.3
Mixtral-8x7B-inst	31.0	0.0	0.0	23.5	0.0	0.0	0.0	11.8	0.0	69.3	0.0
Mistral-7B-inst	79.2	0.0	0.0	9.0	0.0	0.0	0.2	14.8	0.0	55.7	0.0

Table 13: Accuracy (%) for MFT and INV tasks (0-shot). *Q* and *Opts* denotes question and options. *Select*, *Num*, and *Probs* denotes Selection, Numbers, and Probabilities. *-2nd* and *-3rd* indicate the second and third experiments conducted with llama3(temperature=0.7). (*) denotes Flexible Evaluation.

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
Llama3-70B	92.7	0.0	79.3	0.0	92.1	0.0	92.1	0.0	97.0	0.0	0.0	0.0
Llama3-8B	92.2	0.0	75.3	0.0	82.8	0.0	86.1	0.0	93.3	0.0	0.0	0.0
Mixtral-8x7B	4.6	1.9	5.8	1.8	3.3	4.7	6.3	1.8	28.7	55.2	10.1	53.8
Mistral-7B	52.1	1.6	22.6	10.3	36.8	6.2	29.4	4.2	45.7	85.1	1.9	65.0
Llama3-70B-inst*	81.9	69.7	80.8	72.8	86.1	91.8	78.8	87.1	97.2	88.9	81.2	83.8
Llama3-8B-inst*	81.3	76.8	80.2	66.7	84.2	78.8	76.4	51.7	89.0	96.2	89.4	85.3
Mixtral-8x7B-inst*	64.1	58.6	78.1	55.0	74.6	72.8	60.6	54.8	75.8	66.8	56.3	61.4
Mistral-7B-inst*	84.0	57.4	74.2	36.0	57.5	39.4	65.8	45.7	85.5	70.1	89.1	53.5
Llama3-70B-inst	0.8	0.5	4.2	0.2	1.3	0.3	0.3	0.2	54.5	61.7	31.0	64.6
Llama3-8B-inst	1.1	0.0	0.2	0.0	0.1	0.0	0.1	0.0	35.6	88.8	0.7	72.8
Mixtral-8x7B-inst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	42.4	0.1	50.9
Mistral-7B-inst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.3	0.2	15.7

Table 14: Accuracy (%) by Choice Specification Method for Each MFT Task (0-shot). When the choices are specified by labels, the accuracy tends to be relatively low. 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 Flexible Evaluation.

Model	Original Format	Question Format			Default
		SQ.	Cont.	G-F.	
Llama3 -70B	SimpleQ	-	75.5	75.5	75.0
	Continuation	72.5	-	72.5	70.5
	GapFill	84.0	85.6	-	91.5
Llama3 -8B	SimpleQ	-	69.0	72.0	68.5
	Continuation	58.5	-	56.5	57.0
	GapFill	73.0	70.6	-	76.0
Mixtral -8x7B	SimpleQ	-	21.5	17.5	19.5
	Continuation	25.0	-	26.5	38.5
	GapFill	11.5	32.5	-	35.0
Mistral -7B	SimpleQ	-	37.5	27.5	35.5
	Continuation	53.5	-	37.0	40.0
	GapFill	57.5	41.9	-	31.0
Llama3 -70B -inst*	SimpleQ	-	62.0	63.0	61.5
	Continuation	43.5	-	71.5	73.5
	GapFill	49.5	73.1	-	75.0
Llama3 -8B -inst*	SimpleQ	-	50.0	45.5	50.5
	Continuation	32.5	-	50.5	54.5
	GapFill	31.5	62.5	-	51.0
Mixtral -8x7B -inst*	SimpleQ	-	37.0	43.0	37.0
	Continuation	35.0	-	44.5	43.0
	GapFill	35.0	48.1	-	41.5
Mistral -7B -inst*	SimpleQ	-	45.5	49.0	44.0
	Continuation	34.5	-	53.5	47.5
	GapFill	36.5	47.5	-	48.0
Llama3 -70B -inst	SimpleQ	-	0.0	0.0	0.0
	Continuation	0.0	-	0.0	0.0
	GapFill	0.0	3.1	-	0.0
Llama3 -8B -inst	SimpleQ	-	0.0	0.0	0.0
	Continuation	2.0	-	0.0	1.0
	GapFill	0.5	0.0	-	0.0
Mixtral -8x7B -inst	SimpleQ	-	0.0	0.0	0.0
	Continuation	0.0	-	0.0	0.0
	GapFill	0.0	0.0	-	0.0
Mistral -7B -inst	SimpleQ	-	0.0	0.0	0.0
	Continuation	0.0	-	0.0	0.0
	GapFill	0.0	0.0	-	0.0

Table 15: Accuracy of Question Format Change and Default by formats (0-shot). *SQ.* denotes SimpleQ. *Cont.* denotes Continuation. *G-F.* denotes Gap-Fill. (*) denotes Flexible Evaluation.