

Is Large Language Model Performance on Reasoning Tasks Impacted by Different Ways Questions Are Asked?

Seok Hwan Song, Mohna Chakraborty, Qi Li, Wallapak Tavanapong

Department of Computer Science, Iowa State University, Ames, Iowa, USA

{song92, mohnac, qli, tavanapo}@iastate.edu

Abstract

Large Language Models (LLMs) have been evaluated using diverse question types, e.g., multiple-choice, true/false, and short/long answers. This study answers an unexplored question about *the impact of different question types on LLM accuracy on reasoning tasks*. We investigate the performance of five LLMs on three different types of questions using quantitative and deductive reasoning tasks. The performance metrics include accuracy in the reasoning steps and choosing the final answer. **Key Findings:** (1) Significant differences exist in LLM performance across different question types. (2) Reasoning accuracy does not necessarily correlate with the final selection accuracy. (3) The number of options and the choice of words, influence LLM performance.

1 Introduction

Performance of Large Language Models (LLMs) on reasoning tasks has been extensively studied using diverse benchmarks (Weston et al., 2016; Cobbe et al., 2021; Huang and Chang, 2023). Several benchmarks use short answer questions (SAQs) where the LLMs generate a response, often accompanied by an explanation. Multiple-choice questions (MCQs) are widely used because they allow for simple assessment (Li et al., 2024). LLMs select the correct option among the given options. For True or False questions (TFQs), LLMs judge if the given statement/question is true or false. Factors that may influence LLM performance include the number of options, the sensitivity of word choices in MCQs, and whether “True” or “False” is the correct answer in TFQs.

Assessing LLM performance can be challenging due to the diverse question types. SAQs have the advantage of analyzing reasoning steps with clear answers. For MCQs, LLMs need to understand the options correctly and choose the right option.

LLMs may overlook the reasoning steps. TFQs require LLMs to understand true or false accurately.

Motivated by these challenges, we evaluate LLM performance on reasoning questions across different question types on two reasoning tasks, *quantitative reasoning task and a deductive reasoning task*. Quantitative reasoning datasets were evaluated to benchmark multiple existing LLMs (OpenAI, 2023; Dubey et al., 2024; Team et al., 2024), and it contains arithmetic calculation. We also chose deductive reasoning tasks to evaluate LLMs’ logical reasoning without arithmetic calculations.

Our study focuses on assessing the suitability of these question types for benchmarking purposes and gaining insights into the behavior of LLMs. We propose *final selection accuracy* and *reasoning accuracy* for performance assessment. Final selection accuracy evaluates only the final answer in the output, such as the selected option in MCQs or TFQs. Reasoning accuracy assesses the correctness of the reasoning steps leading to the final answer. Checking reasoning accuracy is time-consuming since it cannot be automated. Consequently, we introduce the following research questions:

1. (RQ1) Do question types (SAQs, MCQs, and TFQs) impact the final selection accuracy of LLM?
2. (RQ2) Do question types impact the reasoning accuracy of LLM?
3. (RQ3) What are the patterns of incorrect outputs by LLMs on different question types?
4. (RQ4) What factors of MCQs impact LLM performance?
5. (RQ5) What factors of TFQs influence LLM performance?

Contribution #1: The aforementioned research questions and the experimental design to answer

the questions. We carefully designed the options for MCQs and TFQs. To our best knowledge, RQ1, RQ2, RQ3, and RQ5 have not been investigated. RQ4 was partially investigated in (Zheng et al., 2023) but with the following differences: (1) different numbers of options, (2) the wrong options in our MCQs covering a wide range of mistakes, and (3) the inclusion of the “Something else” option that enlarges the solution space in reasoning in our MCQs.

Contribution #2: New key findings from evaluations of five LLM models, two closed-source LLMs, and three open-source LLMs on quantitative reasoning problems selected from the GSM8K dataset (Cobbe et al., 2021) and deductive reasoning problems selected from the bAbI dataset (Weston et al., 2016). We will share our code and benchmark publicly¹. Some key findings are as follows. (1) There are more statistically significant differences in the final accuracy across different question types compared to the reasoning accuracy. (2) Reasoning accuracy does not necessarily correlate with final selection accuracy. (3) LLMs perform better with “True” as correct answers than “False” in TFQs. For MCQs, the LLM performances vary significantly based on the number of questions. The insight is potentially useful to guide future benchmark development and improve LLM performance.

2 Related Work

Diverse benchmarks (Cobbe et al., 2021; Patel et al., 2021) were used to evaluate LLMs (OpenAI, 2023; Dubey et al., 2024; Team et al., 2024) and prompting methods (Song and Tavanapong, 2024). Several datasets (Hendrycks et al., 2021; Bhakthavatsalam et al., 2021; Sakaguchi et al., 2021) include MCQs (Zheng et al., 2023). Moreover, factors influencing performance on MCQs have been studied, for instance, selection biases by order of options (Zheng et al., 2023; Li et al., 2024; Wang et al., 2024), and types of token options (Zheng et al., 2023). Balepur et al. (2024) found that LLMs can solve MCQs without knowing the question after some few-shot examples. Li et al. (2024) claims that the LLM performance of MCQs is better than those of long-form generation questions on the CARE-MI dataset. They do not find a significant difference in the different number of options

¹Code and benchmark: <https://github.com/NRT-D4/LLM-bias-questiontype>

for MCQs. In their study, the number of options varies from 2 to 4 without “Something else” as a possible answer. Our MCQs contain 5 options and 11 options with “Something else” as the final answer.

3 Benchmark Construction

Our goals for designing the benchmark include (1) ensuring that multiple-choice and True-or-False questions are as difficult as short-answer questions in terms of reasoning; (2) evaluating LLM accuracy in both choosing the final answers and its reasoning steps; and (3) answering the three research questions while avoiding the bias issues mentioned in the related work.

3.1 Reasoning Tasks and Source Datasets

- **Quantitative reasoning with GSM8K (Cobbe et al., 2021):** GSM8K is one of the standard datasets used for evaluating LLM quantitative reasoning performance. The dataset has grade school-level math word problems in short-answer question types. For each problem, the answer and steps in deriving the answer are included.
- **Deductive reasoning with bAbI-Factoid QA with three supporting facts:** This dataset is a subset of the bAbI dataset (Weston et al., 2016). Each problem has three supporting facts in the short-answer question type.

We randomly selected 300 reasoning problems from each dataset, considering the API budget for the closed-source LLMs, and the time-consuming manual annotation to evaluate LLM reasoning performance. To add reasoning complexity to deductive reasoning tasks, for each selected problem, we added two additional facts irrelevant to the question of the problem. These facts were randomly selected from the other questions in the bAbI dataset. We denoted the dataset with the 300 selected GSM8K problems as GSM8K300 and the modified bAbI dataset as bAbI300. See examples in A.3.

3.2 Prompt Designs to Format Question Types

Given each reasoning problem, different prompt structures are used to format the problem into corresponding SAQ, MCQ, and TFQ question types. Figure 1 shows our example prompts.

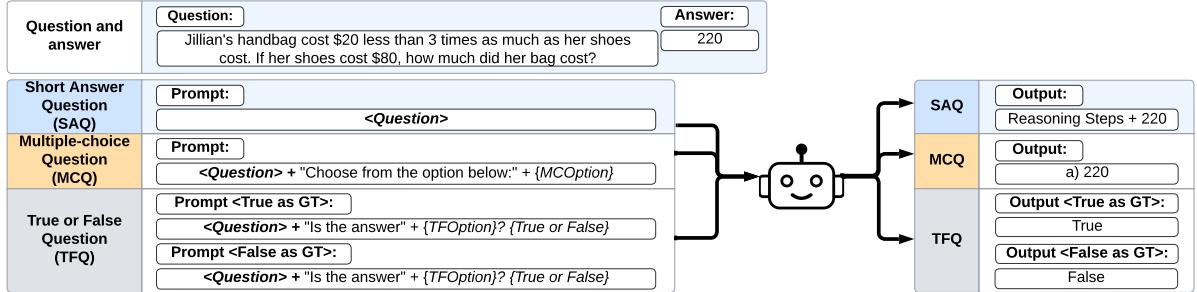


Figure 1: Examples of different types of questions generated from the original problem. Variables in italics are replaced by actual values; “+” indicates concatenation. *MCOption* represents multiple choices options; *TFOpt* is either a correct or incorrect answer for TFQs.

3.2.1 Short Answer Questions (SAQs)

The prompt structure is simple. It only has the question without any other instructions to LLMs. They return their reasoning steps and the final answer.

3.2.2 Multiple Choice Questions (MCQs)

This type of question contains a question, a brief instruction, and a list of options. Each option contains a token (e.g., “a”, “b”) followed by a value. We investigate several design factors but fix the token format since previous work has shown that the token formats (i.e., a/b/c/d, 1/2/3/4, and (A)/(B)/(C)/(D)) do not result in statistically significant differences in LLM performance (Zheng et al., 2023).

Number of options: Previous work found that LLM performances on 2, 3, and 4 option MCQs are not statistically significant (Li et al., 2024). We created 5-option and 11-option MCQs. The 11-option MCQs are rare but included in the Numersense dataset (Lin et al., 2020). We aim to test whether the significant difference in the number of options leads to observed notable performance variations. If LLMs perform actual calculations through reasoning steps, the number of options should not matter. However, if LLMs guess or test each option step-by-step, the chance of guessing correctly with eleven options is significantly less.

“Something else” or (SEO) as the last option value: Our rationale is to ensure that LLMs can select an option for their reasoning on our multiple-choice questions when SEO is included as an option, just like in short-answer questions.

Positions of correct answers: A multiple-choice question has only one correct option among all positions, including the final option “Something else.” LLMs were found to have *selection bias* defined as the bias to select specific options as an-

swers (Zheng et al., 2023). We investigate three approaches for placing the correct answers.

1. **Uniform across all positions (U):** Across all the problems, each option, including SEO, has an equal chance of being the correct answer. For instance, for the dataset with 300 reasoning problems, with five-option multiple-choice questions, 60 problems have the first option as the correct answer; 60 other problems have the correct answers in the second option, and so on.
2. **Uniform across all non-SEO positions (U-SEO):** Each option, except SEO, has an equal chance of being the correct answer. For the same example, the first option is the correct answer for 75 problems. The second option is the correct answer for 75 other problems, and so on. SEO positioned at the end has an incorrect answer.
3. **Only SEO (oSEO):** The correct answer is SEO for all the reasoning problems.

Option values for wrong answers: For the quantitative reasoning task, we designed three types of wrong options for each question. In Type 1, the wrong answer is 5% different from the ground truth, whereas in Type 2, the wrong option is 20% different. For Type 3, we used the numeric values from the question to fill in a generated equation with random operators and check that the result of the equation is not the same as the ground truth before using the result as the wrong option. When there were too few numeric values in a reasoning problem to generate different wrong answers, we manually created wrong answers. A 5-option MCQ has three wrong options, one for each type,

whereas an 11-option MCQ has three wrong options for each type. For the deductive reasoning task, the wrong option values were selected from a predefined list. For a reasoning problem that SEO is the correct answer, another wrong option value is needed. For quantitative reasoning, we randomly generated the wrong option value that has the same number of digits as the ground truth. For deductive reasoning, the wrong value was randomly selected from the predefined list.

3.3 True or False Questions (TFQs)

This question type has either a correct or incorrect answer in the question. The output is either true or false. We designed two formats of TFQ questions, as shown below, to test whether LLMs are sensitive to how the questions are asked. We replaced `{answer}` and `{instruction}` variables with values for each reasoning problem when prompting LLMs.

- **Question format:**

“Is the answer `{answer}`? + `{instruction}`”

- **Statement format:**

“The answer of the question is `{answer}`. + `{instruction}`”

The `{answer}` variable was substituted with the correct or incorrect answer. For GSM8K300, the incorrect answer was randomly picked from numbers with the same number of digits of the correct answer. For bAbI300, the incorrect answer was randomly selected from the list of options without the correct answer. We considered two different values for the `{instruction}` variable: (1) “True or False” and (2) “Solve the question first and choose True or False.” We found that LLMs tend to select the true or false before solving the question. Therefore, we added the second instruction. We also studied the impact of using “Yes or No” instead of “True or False” on LLM performance.

3.4 Evaluation Metrics

We evaluate LLM performance using two metrics: final selection accuracy and reasoning accuracy.

Final selection (FS) accuracy is the ratio of the number of reasoning problems an LLM gives the correct answers to the total number of reasoning problems. The correct answer for each SAQ problem is the ground truth. For MCQs, the correct answer includes the correct token with the exception that if the correct answer is “Something else,”

the token is not required. For TFQs, the final selection answer should simply be either “True” or “False.” Calculation of FS accuracy was automated using string matching between the outputs and the ground truths.

Reasoning accuracy is the ratio of the number of reasoning problems an LLM gives correct reasoning to the total number of reasoning problems. We used a strict criterion that the LLM needs to output all the reasoning steps correctly to be considered as giving the correct reasoning. This process cannot be automated. **Manual Evaluation:** The first author trained three students. Each person evaluated distinct subsets of the complete LLM outputs group by the LLM model to keep the evaluations within the same model consistent. All evaluators communicated frequently to maintain evaluation consistency. The main author randomly verified the evaluation results. For SAQs, the reasoning accuracy is the same as the FS accuracy.

3.5 Models and Experiments

We chose OpenAI GPT models (gpt-4o and gpt-3.5-turbo) to represent large closed-source LLMs. We selected Llama3 8B (llama-8B-instruct), Llama3.2 1B (llama-1B-instruct), and Gemma 7B (gemma-7B-instruct) to represent small open-source LLMs for the repeatability of experiments. The default configurations used greedy decoding (i.e., temperature = 0). For each reasoning problem and each configuration, an API call was requested to that LLM once. The reasoning and final outputs were saved. We did not apply self-consistency since we needed to analyze reasoning accuracy. The experiments were conducted from late October to early December of 2024.

3.6 Statistical Tests and Notations

In all performance comparisons for RQ1, RQ2, RQ4, and RQ5, we used the two-sided paired Wilcoxon signed-rank test (Wilcoxon, 1992) to check for statistical significance in LLM performance differences between two different question types. The baseline can be different for different performance comparisons. The highlighted cells (both cyan and pink) in all the tables indicate statistically significant differences from the baseline (p -values < 0.05). Cyan and pink colors highlight that the baseline is better or worse than the compared method, respectively. Different positions of the correct answers of MCQs (U, U-SEQ, and oSEQ) are described in Section 3.2.2. In Tables 2-6,

Final Selection (FS)	Models	GSM8K300						bAbI300					
		SAQ	MCQ (5 Options)			TFQ (Question)		SAQ	MCQ (5 Options)			TFQ (Question)	
			U	U-SEO	oSEO	True	False		U	U-SEO	oSEO	True	False
	GPT-4o	92.00	89.45	91.33	97.67	93.00	97.33*	82.33	79.00	80.00	83.67	67.67	97.00 [†]
	GPT-3.5	79.60	69.67	83.67	16.67	75.67	96.67*	52.67	30.00	38.33	10.67	16.33	92.67 [†]
	Gemma	58.67	54.33	67.33	2.00	98.33	10.00	36.33	39.33	49.67	0.00	99.33	19.00
	Llama-8B	75.33	66.67	77.33	19.00	73.67	86.33*	53.67	59.00	69.67	21.00	77.33	88.00 [†]
	Llama-1B	48.00	11.67	10.67	5.33	47.67	37.00*	37.67	20.67	14.33	23.67	36.33	19.33
Reasoning (R)	GPT-4o	92.00	93.67	93.67	92.67	95.00	94.00	82.33	74.33	81.67	41.81	59.33	68.33
	GPT-3.5	79.60	81.00	80.67	80.00	78.00	79.33	52.67	53.33	60.67	37.67	48.67	53.67
	Gemma	58.67	54.85	60.67	54.00	57.00	45.67	36.33	38.00	45.30	0.67	30.67	2.67
	Llama-8B	75.33	81.00	81.00	76.67	76.67	77.33	53.67	66.33	70.33	48.00	74.67	57.67
	Llama-1B	48.00	47.00	46.33	42.00	53.67	49.33	37.67	28.33	36.33	17.33	26.67	20.33

Table 1: Accuracy on SAQ (baseline), MCQ, and TFQ question types. The “True” and “False” columns indicate that the correct answers are true and false, respectively. See Section 3.6 for the notations and significance of color highlights.

the difference (denoted as Δ) is the LLM performance of the baseline less that of the compared question type. Q and S for TFQs denote the question and the statement formats, respectively, where “+st” indicates the instruction that asks to solve the problem first before answering the question. The term “LLMs” used in analyzing experimental results refers to the specific LLMs under study.

4 RQ1: Do question types impact the final selection accuracy of LLM?

Table 1 (top) shows FS accuracy for commonly used 5-option MCQs and TFQs with the question format. Our analysis description focuses on the cases with *statistically significant differences*.

4.1 Quantitative reasoning with GSM8K300

SAQ vs MCQ: FS accuracy on SAQ is better than MCQ in 53% of the cases (8/15), but worse than MCQ in 13.3% of the cases (2/15). SAQ wins over MCQ (o-SEO) in 80% of the cases (4/5), suggesting the difficulty in choosing the correct option with SEO as the correct answer.

SAQ vs TFQ: FS accuracy between the two question types is statistically significant in 6 of 10 cases (60%). Four of these cases have TFQ (False) winning over SAQ. See Table 1 (top) GSM8K300. Both GPT models perform worse on SAQs.

MCQ vs TFQ: FS accuracy on TFQ (False) is at least 7.88 higher than that of MCQ (U) in the four cases indicated by * in Table 1.

4.2 Deductive reasoning with bAbI300

SAQ vs MCQ: SAQ wins over MCQ and loses in as many statically significant cases as those on GSM8K300. However, most models perform worse on this dataset than on GSM8K300, for instance, GPT-3.5 on MCQ (U-SEO).

SAQ vs TFQ: FS accuracy of all LLMs is influenced positively or negatively with TFQs with “False” as the correct answer.

MCQ vs TFQ: The symbol [†] in Table 1 indicates cases where TFQ (False) wins over MCQ (U) by at least 20 percent.

5 RQ2: Do question types impact the reasoning accuracy of LLM?

Table 1 (bottom) shows reasoning accuracy for 5-option MCQs and TFQs with the question format.

5.1 Quantitative reasoning with GSM8K300

SAQ vs MCQ: 80% of the cases (12/15) do not show statistically significant reasoning accuracy differences. Reasoning accuracy difference positively co-occurs with FS accuracy difference in only one case, Llama-1B on MCQ (oSEO).

SAQ vs TFQ: 70% of the cases (7/10) are not significant. The reasoning accuracy mostly does not correlate with FS accuracy. When “False” is the correct answer, LLMs need not be as accurate to choose the final correct answer. Gemma shows the opposite trend, with SAQ winning over TFQ (False).

MCQ vs TFQ: Differences in reasoning accuracy are less pronounced compared to FS accuracy.

5.2 Deductive reasoning with bAbI300

SAQ vs MCQ: SAQ wins over MCQ in 46.67% (7/15) and loses in 20 % (3/15) of the cases. Four cases of reasoning accuracy difference co-occur with FS accuracy difference.

SAQ vs TFQ: SAQ wins over MCQ in 50% (5/10) and loses in 10% (1/10) of the cases. Four cases of reasoning accuracy difference co-occur with FS accuracy difference.

MCQ vs TFQ: The notable case includes MCQ (U) winning over TFQ (False) by 32.67 on Gemma.

Summary: Reasoning accuracy does not always correlate with FS accuracy.

6 RQ3: What are the patterns of the incorrect outputs by LLMs on different question types?

This analysis focuses only on incorrect output cases where either the final selection, the reasoning, or both are incorrect. Figure 2 shows examples of different patterns. Our analysis reveals distinct patterns of mistakes for different types of questions. Figure 3 shows statistics of the patterns by different models on 5-option MCQs (U) and TFQ (Question-False).

6.1 Correct Final Selection but Wrong Reasoning

The following patterns cause LLMs to perform better on MCQs or TFQs than SAQs.

A. Good with guessing: LLMs make incorrect reasoning steps but guess the final answer correctly for MCQs and TFQs. Regardless of whether “True” or “False” is the correct answer of TFQs, LLMs sometimes make a correct guess first before outputting their reasoning. This pattern occurs in approximately 96% of incorrect outputs with Gemma on TFQ in Figure 3(b).

B. Proximity helps correctness: LLMs make nearly correct reasoning and calculations. With slightly off results, LLMs can still choose the correct option. This pattern appears in approximately 1% of incorrect outputs, with Gemma on MCQs and Llama-8B on TFQs.

C. Advantage of “Something else”: When this option is the correct answer, LLMs choose correctly sometimes, even with wrong reasoning steps. Around 11% of incorrect outputs from GPT-4o on MCQs are in this category, as shown in Figure 3(a).

6.2 Wrong Final Selection but Correct Reasoning

LLMs sometimes return the wrong final selection after the correct reasoning, which leads to better LLM performance on SAQs than the other question types. We categorize the failure cases as follows.

D. Incorrect option selection: LLMs perform a valid calculation but choose the wrong option due to incorrect decimal or unit conversion for MCQs or an incorrect guess on TFQs before reasoning. Figure 3(a) shows approximately 9% of Llama’s outputs on MCQs select incorrect options.

E. Misunderstanding “Something else”: LLMs fail to select “None of the above” or “Something else” in MCQs when appropriate, repeatedly recalculating to match its output to a specific numerical option. Figure 3(a) shows Gemma and Llama misunderstand “SEO” in approximately 29% and 14% of the outputs on MCQs, respectively.

F. Missing final selection output: LLMs sometimes fail to select the correct option despite completing the correct reasoning steps. This pattern occurs primarily with GPT-4o on MCQs and TFQs.

6.3 Wrong Final Selection and Wrong Reasoning

The failure cases are categorized as follows.

G. Faulty reasoning leading to incorrect answers: Due to errors in the reasoning or calculation steps, LLMs arrive at an incorrect final answer. Figure 3(a) shows approximately 40% by GPT4o on MCQ and 65% by Llama on TFQ outputs.

H. Stopping in the middle of reasoning: LLMs stop their reasoning process in the middle, resulting in incorrect reasoning steps and missing the final selection. Approximately 10% of Llama’s outputs have this pattern on both question types.

I. Selecting different wrong options with wrong reasoning steps: LLMs have wrong reasoning steps, leading to an incorrect answer. This pattern occurs primarily around 47% with Gemma on MCQs.

J. Selecting nothing with wrong reasoning steps: Only 5% of Llama’s outputs on MCQs are this pattern.

7 RQ4: What factors of MCQs influence LLM performance?

7.1 Positions of Correct Answers

Table 1 shows that the LLM performances fluctuate depending on the position of the correct answers.

Prompt	Jillian's handbag cost \$20 less than 3 times as much as her shoes cost. If her shoes cost \$80, how much did her bag cost? Choose from the option below: a)533.33 b)231.0 c)264.0 d)220.0 e)Something else	Answer	d)220
Correct Final Selection & Wrong Reasoning			
	Let's first calculate 3 times the cost of Jillian's shoes: $3 * \$80 = \240		Let's first calculate 3 times the cost of Jillian's shoes: $3 * \$80 = \240
	Now, we know that Jillian's handbag cost \$20 more than 3 times the cost of her shoes. So, we add \$20 from \$240: $\$240 + \$20 = \$260$ Therefore, Jillian's handbag cost \$260.		Now, we know that Jillian's handbag cost \$20 less than 3 times the cost of her shoes. So, we subtract \$20 from \$240: $\$240 - \$20 = \$220$ Therefore, Jillian's handbag cost \$220.
	Wrong Reasoning So, the closest answer is d) \$220.0	Correct Final Selection	Correct Reasoning So, the closest answer is b) \$231.0
Wrong Final Selection & Correct Reasoning			
			So, the closest answer is Wrong Final Selection
Wrong Final Selection & Wrong Reasoning			
			So, the closest answer is Wrong Final Selection

Figure 2: Examples of different patterns of incorrect outputs by LLMs for MCQ questions

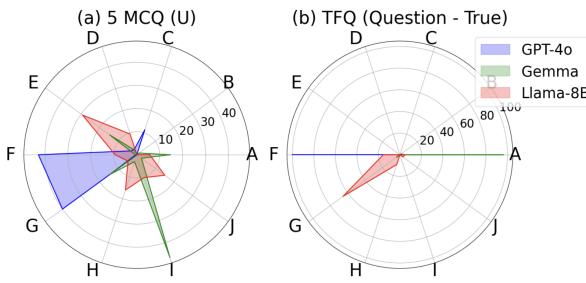


Figure 3: Percent of incorrect pattern outputs by LLMs (GPT-4o, Gemma, and Llama-8B) on (a) 5 MCQ (U) from 0 to 50% and (b) TFQ (Question - True) from 0 to 100%. A-J indicate the letters in front of each pattern.

For most cases, except for GPT-4o and Llama-1B, the models perform best without SEO as a correct option (U-SEO). Using only SEOs as the correct answers (oSEO) results in a significant performance drop on MCQs. The models sometimes selected SEOs as their final selection but struggled to select SEOs when their final reasoning answer did not match one of the provided options. They often chose the option with the number closest to their computed output rather than selecting SEOs.

7.2 Number of Options

Using 2, 3, or 4 options in MCQs does not significantly influence accuracy (Li et al., 2024). We investigate whether a larger number of options would change the finding on the FS accuracy. To assess the statistical significance of differences between the accuracies of 5 and 11-option MCQs, Recall the statistical test discussion in Section 3.6. **GSM8K300:** Table 2 shows no statistically significant differences between the 5-option and 11-option MCQs for GPT-4o and the Llama models. For the other models, the FS accuracy on the 5-option MCQs is better in 4 out of 12 cases. Gemma shows the largest absolute differences in both FS and reasoning accuracy. The 5-option

Models	GSM8K300		bAbI300		
	Δ FS	Δ R	Δ FS	Δ R	
GPT-4o	U	-1.22	0.00	4.67	-1.00
	U-SEO	-0.34	0.00	3.00	-4.33
	oSEO	2.00	-0.73	7.67	1.81
GPT-3.5	U	3.34	0.00	-7.67	6.00
	U-SEO	12.00	1.67	1.66	2.33
	oSEO	5.33	0.33	8.00	17.00
Gemma	U	12.00	-4.82	16.66	16.00
	U-SEO	21.66	5.00	26.00	19.66
	oSEO	-2.33	-0.67	-5.33	-2.66
Llama-8B	U	5.00	1.00	1.00	2.00
	U-SEO	3.00	-1.33	6.00	-1.33
	oSEO	5.00	-0.33	6.00	4.33
Llama-1B	U	3.34	2.00	12.34	2.00
	U-SEO	3.67	0.66	7.00	5.66
	oSEO	4.33	-2.00	21.34	-5.67

Table 2: Accuracy differences on 5-option MCQs (baseline) vs 11-option MCQs. Recall the notations and significance of color highlights in Section 3.6.

MCQs clearly win on Gemma with and without “Something else” as the correct answer. **bAbI300:** All LLMs exhibit statistically significant differences in FS accuracy. The 5-option MCQs generally win over 11-option MCQs. Additionally, Llama-8B does not show a statistically significant difference in reasoning accuracy.

7.3 Something else vs None of the above

For this set of experiments, we replaced “Something else” with “None of the above” option to investigate the impact of word choices. Table 3 shows the FS accuracy of GPT-3.5 improves in all cases with “None of the above” as the final option. However, the statistically significant improvements are limited to the 5-option MCQs when SEO is the correct answer for one-fifth of the reasoning problems (U) and all the problems (oSEO). These results indicate that LLMs are sensitive to variations in the wording of MCQ options.

#Options		Something else	None of the above	Δ FS
5 opts	U	69.67	74.33	-4.66
	U-*	83.67	85.33	-1.66
	o*	16.67	32.33	-15.66
11 opts	U	66.33	68.00	-1.67
	U-*	71.67	73.67	-2.00
	o*	11.33	12.00	-0.67

Table 3: FS accuracy differences of GPT-3.5 on MCQs from GSM8K300 with “Something else” (baseline) vs “None of the above” as the final option; “*” indicates either SEO or “None of the above.” Recall the notations and significance of color highlights in Section 3.6.

Models	GSM8K300		bAbI300		
	Δ FS	Δ R	Δ FS	Δ R	
GPT-4o	Q-S	0.67	1.33	11.67	-2.00
	Q-S+st	-0.67	-0.98	-9.66	-10.00
GPT-3.5	Q-S	-1.33	-2.67	-50.67	-29.00
	Q-S+st	-2.00	-1.34	-33.00	-24.34
Gemma	Q-S	5.33	-1.67	-0.67	-52.33
	Q-S+st	-12.33	-6.13	-32.00	-55.66
Llama-8B	Q-S	-4.66	-3.33	-1.00	-7.00
	Q-S+st	-15.00	-3.00	-22.00	-21.33
Llama-1B	Q-S	-7.66	1.67	-11.67	-30.00
	Q-S+st	-4.33	-0.66	-6.67	-10.33

Table 4: Accuracy differences on TFQs using “True” as the correct answer of the question format (Q) (baseline) vs the statement format (S). Recall the notations and significance of color highlights in Section 3.6.

8 RQ5: What factors of TFQs influence LLMs’ performance?

8.1 Question Format vs Statement Format

Table 4 shows influences of different TFQ formats. FS accuracy is worse for the question format than that of the statement format, except for GPT4-o on bAbI300 and Gemma on GSM8K300. Reasoning accuracy differences do not always correlate with FS accuracy differences.

8.2 True vs False as Correct Answers

Table 5 shows the results. **GSM8K300:** On FS accuracy, GPT-3.5 and Llama-8B tend to perform significantly better with “False” as correct answers, while Gemma and Llama-1B perform better on TFQs with “True” as correct answers. For GPT4-o, there are no significant differences in reasoning and FS accuracy. **bAbI300:** On FS accuracy, GPT models tend to perform significantly better with “False” as correct answers, while Gemma performs better on TFQs with “True” as correct answers. LLMs mostly reason better with TFQs with “True” as correct answers.

Models	GSM8K300		bAbI300		
	Δ FS	Δ R	Δ FS	Δ R	
GPT-4o	Q	-4.33	1.00	-29.33	-9.00
	Q+st	-4.00	0.34	-15.33	17.33
	S	-5.00	0.02	-43.67	-14.67
	S+st	-3.33	0.65	2.33	27.67
GPT-3.5	Q	-21.00	-1.33	-76.34	-5.00
	Q+st	-16.00	0.00	-27.67	16.00
	S	-18.33	1.34	-20.33	26.34
	S+st	-13.67	3.34	4.00	67.34
Gemma	Q	88.33	11.33	80.33	28.00
	Q+st	12.67	3.87	33.67	18.00
	S	69.67	23.67	82.67	79.00
	S+st	60.33	67.00	99.00	89.33
Llama-8B	Q	-12.66	-0.66	-12.67	17.00
	Q+st	-18.00	1.67	-18.67	15.67
	S	-2.34	3.67	-6.00	36.00
	S+st	9.33	5.34	15.33	44.66
Llama-1B	Q	10.67	4.34	-8.67	6.34
	Q+st	34.67	2.34	1.33	31.67
	S	8.33	15.00	0.33	30.34
	S+st	44.00	12.00	3.67	30.00

Table 5: Accuracy differences on TFQs with “True” (baseline) vs “False” as the correct answers. Recall the notations and significance of color highlights in Section 3.6.

8.3 True or False vs Yes or No

Are LLMs sensitive to the choice of words for TFQs? In this set of experiments, we replaced “True or False” with “Yes or No” in the prompts, keeping everything else identical to the original TFQs prompts. Table 6 shows the FS accuracies of GPT-3.5 drops with statistical significance when “Yes or No” instead of “True or False” in all scenarios except when the question format is used together with the guiding instruction to solve the problem first (+st). Using “Yes or No” has a performance drop of 15 and 26.67% without the guiding instruction but is only around 0.33% otherwise.

Types	True or False	Yes or No	Δ FS
Q	75.67	49.00	26.67
Q+st	80.33	80.00	0.33
S	77.00	62.00	15.00
S+st	82.33	87.67	-5.34

Table 6: GPT-3.5’s FS accuracy on TFQs between “True or False” (baseline) and “Yes or No” using GSM8K300. Recall the notations and significance of color highlights in Section 3.6.

9 Conclusions and Future Work

LLM performance on different question types fluctuates. Our experimental results show that the reasoning accuracy is not always correlated with the final selection accuracy for multiple-choice ques-

tions and True or False questions. To improve LLM performance on these question types, it is important to improve both reasoning accuracy and the accuracy of selecting the final answer. The trends in significant differences vary across different reasoning tasks. For MCQs, factors such as the position of correct answers, the number of options, and the selection of words in options impact LLM performance. In TFQs, factors such as type of instruction, whether “True” or “False” is the correct answer, and the selection of words influence LLM performance.

10 Limitations

This paper presents a comprehensive evaluation of the impact of different question types on LLM performance, but with the following limitations. The experiments do not involve prompting methods or few-shot learning strategies. The proposed measure of reasoning accuracy does not account for accuracy at individual reasoning steps, nor does it consider how the number of reasoning steps affects performance. While deeper insights could be obtained through such a study, it would require substantially more manual labeling effort. Lastly, due to computational constraints, the experiments were conducted on subsets of the two datasets. Nevertheless, we conducted statistical tests to evaluate the impact of question types on LLM performance.

11 Ethical Considerations and Potential Risk

The findings presented in this study are based on experiments conducted on five different large language models for quantitative and deductive reasoning tasks on standard open-source datasets under a zero-shot setting. Each of these language models has been trained on a distinct corpus with a specific training objective, making the obtained outputs dependent on the experimental setup. Our objective is to empirically demonstrate that LLM performance fluctuates across different question types. This variation highlights the need to improve both reasoning accuracy and the accuracy of selecting the final answer to enhance overall performance. Consequently, the results and analysis may not generalize to other large language models or to the same models if fine-tuned on the same or different datasets.

12 Acknowledgement

This work is partially supported by the NSF Grant No. 2152117. The findings, interpretations, and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the funding agency. The authors gratefully acknowledge Owen Harty, Vicky Lee, Noah Shpardsen, and Yonas Sium for their invaluable assistance with annotation.

References

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or abduction: How do LLMs answer multiple-choice questions without the question?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. [Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge.](#) *CoRR*, abs/2102.03315.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. [Piqa: Reasoning about physical commonsense in natural language.](#) In *AAAI Conference on Artificial Intelligence*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions.](#) 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. [Transformers as soft reasoners over language.](#) *arXiv preprint arXiv:2002.05867*.

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

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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. **The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Jörg Frohberg and Frank Binder. 2022. **CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. **Choice of plausible alternatives: An evaluation of commonsense causal reasoning**. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. **Towards reasoning in large language models: A survey**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. **TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. **Mawps: A math word problem repository**. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. **Natural questions: A benchmark for question answering research**. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms? *arXiv preprint arXiv:2403.17752*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. **Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Leora Morgenstern, Ernest Davis, and Charles L. Ortiz. 2016. **Planning, executing, and evaluating the winograd schema challenge**. *AI Magazine*, 37(1):50–54.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. **Ms marco: A human generated machine reading comprehension dataset**.

OpenAI. 2023. **Gpt-4 technical report**. *Preprint*, arXiv:2303.08774.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. **The LAMBADA dataset: Word prediction requiring a broad discourse context**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattacharya, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. **Reasoning with language model prompting: A survey**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. **Resolving complex cases of definite pronouns: The Winograd schema challenge**. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana. Association for Computational Linguistics.

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

Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022. Robustlr: A diagnostic benchmark for evaluating logical robustness of deductive reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9614–9631.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Seok Hwan Song and Wallapak Tavanapong. 2024. How much do prompting methods help llms on quantitative reasoning with irrelevant information? In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2128–2137.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. *GLUE: A multi-task benchmark and analysis platform for natural language understanding*. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. “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 and virtual meeting. Association for Computational Linguistics.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. *Crowdsourcing multiple choice science questions*. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. *Towards ai-complete question answering: A set of prerequisite toy tasks*. Cited by: 177.

Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. *A broad-coverage challenge corpus for sentence understanding through inference*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

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. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Related Work about Datasets

LLMs have been tested with various reasoning tasks: mathematical reasoning, deductive reasoning, causal reasoning, visual reasoning, and commonsense reasoning, just to name a few (Qiao et al., 2023). Benchmarking is one of the most widely adopted methods to evaluate LLM performance on these tasks. We list existing datasets on two reasoning tasks and question types used in this study.

A.1.1 Quantitative and Deductive Reasoning Tasks

Math Word Problems datasets (e.g., MAWPS (Koncel-Kedziorski et al., 2016), SVAMP (Patel et al., 2021), and GSM8K (Cobbe et al., 2021)) are commonly used benchmarks for quantitative reasoning tasks. Recent datasets, e.g., GSM-IC

(Shi et al., 2023) and MPN (Song and Tavanapong, 2024)), include irrelevant information to increase reasoning complexity. Some benchmarks target classical logical reasoning, including deductive reasoning tasks. Example datasets include reasoning benchmark in philosophy and logic (Yu et al., 2024), bAbI (Weston et al., 2016), RuleTaker (Clark et al., 2020), ProofWriter (Tafjord et al., 2020), and RobustLR (Sanyal et al., 2022).

A.1.2 Question Types

Question types can be broadly categorized into short/long answer questions, multiple-choice questions, and true/false or yes/no questions. We reviewed existing LLM benchmark datasets listed in the Github². Most of the datasets have only one question type. A few datasets (Bisk et al., 2019; Nguyen et al., 2016) have more than one question type. Multiple-choice questions appear most frequently. True/false or yes/no questions appear least often. Short/long answer questions are included in diverse benchmarks for Math Word Problems, reading comprehension, and language understanding.

Examples of datasets that include different question types are listed below.

Multiple-choice questions: MMLU (Hendrycks et al., 2021), ARC (Bhakthavatsalam et al., 2021), HellaSwag (Zellers et al., 2019), MultiNLI (Williams et al., 2018), Winogrande (Sakaguchi et al., 2021), PDP (Morgenstern et al., 2016), SuperGLUE-WSC (Sakaguchi et al., 2021), DPR (Rahman and Ng, 2012), KnowRef (Emami et al., 2019), COPA (Gordon et al., 2011), Winogender (Rudinger et al., 2018), SciQ (Welbl et al., 2017), CRASS (Frohberg and Binder, 2022), PIQA (Bisk et al., 2019), CommonsenseQA, Social IQa, HellaSWAG (Zellers et al., 2019), and Numersense (Lin et al., 2020). Three to five options are common, whereas two and eleven options are less common.

Short/long answer questions: GSM8K (Cobbe et al., 2021), SVAMP, MAWPS, Natural Questions (Kwiatkowski et al., 2019), GLUE (Wang et al., 2018), LAMBADA (Paperno et al., 2016), SuperGLUE (Wang et al., 2019), TriviaQA (Joshi et al., 2017), DROP (Dua et al., 2019), MS-MARCO (Nguyen et al., 2016), and PIQA.

True/false or yes/no questions: BoolQ (Clark et al., 2019), MS-MARCO (Nguyen et al., 2016).

Most existing datasets mostly adopt one question

type. A few have more than one question type: PIQA and MS-MARCO.

A.2 Methods to Generate Wrong Options

To generate wrong answers for the deductive reasoning task, our program randomly picked three values from the five possible wrong answers for the 5-option MCQs. These values are ‘kitchen’, ‘bedroom’, ‘garden’, ‘bathroom’, ‘hallway’, and ‘office’. For the 11 option MCQs, nine values were randomly selected from the eleven values of wrong answers, including ‘beds’, ‘baths’, ‘table’, ‘desk’, and ‘livingroom’ in addition to the five words used to generate wrong answers for the 5-option MCQs.

When SEO is the correct answer, the ground truth needs to be removed from the options. For the quantitative reasoning task, we randomly chose a number with the same number of digits. While four or ten values are selected from the possible answers for the deductive reasoning task.

A.3 Examples of Datasets

See Tables 7 and 8.

A.4 Prompt Designs to Format Question Types

See Tables 9 to 14.

A.5 Complete Accuracy Report for All Configurations Except for “None of the above” for MCQs and “Yes or No” for TFQs

See Tables 15 to 16.

²https://github.com/leobeeson/llm_benchmarks

Question:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

Ground Truth: 23

Table 7: Example of GSM8K300

Question:

Daniel journeyed to the office. John traveled to the office.

Mary went back to the garden. Mary went to the bedroom. Mary put down the milk.

Where was the milk before the bedroom?

Ground Truth: garden

Table 8: Example of bAbI300

Prompt:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

Output from GPT4o:

Chatty prepared 3 dozen eggs.

Since 1 dozen equals 12 eggs, 3 dozen eggs is:

$3 \times 12 = 36$ eggs

She has 4 children, and each child should receive the same number of eggs.

To find out how many eggs each child receives,

divide the total number of eggs by the number of children:

$$\frac{36}{4} = 9$$

Each child receives 9 eggs.

Table 9: Example of SQA

Prompt:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

Choose from the option below:

a)41.0 b)27.6 c)24.15 d)1350.0 e)Something else

Output from GPT4o:

Let's solve the problem step by step.

1. **Red Cars**: The rink has 12 red cars.

2. **Green Cars**:

They have 2 fewer green cars than red cars. So, the number of green cars is:

$$12 - 2 = 10$$

3. **Blue Cars**:

They have 3 times the number of blue cars as green cars. So, the number of blue cars is:

$$3 \times 10 = 30$$

4. **Total Cars**:

The total number of cars is 75.

5. **Yellow Cars**:

Let y be the number of yellow cars. The equation for the total number of cars is:

$$12 + 10 + 30 + y = 75$$

Simplifying this equation:

$$52 + y = 75$$

Solving for y :

$$y = 75 - 52 = 23$$

So, the rink has 23 yellow cars. The correct answer is "e) Something else".

Table 10: Example of MCQ

Prompt <Question>:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

Is the answer 23? Choose one of these. True or False

Output from GPT4o:

True.

Chatty prepared 3 dozen eggs, which is equal to $3 \times 12 = 36$ eggs.

If these 36 eggs are distributed equally among her 4 children, each child would receive:

$$\frac{36}{4} = 9$$

So, each child receives 9 eggs. Therefore, the statement "each child receives 9 eggs" is true.

Table 11: Example of TFQ Question

Prompt <Question + Step instruction>:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

Is the answer 23? Solve the question first and choose. True or False

Output from GPT4o:

To solve the problem, we first need to determine the total number of eggs.

Chatty prepared 3 dozen eggs. Since one dozen equals 12 eggs, 3 dozen eggs is:

$$3 \times 12 = 36 \text{ eggs}$$

Chatty has 4 children, and each child should receive the same number of eggs.

To find out how many eggs each child receives,

we divide the total number of eggs by the number of children:

$$\frac{36}{4} = 9$$

So, each child receives 9 eggs.

Therefore, the statement "Is the answer 9?" is True.

Table 12: Example of TFQ Question + Step instruction

Prompt <Statement>:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

The answer of the question is 23. True or False

Output from GPT4o:

True.

Chatty prepared 3 dozen eggs, which is equal to $3 \times 12 = 36$ eggs.

If these 36 eggs are divided equally among her 4 children, each child would receive:

$$\frac{36}{4} = 9$$

Therefore, each child receives 9 eggs.

The statement that each child receives 9 eggs is true.

Table 13: Example of TFQ Statement

Prompt <Statement + Step instruction>:

Chatty prepared 3 dozen eggs for her 4 children's Easter activity.

Assuming each child gets the same number of eggs, how many eggs does each child receive?

The answer of the question is 23. Solve the question first and choose. True or False

Output from GPT4o:

To solve the problem, we start by determining the total number of eggs Chatty prepared.

Since 1 dozen equals 12 eggs, 3 dozen eggs is:

$$3 \times 12 = 36 \text{ eggs}$$

Chatty has 4 children, and each child should receive the same number of eggs.

To find out how many eggs each child receives,

we divide the total number of eggs by the number of children:

$$\frac{36}{4} = 9 \text{ eggs per child}$$

Therefore, each child receives 9 eggs.

The statement that the answer to the question is 9 is true.

So, the answer is: True.

Table 14: Example of TFQ Statement + Step instruction

Accuracy	Models	SAQ (baseline)	MCQ						TFQ							
			5 options			11 options			True as correct answer			False as correct answer				
			U	U-SEO	oSEO	U	U-SEO	oSEO	Q	Q+st	S	S+st	Q	Q+st	S	S+st
Final Selection	GPT-4o	92.00	89.45	91.33	97.67	91.67	91.67	95.67	93.00	93.33	92.33	94.00	97.33	97.33	97.33	97.33
	GPT-3.5	79.60	69.67	83.67	16.67	66.33	71.67	11.33	75.67	80.33	77.00	82.33	96.67	96.33	95.33	96.00
	Gemma	58.67	54.33	67.33	2.00	42.33	45.67	4.33	98.33	77.67	93.00	90.00	10.00	65.00	23.33	29.67
	Llama-8B	75.33	66.67	77.33	19.00	61.67	74.33	14.00	73.67	64.33	78.33	79.33	86.33	82.33	80.67	70.00
	Llama-1B	48.00	11.67	10.67	5.33	8.33	7.00	1.00	47.67	65.00	55.33	69.33	37.00*	30.33	47.00	25.33
Reasoning	GPT-4o	92.00	93.67	93.67	92.67	93.67	93.67	93.40	95.00	93.67	93.67	94.65	94.00	93.33	93.65	94.00
	GPT-3.5	79.60	81.00	80.67	80.00	81.00	79.00	79.67	78.00	81.33	80.67	82.67	79.33	81.33	79.33	79.33
	Gemma	58.67	54.85	60.67	54.00	59.67	55.67	54.67	57.00	60.87	58.67	67.00	45.67	57.00	35.00	46.67
	Llama-8B	75.33	81.00	81.00	76.67	80.00	82.33	77.00	76.67	79.67	80.00	82.67	77.33	78.00	76.33	77.33
	Llama-1B	48.00	47.00	46.33	42.00	45.00	45.67	44.00	53.67	50.67	52.00	51.33	49.33	46.00	37.00	39.33

Table 15: Accuracy on SAQ (baseline), MCQ, and TFQ question types with GSM8K300. Recall the notations and significance of color highlights in Section 3.6.

Accuracy	Models	SAQ (baseline)	MCQ						TFQ							
			5 option			11 option			True			False				
			U	U-SEO	oSEO	U	U-SEO	oSEO	Q	Q+st	S	S+st	Q	Q+st	S	S+st
Final Selection	GPT-4o	82.33	79.00	80.00	83.67	64.33	77.00	76.00	67.67	82.67	56.00	92.33	97.00	98.00	99.67	90.00
	GPT-3.5	52.67	30.00	38.33	10.67	37.67	36.67	2.67	16.33	67.00	67.00	100.00	92.67	94.67	87.33	96.00
	Gemma	36.33	39.33	49.67	0.00	22.67	23.67	5.33	99.33	68.00	100.00	100.00	19.00	34.33	17.33	1.00
	Llama-8B	53.67	59.00	69.67	21.00	58.00	63.67	15.00	77.33	64.33	78.33	86.33	88.00	83.00	84.33	71.00
	Llama-1B	37.67	20.67	14.33	23.67	8.33	7.33	2.33	36.33	49.00	48.00	55.67	45.00	47.67	39.67	52.00
Reasoning	GPT-4o	82.33	74.33	81.67	41.81	75.33	86.00	40.00	59.33	82.67	61.33	92.67	68.33	65.33	76.00	65.00
	GPT-3.5	52.67	53.33	60.67	37.67	47.33	58.33	20.67	48.67	69.33	77.67	93.67	53.67	53.33	51.33	26.33
	Gemma	36.33	38.00	45.30	0.67	22.00	25.67	3.33	30.67	34.67	83.00	90.33	2.67	16.67	4.00	1.00
	Llama-8B	53.67	66.33	70.33	48.00	64.33	67.33	43.67	74.67	66.00	81.67	87.33	57.67	50.33	45.67	42.67
	Llama-1B	37.67	28.33	36.33	17.33	31.00	30.67	23.00	26.67	40.67	56.67	55.00	20.33	19.06	26.33	26.67

Table 16: Accuracy on SAQ (baseline), MCQ, and TFQ question types with bAbI300. Recall the notations and significance of color highlights in Section 3.6.