Choosing 'Right' from Wrong: A Closer Look at Selection Bias in Spatial Multiple-Choice Questions in Large Multimodal Models

Supplementary Material

This document is Supplementary Material to Zeno et al., *"Choosing 'Right' from Wrong: A Closer Look at Selection Bias in Spatial Multiple-Choice Questions in Large Multimodal Models"*, 2025.

Hypothesis Testing Details

For analysis, we interpret each confusion matrix as a contingency table and calculate the corresponding probability matrix by normalizing all cells to sum to 1. The probabilities correspond a multinomial joint distribution $P(X, Y)$, where X is the correct option (rows) and Y is the selected option (columns). We use the notation p_{ij} to reference element i, j in the probability matrix, where the correct option is i and the selected option is j. Let $p_{j|i} = \frac{p_{ij}}{p_{i\bullet}}$ denote the conditional probability $P(Y = j|X = i)$, where $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ is the marginal distribution for X. Likewise, let $p_{i|j} = \frac{p_{ij}}{p_{\bullet j}}$ denote the conditional probability $P(X = i|Y = j)$, where $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ is the marginal distribution for Y [1].

Selection Bias We can first verify if LMMs also suffer from selection bias by testing if the probability of selecting each option is *uniform* (H_0: $p_{\bullet 1} = \ldots = p_{\bullet J}$). In our evaluation, the proportion of times that each option is selected corresponds to the marginal distribution for Y. However, a hypothesis test on the Y marginals, as in the work of [30], would ignore any associations between X and Y. This would assume that the model is equally biased to a particular option across all alternate options.

Therefore, we perform additional tests to determine if any selection-bias found is input dependent (i.e., depends on the correct option ID). Given that we do our experiments on a dataset with balanced classes (i.e., equal ratios for correct option IDs), the LMM is biased towards a particular option regardless of the correct option if the variables X and Y are *independent* (H_0: $p_{ij} = p_{i\bullet}p_{\bullet j}$ $\forall i, j$).

"Least incorrect" answer In the datasets we use for our experiments, there is no "best" wrong answer. We instead investigate LMMs are consistent when they choose an incorrect option. If this behavior is present in LMMs, then we would expect the marginal distributions to be *homogeneous* (H_0: $p_{i\bullet} = p_{\bullet i}$ $\forall i$).

We use all possible option-order permutations in our experiments, so we would also expect the joint distribution to also be *symmetric* around the diagonal if the model selects consistently a supposed "least incorrect" option (H_0: $p_{ij} = p_{ji}$ $\forall i, j$).

Large Multimodal Model (LMM) Multiple-Choice Prompts

Shown below is an example of the text portion of a prompt—and the text response—in our main experiment. Each prompt of this type is provided to the model with an accompanying image from the What's Up benchmark [13], where there is exactly one correct answer among the multiple-choice options. As stated in the paper, there are four image configurations for a pair of objects (e.g., in the example below, an armchair and beer bottle), and each one is tested using all permutations of the options list.

Example Prompt: MCQ (A/B/C/D)

Given the image, which caption below is correct?
(A) A beer bottle on a armchair
(B) A beer bottle to the left of a armchair
(C) A beer bottle to the right of a armchair
(D) A beer bottle under a armchair

The correct answer is (

Example response: *The correct answer is (A) A beer bottle on a armchair.*

This example shows a case where the option ID symbols are A/B/C/D. In prompts with alternative symbol sets a/b/c/d and 1/2/3/4, the corresponding characters inside the parentheses in the prompt are replaced.

[1] The earlier notation of p_j in the accompanying paper corresponds to $p_{\bullet j}$ here.

Option ID Symbols	Architecture	LLM	Independence	Homogeneity	Symmetry	Uncertainty
a/b/c/d	LLaVA-1.6	Mistral-7B	5,196.7	2,221.2	1,834.9	0.174
		Vicuna-7B	4,747.8	1,334.5	1,187.8	0.155
		Vicuna-13B	3,670.1	2,380.3	1,940.4	0.128
		Llama3-8B	14,822.1	2,264.3	1,857.1	0.496
	Llama3.2-Vision	Llama3.1-11B	12,570.6	1,348.1	1,265.5	0.400
	InstructBLIP	Flan-T5 XXL	3,330.0	833.3	788.7	0.108
	Qwen2-VL	Qwen2-7B	4,587.8	373.2	422.8	0.146
	Qwen2.5-VL	Qwen2.5-3B	4,091.4	756.8	710.6	0.130
		Qwen2.5-7B	5,947.9	840.2	778.5	0.189
A/B/C/D	LLaVA-1.6	Mistral-7B	5,501.3	1,951.1	1,639.7	0.183
		Vicuna-7B	4,380.9	348.0	353.9	0.139
		Vicuna-13B	4,601.1	1,999.4	1,671.3	0.158
		Llama3-8B	14,865.7	2,027.8	1,722.4	0.501
	Llama3.2-Vision	Llama3.1-11B	13,443.0	1,122.3	1,085.8	0.426
	InstructBLIP	Flan-T5 XXL	3,691.4	3,742.1	2,746.1	0.127
	Qwen2-VL	Qwen2-7B	3,485.0	5,438.5	3,519.6	0.134
	Qwen2.5-VL	Qwen2.5-3B	3,829.8	1,291.0	1,146.7	0.123
		Qwen2.5-7B	5,593.7	515.3	491.6	0.176
1/2/3/4	LLaVA-1.6	Mistral-7B	5,107.2	607.4	575.1	0.162
		Vicuna-7B	4,331.9	971.6	912.2	0.142
		Vicuna-13B	4,179.3	926.3	853.1	0.137
		Llama3-8B	13,484.6	2,265.7	1,897.9	0.456
	Llama3.2-Vision	Llama3.1-11B	12,672.3	1,609.2	1,463.3	0.408
	InstructBLIP	Flan-T5 XXL	3,113.6	3,277.7	2,482.1	0.106
	Qwen2-VL	Qwen2-7B	4,669.2	367.4	362.0	0.148
	Qwen2.5-VL	Qwen2.5-3B	3,996.8	149.9	185.3	0.127
		Qwen2.5-7B	5,197.2	372.0	365.4	0.163

Table S1. Hypothesis tests performed on contingency matrices of MCQ experiments on the What's Up dataset. Table shows the χ^2 Statistics ($p \ll 10^{-4}$) for Independence, Marginal Homogeneity (Bhapkar's), Symmetry (McNemar-Bowker), and Uncertainty coefficient values (Thiel's U) for all option ID symbols a/b/c/d, A/B/C/D, and 1/2/3/4. Thiel's U measures the degree of association between the variables (X and Y) and it won't penalize a model for predicting the wrong class, as long as it is consistent (e.g., simply rearranging the classes). **LLaVA-1.6, Llama3-8B** and **Llama3.2-Vision, Llama3.1-11B** have a stronger association between the selected option and correct option in the prompt, as can be observed by their higher weights on the main diagonal of their confusion matrices.

Option ID Symbols	**Architecture**	**LLM**	$p_{\bullet 0}$	$p_{\bullet 1}$	$p_{\bullet 2}$	$p_{\bullet 3}$
a/b/c/d	LLaVA-1.6	Mistral-7B	0.47	0.24	0.14	0.14
		Vicuna-7B	0.22	0.35	0.32	0.11
		Vicuna-13B	0.06	0.41	0.33	0.20
		Llama3-8B	0.11	0.38	0.31	0.20
	Llama3.2-Vision	Llama3.1-11B	0.41	0.17	0.21	0.22
	InstructBLIP	Flan-T5 XXL	0.39	0.21	0.16	0.24
	Qwen2-VL	Qwen2-7B	0.31	0.17	0.28	0.23
	Qwen2.5-VL	Qwen2.5-3B	0.15	0.33	0.30	0.22
		Qwen2.5-7B	0.15	0.28	0.35	0.22
A/B/C/D	LLaVA-1.6	Mistral-7B	0.45	0.24	0.18	0.13
		Vicuna-7B	0.21	0.24	0.34	0.21
		Vicuna-13B	0.05	0.32	0.33	0.29
		Llama3-8B	0.09	0.35	0.33	0.24
	Llama3.2-Vision	Llama3.1-11B	0.38	0.17	0.22	0.23
	InstructBLIP	Flan-T5 XXL	0.58	0.14	0.16	0.12
	Qwen2-VL	Qwen2-7B	0.52	0.37	0.06	0.05
	Qwen2.5-VL	Qwen2.5-3B	0.10	0.30	0.30	0.30
		Qwen2.5-7B	0.17	0.29	0.33	0.21
1/2/3/4	LLaVA-1.6	Mistral-7B	0.36	0.26	0.21	0.17
		Vicuna-7B	0.30	0.33	0.25	0.12
		Vicuna-13B	0.13	0.37	0.28	0.22
		Llama3-8B	0.08	0.38	0.31	0.23
	Llama3.2-Vision	Llama3.1-11B	0.42	0.22	0.15	0.21
	InstructBLIP	Flan-T5 XXL	0.44	0.36	0.08	0.11
	Qwen2-VL	Qwen2-7B	0.22	0.34	0.24	0.19
	Qwen2.5-VL	Qwen2.5-3B	0.22	0.29	0.28	0.21
		Qwen2.5-7B	0.18	0.25	0.32	0.25

Table S2. Y marginal probabilities in contingency matrices of MCQ experiments on the What's Up dataset. Zero-based indices here correspond to the various option ID symbols a/b/c/d, A/B/C/D, and 1/2/3/4. On all models, the proportion of times each option label was selected was statistically nonuniform (χ^2 Test, $p \ll 10^{-4}$). Here, $p_{\bullet j}$ corresponds to earlier notation p_j used in paper for conciseness.

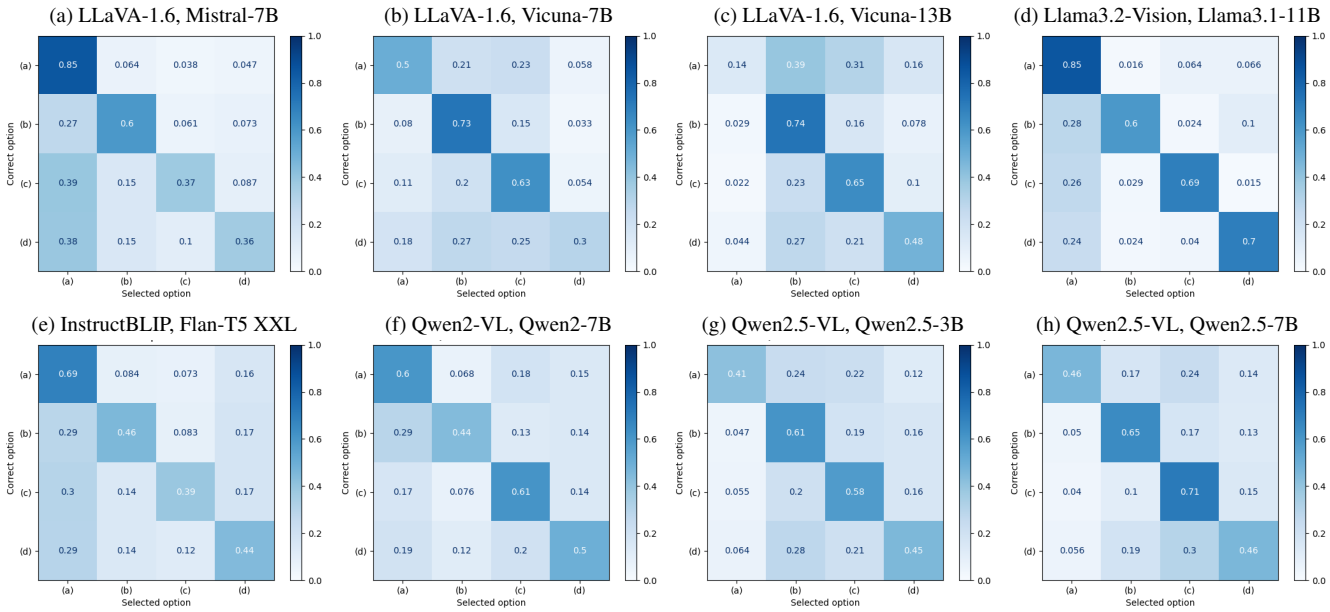

Figure S1. Confusion Matrices (row-normalized) for MCQ (a/b/c/d), What's Up dataset. The main diagonal values in the confusion matrices correspond to accuracy for each option label; since all ordering permutations are tested, the overall accuracy of each model is the average of its diagonal values. Results for option ID symbols a/b/c/d and A/B/C/D are consistent overall. In Figures S1e and S1f, **InstructBLIP, Flan-T5 XXL** and **Qwen2-VL, Qwen2-7B** are notably less biased to choose 'a' versus 'A' (as shown in Figures 3e and 3f in the main paper).

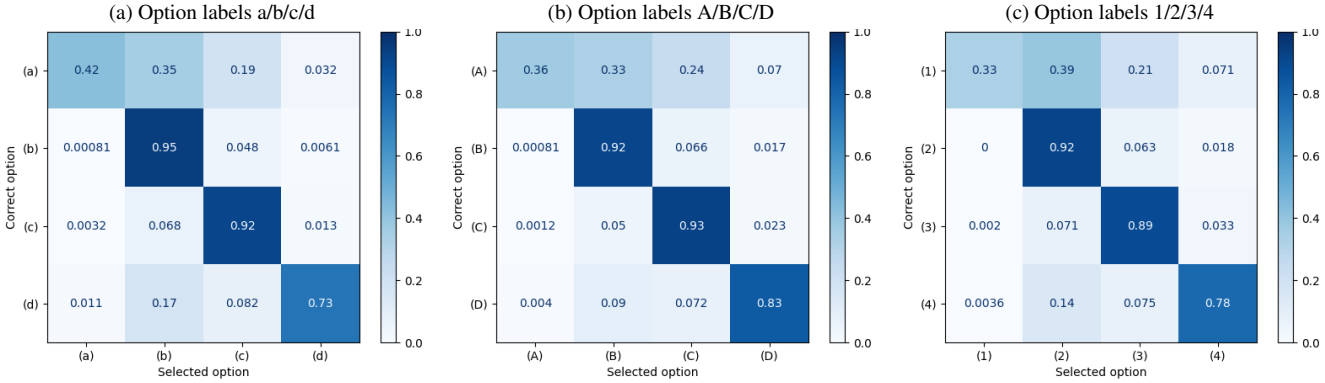

Figure S2. Confusion Matrices (row-normalized) for **LLaVA-1.6, Llama3-8B**, MCQ, What's Up dataset. Diagonal values represent accuracy for each option label. This model has very high accuracy in the main diagonal for all option labels except the first one (options 1, a, and A). In Figure S2c, we see that the model is more likely to select the wrong option when the correct choice is first (top row, option 1), choosing instead options 2 and 3. In Figures S2a and S2b, the model has slightly more probability of choosing correctly the first option (a and A) than the second and third options (b and c, B and C).

Option ID Symbols	Architecture	LLM	Accuracy
a/b/c/d	LLaVA-1.6	Mistral-7B	0.55
		Vicuna-7B	0.54
		Vicuna-13B	0.50
		Llama3-8B	**0.76**
	Llama3.2-Vision	Llama3.1-11B	*0.71*
	InstructBLIP	Flan-T5 XXL	0.49
	Qwen2-VL	Qwen2-7B	0.54
	Qwen2.5-VL	Qwen2.5-3B	0.51
		Qwen2.5-7B	0.57
A/B/C/D	LLaVA-1.6	Mistral-7B	0.56
		Vicuna-7B	0.53
		Vicuna-13B	0.53
		Llama3-8B	**0.76**
	Llama3.2-Vision	Llama3.1-11B	*0.73*
	InstructBLIP	Flan-T5 XXL	0.48
	Qwen2-VL	Qwen2-7B	0.43
	Qwen2.5-VL	Qwen2.5-3B	0.50
		Qwen2.5-7B	0.57
1/2/3/4	LLaVA-1.6	Mistral-7B	0.55
		Vicuna-7B	0.53
		Vicuna-13B	0.52
		Llama3-8B	**0.73**
	Llama3.2-Vision	Llama3.1-11B	*0.71*
	InstructBLIP	Flan-T5 XXL	0.46
	Qwen2-VL	Qwen2-7B	0.54
	Qwen2.5-VL	Qwen2.5-3B	0.52
		Qwen2.5-7B	0.56

Table S3. Model accuracy for all option ID symbols in MCQ experiments on the What's Up dataset. **LLaVA-1.6, Llama3-8B** achieves the highest accuracy on all experiments, followed by **Llama3.2-Vision, Llama3.1-11B**. Despite higher accuracy, both models still show evidence of selection bias for all option ID symbols.