Can multiple-choice questions really be useful in detecting the abilities of LLMs?

Wangyue Li[†♢], **Liangzhi Li**[†♦*], **Tong Xiang**[†♦], **Xiao Liu**[†], **Wei Deng**[♢♡], **Noa Garcia**[♦]
[†]Meetyou AI Lab, [♢]Southwestern University of Finance and Economics, [♦]Osaka University
[♡]Chongqing University of Posts and Telecommunications
{liliangzhi,liuxiao}@xiaoyouzi.com
{alee90792,tongxiang39}@gmail.com
dengwei@swufe.edu.cn, noagarcia@ids.osaka-u.ac.jp

Abstract

Multiple-choice questions (MCQs) are widely used in the evaluation of large language models (LLMs) due to their simplicity and efficiency. However, there are concerns about whether MCQs can truly measure LLM's capabilities, particularly in knowledge-intensive scenarios where long-form generation (LFG) answers are required. The misalignment between the task and the evaluation method demands a thoughtful analysis of MCQ's efficacy, which we undertake in this paper by evaluating nine LLMs on four question-answering (QA) datasets in two languages: Chinese and English. We identify a significant issue: LLMs exhibit an order sensitivity in bilingual MCQs, favoring answers located at specific positions, i.e., the first position. We further quantify the gap between MCQs and long-form generation questions (LFGQs) by comparing their direct outputs, token logits, and embeddings. Our results reveal a relatively low correlation between answers from MCQs and LFGQs for identical questions. Additionally, we propose two methods to quantify the consistency and confidence of LLMs' output, which can be generalized to other QA evaluation benchmarks. Notably, our analysis challenges the idea that the higher the consistency, the greater the accuracy. We also find MCQs to be less reliable than LFGQs in terms of expected calibration error. Finally, the misalignment between MCQs and LFGQs is not only reflected in the evaluation performance but also in the embedding space. Our code and models can be accessed at https://github.com/Meetyou-AI-Lab/Can-MC-Evaluate-LLMs.

Keywords: Natural Language Processing, Large Language Model, Question Answering, Text Generation, Evaluation Methods

1. Introduction

Over the past few years, large language models (LLMs) have exhibited remarkable performance on a wide range of question-answering (QA) tasks (Brown et al., 2020; Kadavath et al., 2022; Robinson and Wingate, 2023). The evaluation of LLMs' strengths and limitations often relies on diverse benchmarks presented in different formats (Singhal et al., 2023; Liu et al., 2023b), domains (Jin et al., 2019; Zhong et al., 2020), and languages (Petroni et al., 2019; Bang et al., 2023). As previous research has shown (Liang et al., 2022; Chang et al., 2023; Li et al., 2023a; Chia et al., 2023), evaluation using benchmarks is essential for the detection and mitigation of various issues such as misinformation (Zheng et al., 2021; Gao et al., 2022), hate speech (ElSherief et al., 2021; Lu et al., 2023), and malicious uses (Xu et al., 2021; Ganguli et al., 2022; Shaikh et al., 2023; Zou et al., 2023). Such mechanisms are critical for safeguarding against harmful content and promoting responsible usage of LLMs in various contexts.

QA benchmarks come in a variety of formats, including *True/False questions* (TFQs) in which models predict whether a statement in the question is correct or not, *multiple-choice questions* (MCQs), in which multiple candidate answers accompany the input question, and *long-form generation questions* (LFGQs), in which a generated answer could span multiple sentences. Among these, multiple-choice is the most popular format as it allows a simple and quick assessment of model performance (Bhakthavatsalam et al., 2021; Ramamurthy and Aakur, 2022; Liu et al., 2023a; Huang et al., 2023). However, MCQs also present several limitations, such as potential misalignment with real-world use cases where LLMs are often required to answer questions in long-from generation format (Nuance, 2023; Bommasani et al., 2021). In addition, LLMs have been shown to be affected by changes in the position of the candidate answers (Zheng et al., 2023; Wang et al., 2023) and their contents (Pezeshkpour and Hruschka, 2023) when answering MCQs. The aforementioned problems highlight the limitations of MCQs benchmarks in evaluating LLMs, which could potentially lead to overestimation of LLMs capabilities.

With the above issues in mind, our motivation is to explore the limitations and characteristics of both MCQs and LFGQs as main evaluation formats in QA tasks. We aim to answer the following research questions:

1. How does the arrangement of options in MCQs influence LLMs' selection of responses?

* Corresponding author.

2. What methodologies can be employed to conduct comprehensive comparative experiments between MCQs and LFGQs? Additionally, what specific aspects should be considered when conducting comparative tests?

The answers to these questions contribute to the understanding and comparison between MCQs and LFGQs as evaluation formats in QA tasks. Given the prevalence of MCQs as the dominant evaluation format, our aim is to thoroughly examine their efficacy. This begins with a detailed exploration of MCQs' capabilities and subsequently extends to a comparative analysis with LFGQs, providing a comprehensive assessment of both formats.

We address the first question by conducting a series of experiments (§3) to reveal the sensitivity of LLMs to answering MCQs by applying slight perturbations to the positional order of the options. We find significant differences between the answers in multiple LLMs (§3.1). We also identify specific patterns of the selected answer according to its position that varies among different LLMs (§3.2). For the second question, we conduct comparative experiments (§4) to quantify the misalignment between MCQs and LFGQs on three different spaces: direct output space (§4.1), token logits space (§4.2), and hidden embedding space (§4.3). By doing this, we aim to gain a deeper understanding of the unique characteristics between the two types of questions.

Our key findings reveal that:

- LLMs exhibit order sensitivity in bilingual MCQs, favoring answers at the first position.
- Answers obtained from MCQs and LFGQs for identical questions have a low correlation.
- Higher consistency does not indicate better model performance.
- The misalignment between MCQs and LFGQs is evident in the evaluation performance as well as in the embedding space.

Overall, our study aims to provide a better understanding of the difference in QA formats in LLM evaluation, uncover underlying patterns, and shed light on the improvement of current methods.

2. Experimental Details

Models We use different models on different experiments, tailoring our choices based on the specific goals of each experiment, as summarized in Table 1. To check whether LLMs are sensitive to the order of the candidate answers (§3), we evaluate three models: **ChatGLM-6B** (Zeng et al., 2023; Du et al., 2022) and two models from the GPT family, namely **GPT-3.5-turbo** (OpenAI, 2023b) and **GPT-4** (OpenAI, 2023a). In comparing MCQs and LFGQs (§4), we again use different models on the three different spaces. For the direct output space (§4.1), we use GPT-3.5-turbo, GPT-4, and ChatGLM-6B, considering both their diversity and performances. For the token logits space (§4.2), we only test GPT-3.5-turbo, as it is the only model that can output token probabilities (Manakul et al., 2023) within three models. Finally, in the embedding space (§4.3), we conduct experiments with models from multiple popular LLM families across various sizes, including **StableLM-Tuned-Alpha-3/7B** (Stability-AI, 2023), **RedPajama-INCITE-Instruct-3B-v1** (Computer, 2023), **Llama-2-7b-chat-hf** (Touvron et al., 2023b), **Dolly-v2-2/7/12B** (Conover et al., 2023), **Vicuna-7b-v1.3** (Chiang et al., 2023), and **Open-llama-3/7B** (Touvron et al., 2023a).

Datasets We conduct experiments on four evaluation benchmarks:

1. **CARE-MI** (Xiang et al., 2023): A Chinese benchmark for evaluating LLM misinformation in the maternity and infant care domain. It includes $1,612$ LFGQs. The questions can also be obtained as MCQs and TFQs from the original MLEC-QA(Li et al., 2021) and MEDQA (Jin et al., 2020) datasets, according to the question generation process of CARE-MI, resulting in each question being formulated in the three formats: MCQ, LFGQ, and TFQ.

2. **M3KE** (Liu et al., 2023a): A dataset with $20,477$ standard Chinese questions for 71 tasks, encompassing all major levels of Chinese education system, including humanities, history, politics, law, education, psychology, science, technology, art and religion in MCQ format. Each question presents four candidate answers.

3. **ARC** (Clark et al., 2018): A dataset with natural, grade-school science questions (authored for human tests) in English. It is the largest public-domain set of this kind with $7,787$ questions. Each question contains four candidate answers.

4. **MATH**: A synthetic dataset randomly generated by a script with simple mathematical questions in English. Each question has four candidate answers.

As shown in Table 2, we ensure data balance by using a similar number of samples from each dataset. For the first research question (§3), we use all the datasets: CARE-MI, M3KE, ARC, and MATH. In the second research question (§4), we use the CARE-MI dataset for the direct output (§4.1) and token logits analysis (§4.2) as it is the only dataset offering the three different QA formats. The ARC dataset is used on the embedding space analysis (§4.3), wherein we extend its MCQs to LFGQs by not presenting the candidate answers to the LLMs.

Model	Are LLMs sensitive to order?		MCQ vs LFGQ		
	Order Sensitivity	Patterns Decomposition	Direct Output	Token Logits	Embeddings
GPT-3.5-turbo (OpenAI, 2023b)	✓	✓	✓	✓	
GPT-4 (OpenAI, 2023a)	✓	✓	✓		
ChatGLM-6B (Zeng et al., 2023)	✓	✓	✓		
Stablelm-tuned-α (Stability-AI, 2023)					✓
RedPajama-INCITE-3B-v1 (Computer, 2023)					✓
Dolly-v2 (Conover et al., 2023)					✓
Vicuna-7b-v1.3 (Chiang et al., 2023)					✓
Open-llama (Touvron et al., 2023a)					✓
Llama-2-7b-chat-hf (Touvron et al., 2023b)					✓

Table 1: Summary of the LLMs used in each of our analyses.

Dataset	Size	Lang.	Format
CARE-MI	344	ZH	MCQ/LFGQ/TFQ
M3KE	299	ZH	MCQ
ARC	291	EN	MCQ
MATH	300	EN	MCQ
Total	1,234	-	-

Table 2: Summary of the evaluation datasets. **Lang.** stands for the language of the datasets.

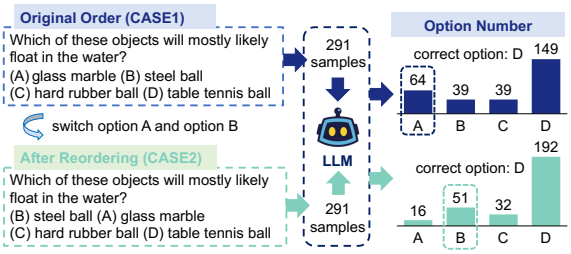

Figure 1: Example of order sensitivity experiments, in which the correct answer is D. In the ARC dataset, when predicting a wrong option (A, B, or C), the LLM prefers the option located in the first position.

The selection of benchmarks is guided by three specific criteria: (1) Source diversity: we aim to conduct our analyses across different domains. (2) Language: We presume language is a potential factor influencing LLMs evaluation performance, so we conduct experiments on two high-resource languages, i.e., Chinese and English. (3) Performance-level: By incorporating benchmarks with varying levels of LLMs demonstrated performance, we aim to better understand how model proficiency influences results in the different QA formats. Additionally, in the token logits space (§4.2), we investigate changing the number of candidate answers to explore their impact on the expected calibration error.

Prompt design To encourage the generation of concise responses, we provide LLMs with prompts both prior to (*pre-prompt*) and following (*post-prompt*) each question in any dataset format. Additionally, for MCQs, each question is accompanied by four candidate options, with only one being correct. The pre-prompt for MCQs is "*Please select a correct option*", while the post-prompt "*Only one option can be selected. No explanation is allowed*". For LFGQs, the post-prompt is "*Just answer in one sentence*", and for TFQs, it is "*Just answer 'yes' or 'no'*". We find that this prompt design facilitates the generation of brief content, aiding subsequent accuracy evaluation (§4.1) and automatic confidence calculation (§4.2).

3. Are LLMs sensitive to the order of candidate answers?

We first investigate how the arrangement of candidate answers in MCQs datasets affects the evaluation of LLMs. We find that LLMs consistently exhibit a strong preference for specific positions when presented with options in different orders, as illustrated in Figure 1.

3.1. Order Sensitivity

To check whether there are significant differences in LLMs' answers when the candidate options are arranged in a different order, we employ the chi-squared test (McHugh, 2013). To isolate the influence of the correct answer, we designate *option D* as the only correct option for all the questions. Then, we establish two scenarios: in **CASE1**, the option order is 'ABCD', and in **CASE2**, it is 'BACD'. Importantly, when arranging the option order, we also rearrange the contents and positions of each candidate option accordingly, rather than simply altering the numbering, as shown in Figure 1.
In the chi-squared test, we set the null hypothesis, H_0, stating that the responses in **CASE1** and **CASE2** originate from the same distribution. The chi-squared statistic is calculated as

$$X^2 = \sum_{i=0}^{N} \frac{(O_i - R_i)^2}{R_i}, \quad (1)$$

where $\sum_{i=0}^{N}$ is the sum of N candidate options , O_i the frequency of each option in **CASE1**, and R_i the frequency of each option in **CASE2**. With the significance test, we can get the p-value P from the

	GPT3.5	GPT4	ChatGLM
CARE-MI			
X^2	144.192	15.660	27.605
P	**0.000	*0.001	**0.000
Acc	0.203	0.637	0.378
Gap	−0.043	−0.029	−0.116
M3KE			
X^2	90.308	20.829	12.377
P	**0.000	**0.000	**0.006
Acc	0.381	0.632	0.411
Gap	−0.030	−0.017	+0.014
ARC			
X^2	36.515	2.681	10.511
P	**0.000	0.443	*0.015
Acc	0.512	0.935	0.553
Gap	+0.148	-0.031	-0.116
MATH			
X^2	25.129	4.513	90.566
P	**0.000	0.211	**0.000
Acc	0.597	0.780	0.480
Gap	+0.000	-0.023	-0.043

Table 3: LLMs' order sensitivity results. The rearrangement of options makes LLMs output different answers.* indicates $P < 0.05$, ** indicates $P < 0.001$, and bold indicates larger than significance level α.

chi-squared probabilities based on the chi-squared statistic and the degrees of freedom, $N - 1 = 3$. Additionally, we can calculate the accuracy gap, which represents the difference between the original accuracy and the accuracy after reordering. Results are shown in Table 3, from which we note the following observations:

1. There is a considerable disparity in LLMs' outputs across the two scenarios. Except for two instances,[1] all the results have a p-value $P < 0.05$, rejecting the null hypothesis and implying that the distribution of answers predicted by the model varies significantly when options A and B are interchanged. This indicates that the order of options significantly influences LLMs' predictions in MCQs datasets.

2. Among the GPT family, the rearrangement of options has a more pronounced effect on GPT-3.5-turbo, with bigger accuracy gaps, than on GPT-4.

3. Higher accuracy can mitigate significant differences in the order arrangement to some extent. Results from GPT-4 on the ARC and the MATH datasets indicate that high accuracies (≥ 0.780) can lead to not rejecting the null hypothesis.

4. There is no evident correlation between the accuracy gap and the original accuracy. A higher accuracy does not necessarily imply a lower gap between the two scenarios.

[1] GPT4 model on the ARC ($X^2 = 2.681$, $P = 0.443$) and the MATH ($X^2 = 4.513$, $P = 0.211$) datasets.

3.2. Pattern Decomposition

Next, we further explore the pattern decomposition of LLMs to investigate potential patterns underlying their sensitivity to order. We propose the following two hypotheses for exploration: (1) LLMs may have different positional preferences due to their different model bases; (2) LLMs may have different positional preferences depending on whether they have previously memorized the contents of the datasets.

We use the same LLMs as in §3.1, as they can provide concise answers to the questions and come from different model bases. Regarding the datasets, CARE-MI, M3KE, and ARC are derived from website documents, while the MATH dataset, synthetically generated by us, ensures that the LLMs have not been exposed to identical questions during training. Results are presented in Table 4, from which we can extract the following conclusions:

1. Within the GPT family, GPT-3.5-turbo and GPT4 exhibit different behavior. When predicting incorrect options (A, B, or C), GPT4 shows a stronger inclination towards the option positioned first compared to GPT-3.5-turbo. Specifically, when option B is presented first, GPT-3.5-turbo tends to lean towards selecting option B. Furthermore, ChatGLM-6B showcases a certain preference for the first two options.

2. The behavior of the LLMs remains consistent across datasets originating from different languages and sources. Hence, we can conclude that the dataset's language or source, regardless of whether the models were previously exposed to them or not, is not the underlying cause of the models' positional preferences.

3.3. Yes, LLMs are sensitive to ordering

Our experiments showed that the order of candidate answers in MCQs significantly impacts LLMs outputs. GPT-3.5-turbo and GPT4 exhibited different preferences, while ChatGLM-6B showed a certain preference for the first two positions. In addition, the positional preferences in each LLM seemed to remain consistent across datasets originating from different languages and sources.

Dataset	CASE	GPT-3.5-turbo			GPT-4			ChatGLM-6B		
		1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
CARE-MI	C1	41	33	26	44	37	19	37	42	21
	C2	83	8	9	61	26	13	61	23	16
M3KE	C1	25	35	40	49	26	25	33	39	28
	C2	75	1	24	49	31	30	46	32	22
ARC	C1	47	27	26	37	31	32	33	28	39
	C2	52	16	32	21	36	43	33	36	31
MATH	C1	22	60	18	41	30	29	17	69	14
	C2	61	3	36	47	28	25	22	57	11

Table 4: Pattern decomposition results. C1 refers to **CASE1**, and C2 to **CASE2**. The numbers indicate the percentage (%) of incorrect options (A, B, or C) that each model selects for each position (1st position, 2nd position, and 3rd position). Deeper background indicates a higher preference for that position.

These findings are problematic because they reveal potential biases and inconsistencies in LLM outputs, which can affect the reliability and accuracy of their responses. Failure to understand and address these preferences may lead to biased recommendations, inaccurate information retrieval, and flawed decision-making. It is important to develop methods to mitigate these effects, as well as to investigate evaluation protocols that are less impacted by positional preferences. In light of these observations, in the next section, we compare MCQs and LFGQs evaluation methods.

4. Multiple Choice vs Long Form Generation

To compare QA evaluation formats and gain a deeper understanding of LLMs evaluation protocols, we expose several LLMs to the same questions presented in different formats. Then, we analyze and compare the results in three spaces: the direct output space (§4.1), the token logits space (§4.2), and the embedding space (§4.3).

4.1. Direct Output

In the direct output space, which refers to the responses generated by the LLMs, accuracy is one of the most common evaluation metrics used for benchmarking purposes and performance assessment. The difference in accuracy between the MCQs and LFGQs formats is the first aspect we consider (§4.1.1). Additionally, we study the relationship between consistency and accuracy (§4.1.2) by exploring whether LLMs, if familiar with a particular concept, tend to generate responses that are similar and encompass consistent factual information (Manakul et al., 2023).

4.1.1. Accuracy

We randomly select 100 samples from the CARE-MI dataset and evaluate GPT4, GPT-3.5-turbo, and ChatGLM-6B on them. For MCQs, accuracy is computed as usual, i.e. if the predicted answer matches the ground truth, it is considered correct. For LFGQs, accuracy is determined through human evaluation, with 0 denoting an incorrect response and 1 a correct one. Figure 2 (top) compares the accuracy between MCQs and LFGQs across the three LLMs. Notably, the accuracies of MCQs are consistently higher than those of LFGQs. This difference can be attributed to the fact that MCQs offer candidate options, facilitating the prediction task. To delve deeper into the analysis, in Figure 2 (bottom), we visualize a matrix in which, for the same question, there are four scenarios: 1) the response is correct in both formats, 2) the response is incorrect in both formats, 3) the response is correct in MCQs but incorrect in LFGQs, and 4) the response is correct in LFGQs but incorrect in MCQs. Results show that there are a relatively large number of questions where the LLMs can respond correctly in MCQs, but fail in the LFGQs format. Furthermore, we quantify the differences in accuracy produced by the two formats with Pearson correlation coefficients. The obtained values are remarkably low: 0.39 for GPT4, 0.7 for GPT-3.5-turbo, and 0.33 ChatGLM-6B, clearly indicating that different versions of the same question yield different answers from the LLM.

4.1.2. Consistency

Next, we explore the relationship between consistency and accuracy. Consistency stands for the degree to which the LLMs provide the same answer when asked the same question multiple times. For example, an answer of 'AAAAB' is considered more consistent than 'BCDAB' when presented with the same question five times. To conduct this evaluation, we first define the quantitative measures for consistency and accuracy in a sequence of responses to a repeated question.

Formally, let $\mathcal{A} = \{A_1, A_2, ..., A_D\}$ be a sequence of answers, where a model is queried D times. Each answer $A_i \in \mathcal{A}$ is selected from a set of N unique options $\mathcal{O} = \{Opt_1, Opt_2, ..., Opt_N\}$.

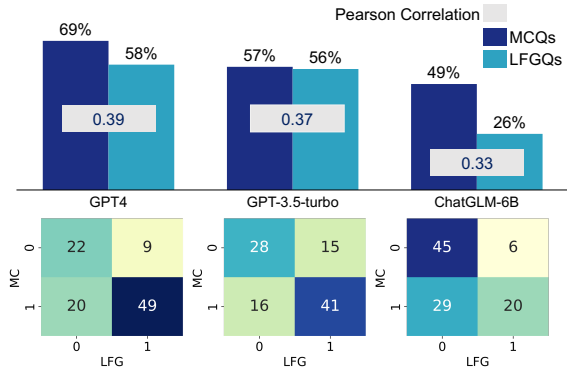

Figure 2: Comparison between MCQs and LFGQs on the CARE-MI dataset. Top: accuracy and pearson correlation. Bottom: MCQ vs LFGQs matrix.

From $\mathcal{A}$, we derive a count sequence $\mathcal{C} = \{\text{COUNT}(Opt_1), \text{COUNT}(Opt_2), ..., \text{COUNT}(Opt_N)\}$, where $\text{COUNT}(Opt_i)$ represents the number of occurrences of Opt_i in $\mathcal{A}$, marking $Opt_{\max}$ as the option with the largest $\text{COUNT}(Opt_i)$. We define sequence consistency K as

$$K(\mathcal{A}) = \frac{1}{D}\text{COUNT}(Opt_{max}) + \frac{1}{D}\sum_{Opt_i \neq Opt_{\max}} \max(0, \text{COUNT}(Opt_i) - 1). \quad (2)$$

As for accuracy, if $A_{ref} \in \mathcal{O}$ is the correct answer, the accuracy for sequence $\mathcal{A}$ can be defined as

$$\text{Acc}(\mathcal{A}) = \frac{1}{D}\text{COUNT}(A_{ref}). \quad (3)$$

In the experiments, we use the same samples as in §4.1.1 and repeat each question five times for each LLM. To compute consistency and accuracy on LFQs, we manually group the long-text generated answers into options so that answers with similar meanings are grouped together under the same option. We also explore the impact of different temperatures, which is the parameter that controls the degree of randomness of the generated text, by using values 0, 0.5, and 1.

Figure 3 shows GPT-3.5-turbo's consistency for MCQs and LFGQs across the three temperature values. Both formats tend to be consistent in their answers. Even when the temperature is increased, consistency does not decrease notably. Between the two formats, LFGQs tend to be more consistent than MCQs. We also calculate the Pearson correlation coefficient between consistency and accuracy. In the case of MCQs, the Pearson correlation coefficient is 0.32, while for LFGQs, the coefficient reaches 0.416, implying that higher consistency does not necessarily mean more correct.

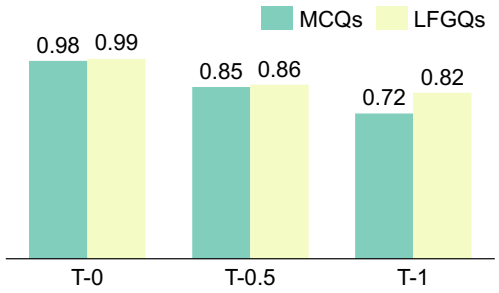

Figure 3: GPT-3.5-turbo's consistency on MCQs and LFGQs with different temperatures.

Our findings suggest that a higher level of consistency indicates a sharper probability distribution of specific knowledge, but it does not guarantee the correctness of the knowledge. Unlike SelfCheckGPT (Manakul et al., 2023), which leverages the idea that the higher the consistency, the higher the correctness, we do not find a direct relationship between consistency and accuracy. We believe this is due to the knowledge required to answer the evaluation dataset. While SelfCheckGPT is evaluated on information from famous individuals (Lebret et al., 2016), we use specialized professional medical datasets.

4.2. Token Logits

To compare MCQs and LFGQs in the token logits space, which is the space of predicted probabilities, we rely on two techniques: unified confidence calculation and expected calibration error.

4.2.1. Unified confidence calculation

One of the mainstream approaches used to analyze why LLMs select specific options when answering MCQs is through token logits (Manakul et al., 2023). GPT-3.5-turbo, for instance, can generate log probabilities for the most probable tokens associated with each output token.[2] However, while there are formulas to calculate confidence for multiple options (Jiang et al., 2021; Holtzman et al., 2021; Lin et al., 2022a), direct utilization of token probability calculations for comparing MCQs with LFGQs is not straightforward.

Since our goal is to compare MCQs and LFGQs, we follow (Jiang et al., 2021) and propose a unified confidence calculation applicable to the three QA formats: MCQs, LFGQs, and TFQs. Let us assume an input question q that makes a LLM generate the set of answers $\mathcal{A}$.

Each answer $A_i \in \mathcal{A}$ contains $|A_i|$ tokens. Each token, denoted as t_k, with $1 \leq k \leq |A_i|$, has a corresponding autoregressive token log probability

[2] https://platform.openai.com/docs/guides/gpt

$P_{log}(t_k|q, t_{<k})$. We first compute the average token log probability of each answer A_i as

$$P_{avg}(A_i|q) = \frac{\sum_{i=1}^{|A_i|} P_{log}(t_k|q, t_{<k})}{\max(1, |A_i|)}. \quad (4)$$

From the initial set of answers $\mathcal{A}$, which may contain duplicates, we consolidate them into z unique answers, denoted as $\mathcal{A}^{uni} = \{A_1^{uni}, A_2^{uni}, ..., A_z^{uni}\}$, where $z \leq D$. For each unique answer, we select the highest log probability observed for any instance of that answer in $\mathcal{A}$, denoted as $P_{log}^{highest}(A_i^{uni}|q)$. Subsequently, we rank the first W unique answers, where $W \leq z$,[3] from $\mathcal{A}^{uni}$ in descending order by their frequency and the corresponding $P_{log}^{highest}(A_i^{uni}|q)$, to filter out excessively similar responses and maintain the diversity in the unique answers. We calculate the standardized confidence for the first W answers as

$$C_N(A_w^{uni}) = \frac{e^{p_{avg}^{higest}}(A_w^{uni}|q)}{\sum_{w=1}^{W} e^{p_{avg}^{higest}}(A_w^{uni}|q)}. \quad (5)$$

Finally, For MCQs, we use regularization matching to combine first W answers with four candidate options, and get the final label (0 or 1) with the sum of corresponding standardized confidence. For LFGQs, we directly get the final label (0 or 1) according to standardized confidence for the first W answers and human labeling for each answer.

4.2.2. Expected Calibration Error

After obtaining the unified confidence, we compute model calibration (Gupta et al., 2006; Ahmed et al., 2020) to test whether a LLM exhibits good calibration across different dataset evaluation formats. A well-calibrated model should provide confidence (i.e., logit) estimates that closely match the actual probability of the correctness of the answer. Inaccurate predictions should correspond to low confidence (i.e, logit) values, whereas accurate predictions should yield high confidence (i.e., logit) values.

In practice, we employ a commonly used metric known as expected calibration error (ECE) (Niculescu-Mizil and Caruana, 2005) to assess the alignment of confidence and accuracy. ECE is computed as the weighted average of the difference between the accuracy and confidence. To measure confidence quantitatively, we divide the $[0, 1]$ interval into multiple bins. Each sample falls into one of these bins based on the model's predicted results. The average model confidence is calculated in each bin, and then compared with the average accuracy of the sample real label in the bin. The absolute value of these two differences can measure the model's confidence. A larger difference indicates lower model confidence. Formally,

[3] $W = 4$ in our experiments.

Format	CA	CARE-MI	M3KE	ARC	MATH
MCQ	4	0.426	0.317	0.492	0.281
MCQ	3	0.329	0.364	0.382	0.259
MCQ	2	0.414	0.427	0.280	0.257
LFGQ	-	0.304	-	-	-
TFQ	2	0.276	-	-	-

Table 5: GPT-3.5-turbo's ECE for different formats and number of candidate answers. CA stands for the number of candidate answers.

$$\text{ECE} = \sum_{b=1}^{\mathcal{B}} \frac{|n_b|}{N} |\text{acc}(b) - \text{conf}(b)|. \quad (6)$$

where b represents the b-th bin, $\mathcal{B}$ represents the total number of bins, n_b represents the number of samples in the b-th bin, $\text{acc}(b)$ represents the average value of the true label of the sample in the b-th bin, $\text{conf}(b)$ represents the average value of the model prediction probability in the b-th bin. In our experiments, we set $\mathcal{B} = 100$.

4.2.3. Results

Within the CARE-MI dataset, we use the three QA formats, MCQs, LFGQs, and TFQs, to compute ECE and reliability. We use confidence scores and true labels to draw reliability diagrams as in (Kängsepp et al., 2022). A reliability diagram closely aligning with the identity line suggests good model calibration, while a significant deviation indicates poor calibration. Results are shown in Figure 4 and in Table 5. LLMs operating on MCQs exhibit the poorest calibration and highest ECE compared to the other two formats. This suggests that the LLMs' predictions in MCQs are not accurately aligned with the true probability of correct answers, indicating overconfidence in their responses. Additionally, we observe that the ECE in TFQs (0.276), which contain only two candidate options, is lower than in MCQs (0.426), which have four candidate options. To investigate the impact of the number of candidate answers, we conduct experiments by varying the number of options in MCQs across the four datasets. We analyze whether the number of options and the domain of each dataset affect ECE. Throughout these experiments, we maintain the correct answer consistently positioned as the last option. As depicted in Table 5, we do not find a clear correlation between these factors and ECE, meaning that the number of options and the domain do not seem to influence LLM's performance.

4.3. Embeddings

Up to this point, our analysis reveals that the misalignment between MCQs and LFGQs answers is evident in both the direct output (§4.1) and the token logits (§4.2). Next, we investigate whether

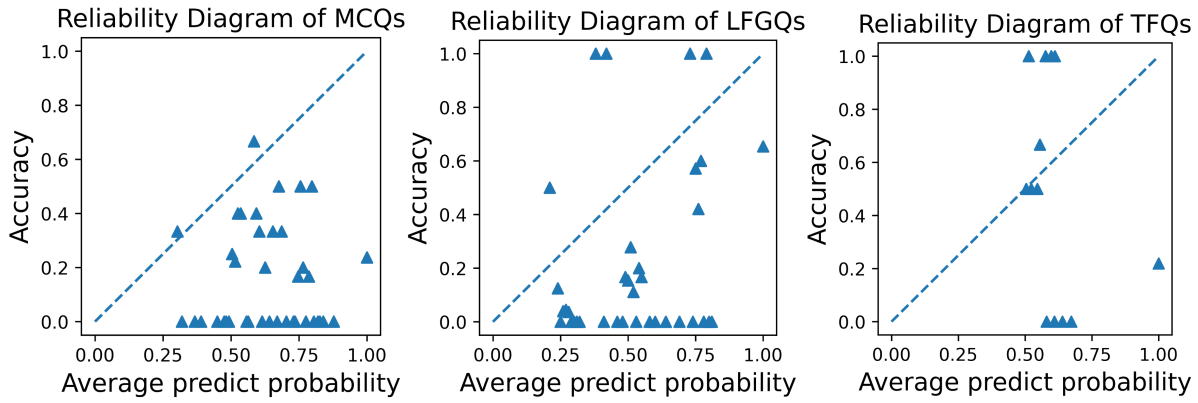

Figure 4: Reliability diagrams for MCQs, LFGQs, and TFQs on the CARE-MI dataset.

this difference is also manifested in the embedding space derived from the hidden states of the models (Burns et al., 2023). We also explore how the embeddings behave under different question formats and models. The technique proposed by Li et al. (2023b) enables the extraction of hidden outputs from the model by collecting the heads of the attention blocks, and use these heads as index to obtain the hidden outputs of each layer from the model.[4] We utilize the hidden outputs of the last token in the input. To thoroughly investigate the distinctions in the embedding space between MCQs and LFGQs themselves as much as possible, unlike the prompt design in the previous experiments, we only set a post-prompt for MCQs, and LFGQs do not contain any prompts. Finally, the hidden outputs have information on the number of input samples, the number of hidden layers, the number of attentions, and the dimensions of heads. Refer to Table 6 in the Appendix (§8) for more details. We randomly select 40 samples from the ARC dataset, each with MCQs and LFGQs formats, and plot t-SNE (Van der Maaten and Hinton, 2008) representations of the hidden embeddings in each layer. Figure 5 shows the visualizations for Llama-2-7b-chat-hf. Other model visualizations are provided in the Appendix (§8). The results show that the embeddings from MCQs and LFGQs display clear separations in some layers of the hidden states. We observe a consistent trend across the various LLMs: in the initial layers, embeddings of the two formats show clear separations. However, as we progress towards the final layers, the embeddings corresponding to MCQs and LFGQs tend to become closer. Additionally, in certain models, the embeddings are distinctly separated in specific middle layers. For instance, in the open-llama-7b model, the embeddings exhibit clear differentiation in the 14th layer. Finally, the representation of embeddings from the same model but different sizes

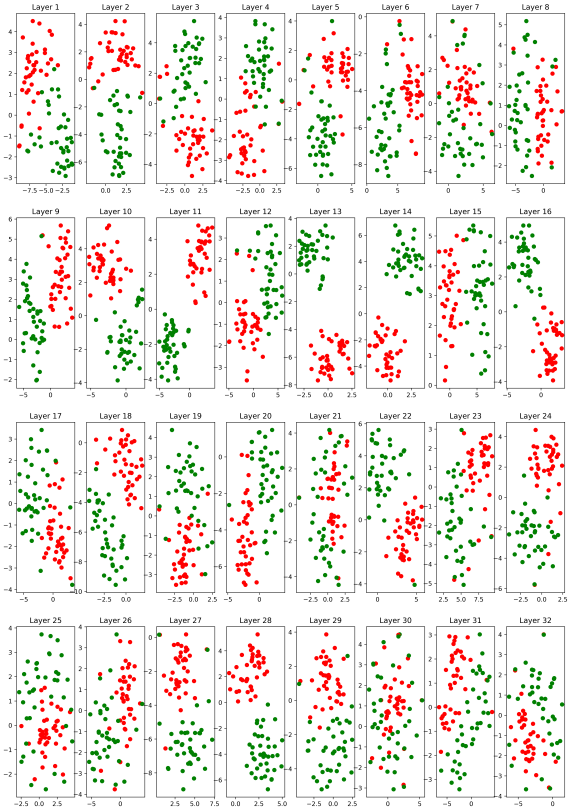

Figure 5: t-SNE visualization for each layer in Llama-2-7b-chat-hf. MCQs in red, LFGQs in green.

can vary, as shown in the embeddings of Dolly-v2-3b and Dolly-v2-7b in Figures 9 and 10 in the Appendix.

4.4. Different QA formats produce different answers

Our experiments showed that different formats of a single question may result in different performances. Moreover, our results challenged the notion that greater consistency leads to higher accu-

[4] https://github.com/davidbau/baukit

racy by closely examining the relationship between them. When comparing MCQs and LFGQs in expected calibration error, prompts from MCQs were the most overconfident in their predictions. Finally, in the embedding space, MCQs and LFGQs representations were clearly separated in some layers of the hidden states.

5. Related Work

5.1. QA Benchmarks

QA is a prevalent evaluation method in natural language processing tasks. With the surge of LLMs, several QA evaluation benchmarks have emerged to asses models' reasoning and fact-retrieval skills (Liang et al., 2022; Chang et al., 2023; Li et al., 2023a; Chia et al., 2023). These QA benchmarks encompass divers dataset formats, including multiple-choice questions (MCQs) (Bhakthavatsalam et al., 2021; Ramamurthy and Aakur, 2022; Liu et al., 2023a; Huang et al., 2023), long-form generation questions (LFGQs) (Zhang et al., 2018; Lin et al., 2022b; Xiang et al., 2023) and True/False questions (TFQs) (Singhal et al., 2023). Many existing QA evaluation benchmarks use relatively simple MCQs formats, in which models can strongly rely to formulate their answers. In addition, previous work has primarily focused on evaluations of MCQs, not considering comparisons between the different formats (Jiang et al., 2021; Lin et al., 2022a; Robinson and Wingate, 2023). In this paper, we focused on conducting a comprehensive comparative analysis between MCQs and LFGQs, thereby enhancing the understanding of the drawbacks and limitations of the different evaluation methods.

5.2. LLMs and Multiple-Choice Questions

Previous work has underscored the sensitivity of LLMs to prompting strategies (Zhao et al., 2021; Singhal et al., 2023) and positional bias (Wang et al., 2023), which pose challenges to model assessment. For instance, (Zheng et al., 2023) showed that GPT-4 tends to favor the candidate answer presented in the first position, leading to unfair evaluation results. Additionally, (Pezeshkpour and Hruschka, 2023) observed that GPT-4 and InstructGPT(Ouyang et al., 2022) perform differently when answer options are rearranged on various benchmarks. We expand upon prior work, which focused on a limited number of models and scenarios, to study and identify general patterns and analyze their underlying causes across diverse datasets and models.

6. Discussion and Conclusion

This paper focused on testing the effectiveness of MCQs evaluating LLMs. Motivated by the observation of consistent preference biases across different datasets with several LLMs, we first conducted a significance test to determine the position of the candidate answers affect LLMs' predictions, resulting in accuracy instability. More specifically, we analyzed how different LLMs have different positional preference patterns on the same dataset, while the preference positional patterns of a particular LLM remained constant across datasets from different sources. In addition, we conducted comparative experiments between MCQs and LFGs in three different spaces to ascertain the advantages and disadvantages of each as evaluation benchmarks.

Recommendations Based on our experiments, we offer a few suggestions for utilizing MCQs and LFGQs formats in LLM evaluation benchmarks:

1. The choice of QA format should be aligned with the type of knowledge being evaluated. Whereas it may be fine to use MCQs for testing general knowledge, in some professional domains—particularly those carrying legal responsibilities, such as the medical field, it is advisable to use LFGQs under human supervision to ensure a more rigorous evaluation.

2. When using MCQs for evaluating LLM, adjusting the number of options, whether decreasing or increasing them, does not necessarily enhance accuracy and confidence. However, regarding order sensitivity, reordering candidate answers for each question and repeating questions can enhance the robustness of the assessment process.

3. Our findings do not indicate a strong correlation between consistency and accuracy in LLMs responses. Therefore, we do not recommend relying on consistency as a tool to enhance performance in LLMs.

4. Given the discrepancy we found between MCQs and LFGQs results, we believe that LFGQs is the best format for evaluating LLM, as it aligns well with real-world use cases. We recommend prioritizing LFGQs format and evaluating LLM from various perspectives, including correctness, completeness, relevance, and interpretability.

We hope that the results presented in the paper and the investigation about order sensitivity and comparative analyses between MCQs and LFQs can inspire future research to improve evaluation benchmarks for LLMs.

Acknowledgment

This work was partly supported by JSPS KAKENHI No.JP22K12091 and the State Key Program of National Nature Science Foundation of China No.61936001.

7. Reference

Mukhtar Ahmed, Shakeel Ahmad, Muhammad Ali Raza, Uttam Kumar, Muhammad Ansar, Ghulam Abbas Shah, David Parsons, Gerrit Hoogenboom, Taru Palosuo, and Sabine Seidel. 2020. Models calibration and evaluation. *Systems modeling*, pages 151–178.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

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.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: towards holistic evaluation of instruction-tuned large language models. *CoRR*, abs/2306.04757.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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. *CoRR*, abs/1803.05457.

Together Computer. 2023. Redpajama-incite-instruct-3b-v1.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit

hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 345–363. Association for Computational Linguistics.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858.

Daniel Gao, Yantao Jia, Lei Li, Chengzhen Fu, Zhicheng Dou, Hao Jiang, Xinyu Zhang, Lei Chen, and Zhao Cao. 2022. KMIR: A benchmark for evaluating knowledge memorization, identification and reasoning abilities of language models. *CoRR*, abs/2202.13529.

Hoshin V Gupta, Keith J Beven, and Thorsten Wagener. 2006. Model calibration and uncertainty estimation. *Encyclopedia of hydrological sciences*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know *When* language models know? on the calibration of language models for question answering. *Trans. Assoc. Comput. Linguistics*, 9:962–977.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Markus Kängsepp, Kaspar Valk, and Meelis Kull. 2022. On the usefulness of the fit-on-the-test view on evaluating calibration of classifiers. *CoRR*, abs/2203.08958.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1203–1213. The Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A chinese multi-choice biomedical question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8862–8874. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *CoRR*, abs/2306.03341.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023a. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

Yugeng Liu, Tianshuo Cong, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2023b. Robustness over time: Understanding adversarial examples' effectiveness on longitudinal versions of large language models. *CoRR*, abs/2308.07847.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM.

Nuance. 2023. Automatically document care with the Dragon Ambient eXperience.

OpenAI. 2023a. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023b. Introducing ChatGPT and Whisper APIs.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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

Priyadharsini Ramamurthy and Sathyanarayanan N. Aakur. 2022. ISD-QA: iterative distillation of commonsense knowledge from general language models for unsupervised question answering. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 1229–1235. IEEE.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4454–4470. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Stability-AI. 2023. [link].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. 2023. CARE-MI: chinese benchmark for misinformation evaluation in maternity and infant care. *CoRR*, abs/2307.01458.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2950–2968. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li,

Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings. In *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, pages 159–168. ACM.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

8. Appendix

In this appendix, we report the visualization of the t-SNE projected embeddings in the other seven models (Figures 6-13) and the hidden embedding space details for each LLMs (Table 6).

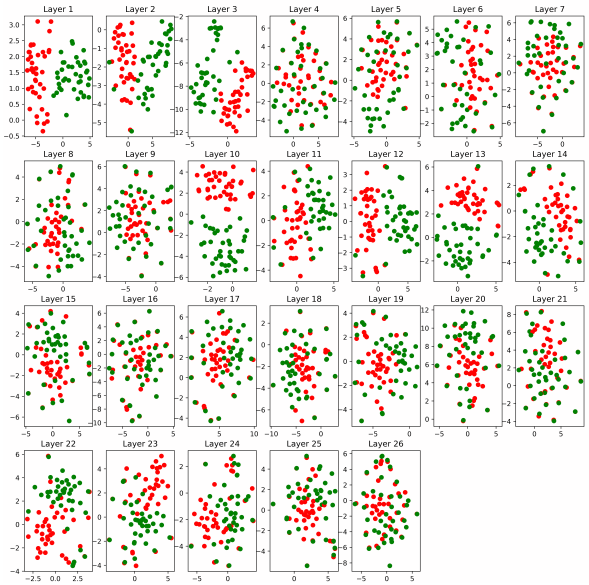

Figure 6: The visualization of t-SNE for each layer in the model Open-llama-3b. The red samples are MCQs, and the samples in green are LFGQs.

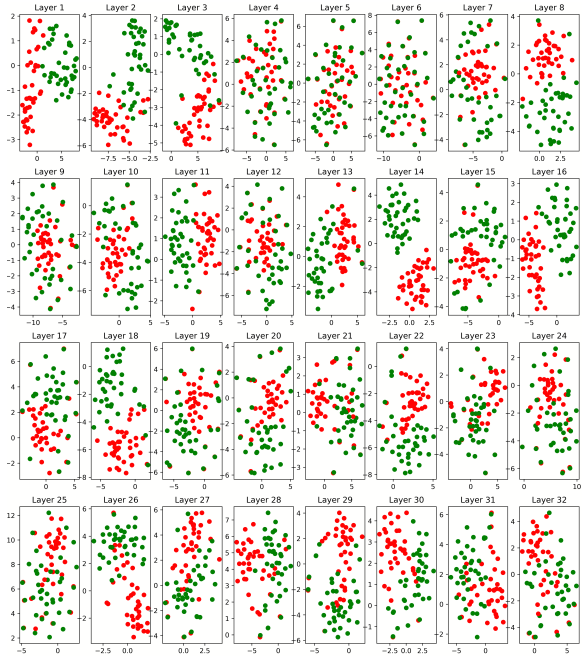

Figure 7: The visualization of t-SNE for each layer in the model Open-llama-7b. The red samples are MCQs, and the samples in green are LFGQs.

Model	hidden layers	attention heads	head dim.
open-llama-3b	26	32	100
open-llama-7b	32	32	128
vicuna-7b-v1.3	32	32	128
dolly-v2-3b	32	32	80
dolly-v2-7b	32	32	128
Llama-2-7b-chat-hf	32	32	128
RedPajama-INCITE-Instruct-3B-v1	32	32	80
stablelm-tuned-α-3b	16	32	128
stablelm-tuned-α-7b	16	32	192

Table 6: Number of hidden layers, number of attention heads, and the head dimensionality of the LLMs.

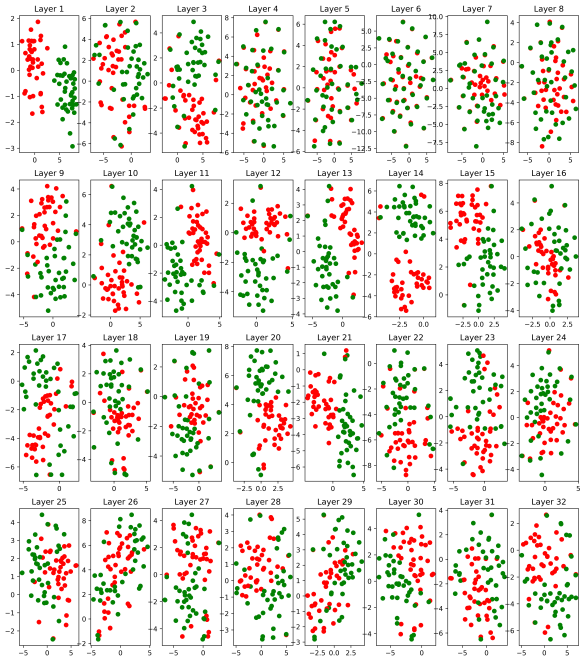

Figure 8: The visualization of t-SNE for each layer in the model Vicuna-7b-v1.3. The red samples are MCQs, and the samples in green are LFGQs.

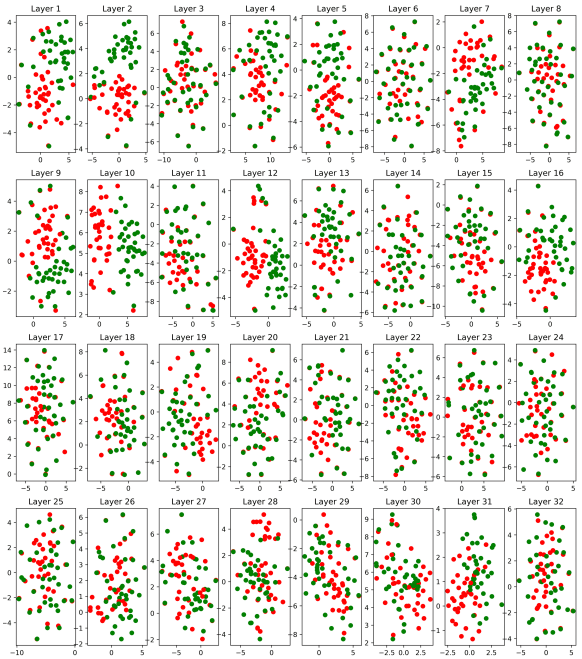

Figure 9: The visualization of t-SNE for each layer in the model Dolly-v2-3b. The red samples are MCQs, and the samples in green are LFGQs.

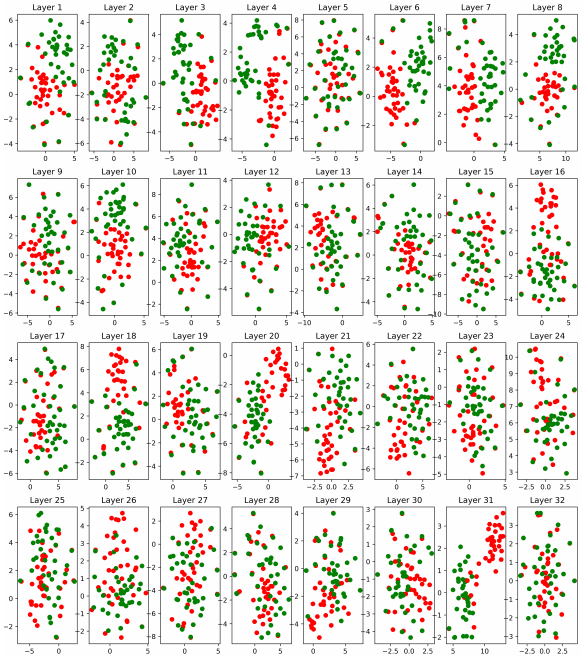

Figure 10: The visualization of t-SNE for each layer in the model Dolly-v2-7b. The red samples are MCQs, and the samples in green are LFGQs.

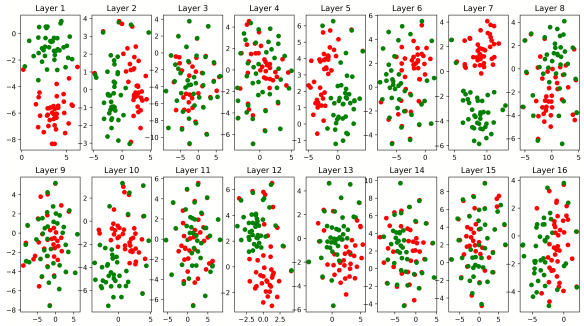

Figure 11: The visualization of t-SNE for each layer in the model Stablelm-tuned-alpha-3b. The red samples are MCQs, and the samples in green are LFGQs.

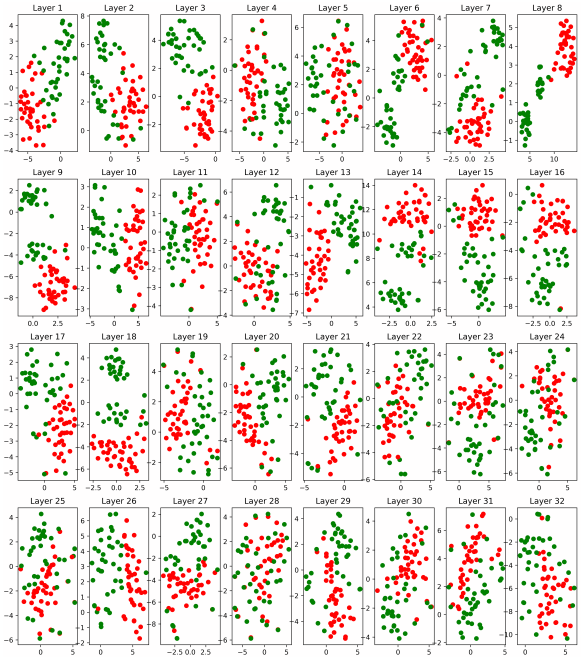

Figure 12: The visualization of t-SNE for each layer in the model RedPajama-INCITE-Instruct-3B. The red samples are MCQs, and the samples in green are LFGQs.

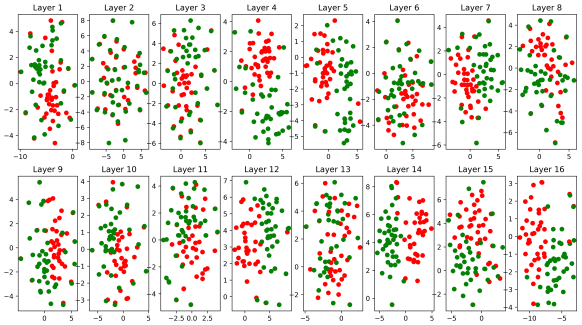

Figure 13: The visualization of t-SNE for each layer in the model Stablelm-tuned-alpha-7b. The red samples are MCQs, and the samples in green are LFGQs.