Journal of Electronic Research and Application, 2026, Volume 10, Issue 5
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Answer Distribution Bias in OmniBench: How Answer-Position Skew Affects Multimodal Large Language Model Evaluation

Sai Wan*
School of Electronic and Information Engineering, Liaoning Technical University (Huludao Campus), Huludao 125105, Liaoning, China

***Corresponding author:** Sai Wan, 15242903593@163.com

Abstract: OmniBench is a widely used tri-modal (image–audio–text) benchmark containing 1,142 four-choice multiple-choice questions. We discover a severe answer-position skew in OmniBench: option D is correct 48.6% of the time ($\chi^2 = 384.34$, $p = 5.46\times10^{-83}$), nearly twice the expected 25%. To test whether this skew distorts evaluation outcomes, we design an option-shuffling experiment: keeping all question content unchanged, we randomly reassign letter labels so that the correct answer is uniformly distributed (D ≈ 25%), then re-evaluate the same models. Results show that accuracy changes significantly in two of three tested models after shuffling (up to 4.20%, $p < 0.01$), demonstrating that unequal answer distribution can significantly bias model evaluation outcomes. Furthermore, we propose a label-free content-scoring evaluation method based on conditional log-probability, which achieves distribution-invariant evaluation (accuracy difference ≤ 0.18%, $p > 0.4$).

Keywords: Multimodal large language models; Benchmark evaluation; Answer distribution bias; Position bias; OmniBench

Online publication: Jun 29, 2026

1. Introduction

Large language models (LLMs) have achieved remarkable progress across a wide range of tasks, and evaluation benchmarks have played a central role in measuring this progress. OmniBench is one of the few benchmarks that simultaneously covers image, audio, and text modalities, and has been widely adopted for evaluating multimodal large language models (MLLMs) [1]. Multimodal benchmarks such as MMMU, MMBench, and MME have established standard evaluation protocols for image-text models, while AudioBench addresses audio understanding [2–5]. OmniBench extends these efforts to the more challenging tri-modal setting by requiring simultaneous comprehension of image, audio, and text [1].

Multiple-choice questions (MCQs) are the dominant format in current LLM evaluation, yet this

format carries a known risk: if the distribution of correct answer positions is non-uniform, it can introduce systematic bias into evaluation results. Pezeshkpour and Hruschka showed that GPT-4 performance can degrade by up to 53% due solely to changes in the positional placement of MCQ options [6]. Zheng *et al.* further demonstrated that LLMs exhibit intrinsic token-level selection bias toward specific option identifiers (A/B/C/D), and proposed PriDe, a prior-debiasing method, as a remedy [7]. Prior work on calibration has also shown that language models carry systematic prior biases toward specific answer choices [8,9]. However, no systematic analysis of answer-position bias has been conducted for tri-modal benchmarks.

This paper makes two contributions: (1) we quantify the answer distribution in OmniBench and demonstrate it is severely skewed; and (2) through a controlled option-shuffling experiment, we directly test whether this skew affects model evaluation outcomes. As a remedy, we propose and validate a label-free content-scoring evaluation method that is robust to answer distribution changes.

2. Answer distribution in OmniBench

OmniBench contains 1,142 four-choice multiple-choice questions. **Table 1** shows the distribution of correct answer positions across all questions. Option D is correct 48.6% of the time, nearly twice the uniform expectation of 25%. The deviation from a uniform distribution is highly significant ($\chi^2 = 384.34$, $p = 5.46\times10^{-83}$). **Figure 1** visualizes this skew.

Table 1. Correct answer distribution in OmniBench (N = 1,142)

Option	Count	Percentage	Deviation from Uniform
A	103	9.0%	−16.0%
B	240	21.0%	−4.0%
C	244	21.4%	−3.6%
D	555	48.6%	+23.6%
Total	1142	100%	$\chi^2 = 384.34$, $p = 5.46\times10^{-83}$

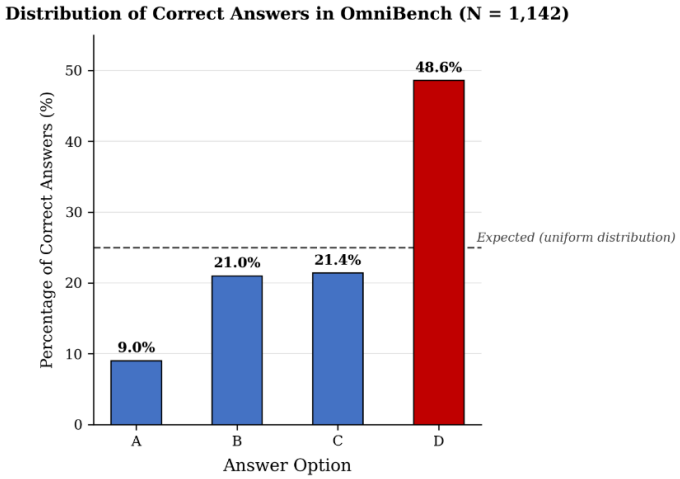

Figure 1. Distribution of correct answers in OmniBench. Option D accounts for 48.6% of correct answers, far exceeding the uniform expectation of 25% (dashed line).

3. Option-shuffling experiment

3.1. Experimental setup

The experiment contains two evaluation conditions as follows:

(1) Original (control): The original OmniBench option ordering is used, with the correct answer at position D 48.6% of the time;

(2) Shuffled (treatment): For each question, the option texts remain unchanged but letter labels are randomly reassigned. After shuffling, D = 24.8% ($\chi^2 = 3.48$, $p = 0.324$), which is not significantly different from a uniform distribution. The only variable is the assignment of letter labels.

All question text, option content, and questions themselves are left unmodified. Experiments are conducted on three Qwen2.5-series models, all evaluated in text-only mode (replacing original image/audio modalities with textual descriptions). Hardware: NVIDIA RTX 4090 (24 GB). All model inference uses greedy decoding (do_sample=False), making outputs fully deterministic, the same input always produces the same output. Therefore, the original evaluation does not need to be repeated; 46.50% is a deterministic result, not a statistical mean.

3.2. Results on Qwen2.5-Omni-7B

After shuffling, accuracy increases significantly by 3.59% ($p = 0.0038$, **). The change cannot be attributed to chance (**Table 2**).

Table 2. Qwen2.5-Omni-7B option-shuffling results (seed = 42)

Condition	D Ratio	Accuracy	Correct/Total	Change
Original	48.6%	46.50%	531/1142	—
Shuffled	24.8%	50.09%	572/1142	+3.59%
McNemar test (Edwards correction): b = 116, c = 75, $\chi^2 = 8.377$, $p = 0.0038$ (**)				

3.3. Robustness check: 10 random seeds

To rule out the possibility that a single shuffling arrangement produces an anomalous result, we repeat the experiment with 10 random seeds s ∈ {0,1,2,3,4,5,6,7,8,42} on Qwen2.5-Omni-7B.

Shuffled accuracy exceeds the original under all 10 seeds. The 95% confidence interval [47.42%, 48.69%] lies entirely above the original accuracy of 46.50%, confirming the finding is highly robust.

3.4. Comparison across three models

Two of the three models show statistically significant accuracy changes after shuffling ($\boldsymbol{p} < 0.01$), with opposite directions. In the original benchmark the three models achieve similar accuracy (45.10%–46.50%); after equalizing the distribution the gap widens to 8.23 percentage points (41.86%–50.09%), revealing a larger observed performance difference that the original skew had masked (Table 3).

The opposite directions of change are consistent with a straightforward hypothesis: a model whose intrinsic rate of selecting option D is lower than the dataset's 48.6% D-ratio is systematically disadvantaged by the original distribution and therefore gains accuracy when the distribution is equalized; conversely, a model that over-selects D profits from the skew and loses accuracy when it is removed. This pattern is consistent with the observed gains for Qwen2.5-Omni-7B (+3.59%) and losses for Qwen2.5-3B (−4.20%) after shuffling (Table 4). Similar directional heterogeneity under position-bias correction has been reported

for text-only benchmarks by Zheng *et al.* [3].

Table 3. Qwen2.5-Omni-7B results across 10 random shuffling seeds (original baseline accuracy: 46.50%).

Seed	Shuffled Acc.	Change	Direction
0	46.67%	+0.18%	↑
1	47.55%	+1.05%	↑
2	47.81%	+1.31%	↑
3	48.07%	+1.58%	↑
4	47.90%	+1.40%	↑
5	48.07%	+1.58%	↑
6	48.60%	+2.10%	↑
7	48.34%	+1.84%	↑
8	47.46%	+0.96%	↑
42	50.09%	+3.59%	↑
Mean ± SD	48.06% ± 0.89%	+1.56%	10/10 ↑
95% CI	[47.42%, 48.69%]	fully above original 46.50%	
One-sample t-test (H1: shuffled mean > original): t = 5.544, p = 0.0002 (***)			

Table 4. Three-model option-shuffling comparison (seed = 42). **: $p < 0.01$; ns: $p \geq 0.05$ (McNemar test, Edwards correction)

Model	Orig. Acc.	Shuf. Acc.	Change	McNemar p	Result
Qwen2.5-Omni-7B	46.50%	50.09%	+3.59%	0.0038 **	Significant increase
Qwen2.5-VL-7B	45.10%	46.15%	+1.05%	0.3961 ns	Not significant
Qwen2.5-3B	46.06%	41.86%	−4.20%	0.0050 **	Significant decrease

4. Content-scoring interface

4.1. Method

The option-shuffling results show that letter-label assignment affects evaluation outcomes. To eliminate this dependency, we propose a label-free content-scoring evaluation method (referred to hereafter as a content-scoring interface for brevity). Rather than prompting the model to output a letter (A/B/C/D), we compute the conditional log-probability of each option's text given the question context (Equation 1), a principle related to earlier calibration approaches [9,10]:

$$\text{score}(\text{option}_i) = \log P(\text{option\_text}_i \mid \text{question\_context}) \quad (1)$$

The option with the highest score is selected as the answer. No letter labels appear anywhere in the prompt. The context template used is:

[Image description]: {image_description}
[Audio description]: {audio_description}

Question: {question}
The answer is:

Since letter labels are absent entirely, the scoring result is structurally independent of how option

letters are assigned; changing the distribution of correct answer labels cannot affect evaluation outcomes by construction.

4.2. Results

With the content-scoring method, changing the option distribution from D = 48.6% to D ≈ 25% causes no significant accuracy change for either model ($p > 0.4$). This contrasts sharply with the letter-based evaluation, where Qwen2.5-3B showed a 4.20% change ($p = 0.005$, **). The method successfully eliminates answer-position bias from evaluation (**Table 5**).

Note that the ≈ 30% content-scoring accuracy is not directly comparable to the ≈ 46% letter-based accuracy, as the two protocols differ structurally; the relevant comparison is within-method stability under distribution change. The 30.65% accuracy still exceeds the 25% chance level, consistent with findings in MMMU-Pro that stricter protocols produce lower but more informative accuracy values [10].

Table 5. Content-scoring interface results vs. option distribution change (seed = 42)

Model	Orig. Acc.	Shuf. Acc.	Change	McNemar *p*	Result
Qwen2.5-3B	30.65%	30.65%	0.00%	0.4795 ns	No change
Qwen2.5-VL-7B	30.65%	30.82%	+0.18%	0.6171 ns	Negligible

5. Conclusion

This paper demonstrates that OmniBench contains a severe answer-position skew (D = 48.6%), and that this skew significantly biases model evaluation results. Through a controlled option-shuffling experiment, we show that merely redistributing letter labels, without changing any question content, causes statistically significant accuracy changes in two of three tested models (up to 4.20%, $p < 0.01$), with opposite directions for different models. The original benchmark masks a real performance gap of 8.23 percentage points between models. We further propose a label-free content-scoring interface that evaluates option quality through conditional log-probability, bypassing letter labels entirely. This interface is validated to be distribution-invariant: accuracy differences after shuffling are ≤ 0.18% ($p > 0.4$). We recommend: (1) benchmark designers should ensure uniform correct-answer distributions; (2) users of OmniBench evaluation results should account for the potential influence of answer distribution skew; (3) content-scoring interfaces provide a principled alternative to letter-based MCQ evaluation.

Disclosure statement

The author declares no conflict of interest.

References

[1] Li Y, Wu B, Zhao F, et al., 2024, OmniBench: Towards the Future of Universal Omni-Language Models, arXiv preprint arXiv:2409.15272.

[2] Yue X, Ni Y, Zhang K, et al., 2023, MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, Proceedings of CVPR 2024. arXiv:2311.16502.

[3] Liu Y, Duan H, Zhang Y, et al., 2023, MMBench: Is Your Multi-modal Model an All-around Player? Proceedings of ECCV 2024. arXiv:2307.06281.

[4] Fu C, Chen P, Shen Y, et al., 2023, MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, arXiv preprint arXiv:2306.13394.

[5] Wang B, Zou X, Lin G, et al., 2024, AudioBench: A Universal Benchmark for Audio Large Language Models, arXiv preprint arXiv:2406.16020.

[6] Pezeshkpour P, Hruschka E, 2023, Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions, Findings of NAACL-HLT 2024. arXiv:2308.11483.

[7] Zheng C, Zhou H, Meng F, et al., 2023, Large Language Models Are Not Robust Multiple Choice Selectors, Proceedings of ICLR 2024 (Spotlight). arXiv:2309.03882

[8] Zhao Z, Wallace E, Feng S, et al., 2021, Calibrate Before Use: Improving Few-Shot Performance of Language Models, Proceedings of ICML 2021. arXiv:2102.09690.

[9] Zhou H, Wan X, Proleev L, et al., 2023, Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering, Proceedings of ICLR 2024. arXiv:2309.17249.

[10] Yue X, Zheng T, Zhang K, et al., 2024, MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark, arXiv preprint arXiv:2409.02813.