Fool Your (Vision and) Language Model With Embarrassingly Simple Permutations

Yongshuo Zong[1], Tingyang Yu[2], Ruchika Chavhan[1], Bingchen Zhao[1] and Timothy Hospedales[1]

[1] University of Edinburgh [2] EPFL

@yongshuozong
yongshuo.zong@ed.ac.uk

ICML International Conference On Machine Learning

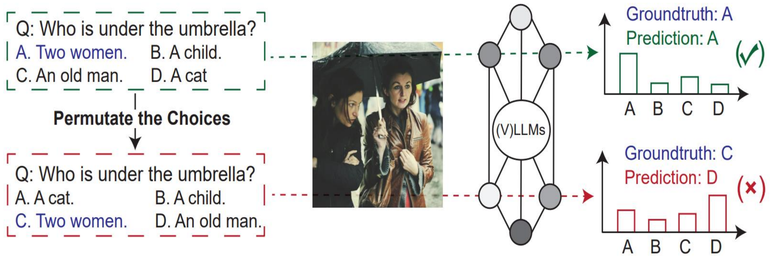
THE UNIVERSITY of EDINBURGH

EPFL

Paper & Code

Background

※ LLMs and VLLMs are increasingly advanced and widely used.
※ Multiple-choice questions are a key evaluation method.



Main Results

LLMs and VLLMs are Not Robust for MCQs

Method	MMLU	ARC-c	BoolQ	SocialIQA	MedMCQA
Llama2-7B	40.91/ 6.17 (34.74 ↓)	47.04/ 7.98 (39.06 ↓)	61.79/ 8.23 (53.56 ↓)	52.00/15.71 (36.29 ↓)	37.96/ 1.60 (36.36 ↓)
Llama2-13B	52.22/18.33 (33.89 ↓)	61.80/21.63 (40.17 ↓)	67.16/38.29 (28.87 ↓)	61.21/34.14 (27.07 ↓)	39.78/ 7.35 (32.43 ↓)
Llama2-70B	64.68/33.16 (31.52 ↓)	80.00/51.50 (28.50 ↓)	76.39/56.21 (20.18 ↓)	71.60/49.85 (21.75 ↓)	49.61/ 7.35 (32.43 ↓)
Vicuna-v1.5	48.57/18.09 (30.48 ↓)	58.37/23.43 (34.94 ↓)	64.04/29.60 (34.44 ↓)	64.99/38.33 (26.66 ↓)	39.28/ 7.67 (31.61 ↓)
Vicuna-v1.5-13B	54.68/26.27 (28.41 ↓)	69.27/38.80 (30.47 ↓)	68.96/42.14 (26.82 ↓)	66.07/44.42 (21.65 ↓)	41.80/11.90 (29.90 ↓)
WizardLM-13B	48.60/15.87 (32.73 ↓)	58.20/21.12 (37.08 ↓)	67.49/42.11 (25.38 ↓)	63.46/31.78 (31.68 ↓)	34.87/ 6.32 (28.55 ↓)
InternLM-7B	45.72/10.45 (35.27 ↓)	56.14/17.34 (38.80 ↓)	65.83/26.41 (39.42 ↓)	59.47/30.30 (29.17 ↓)	32.63/ 2.56 (30.07 ↓)
InternLM-20B	59.14/29.52 (29.62 ↓)	78.28/54.42 (23.86 ↓)	85.20/82.91 (2.29 ↓)	79.48/65.97 (13.51 ↓)	43.61/13.92 (29.69 ↓)
Falcon-7b	31.66/ 2.49 (29.17 ↓)	34.74/ 0.09 (34.65 ↓)	55.35/ 2.66 (52.69 ↓)	36.29/ 0.55 (35.74 ↓)	28.12/ 0.07 (28.05 ↓)
MPT-7B	35.60/ 3.52 (32.08 ↓)	37.76/ 1.06 (36.70 ↓)	58.46/ 7.03 (51.43 ↓)	41.61/ 2.53 (39.08 ↓)	26.31/ 1.60 (24.71 ↓)
GPT-3.5-turbo	64.81/40.39 (24.42 ↓)	82.23/61.55 (20.68 ↓)	87.92/81.35 (6.57 ↓)	70.62/56.29 (14.33 ↓)	52.22/32.07 (20.15 ↓)
Random Chance	25.0	25.0	50.0	33.33	25.0

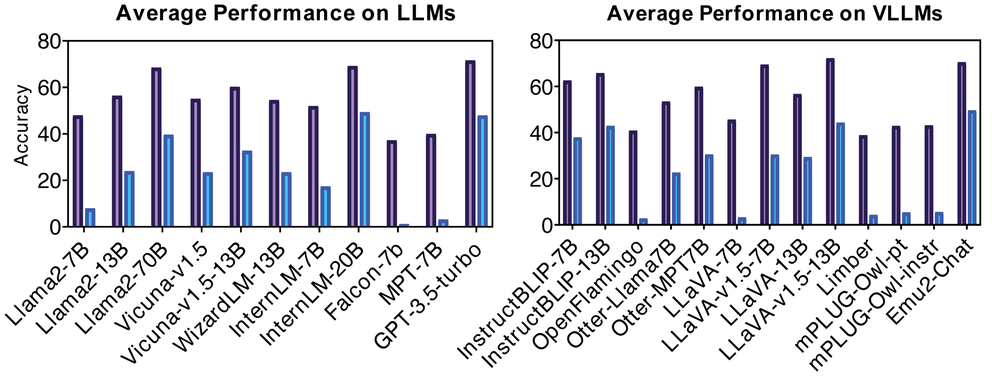
Method	ScienceQA	A-OKVQA	SEED-Bench	MMBench
InstructBLIP-7B	59.46/33.31 (26.15 ↓)	74.06/51.62 (22.44 ↓)	51.61/25.68 (25.93 ↓)	64.91/41.01 (23.90 ↓)
InstructBLIP-13B	64.15/41.84 (22.31 ↓)	77.90/55.38 (22.52 ↓)	53.65/28.79 (24.86 ↓)	67.12/45.49 (21.63 ↓)
OpenFlamingo	39.43/1.37 (38.06 ↓)	46.90/3.58 (43.32 ↓)	37.99/0.87 (37.12 ↓)	38.99/5.18 (33.81 ↓)
Otter-Llama7B	59.92/32.54 (27.38 ↓)	57.99/28.30 (29.69 ↓)	40.77/9.91 (30.86 ↓)	55.24/19.67 (35.57 ↓)
Otter-MPT7B	63.11/31.38 (31.73 ↓)	68.21/43.19 (25.02 ↓)	46.76/10.82 (35.94 ↓)	61.31/36.46 (24.85 ↓)
LLaVA-7B	45.20/2.28 (42.92 ↓)	52.91/ 0.09 (52.82 ↓)	38.36/5.67 (43.03 ↓)	46.03/5.07 (40.96 ↓)
LLaVA-13B	60.63/46.53 (14.10 ↓)	63.14/25.85 (37.29 ↓)	44.00/13.68 (30.32 ↓)	59.13/31.30 (27.83 ↓)
LLaVA-v1.5-7B	67.78/45.61 (22.17 ↓)	81.31/36.24 (45.07 ↓)	59.17/12.44 (46.73 ↓)	69.57/27.39 (42.18 ↓)
LLaVA-v1.5-13B	71.60/52.55 (19.05 ↓)	83.32/47.25 (36.07 ↓)	61.50/18.95 (42.55 ↓)	72.33/58.42 (13.91 ↓)
Limber	49.33/14.03 (35.30 ↓)	39.57/1.22 (38.35 ↓)	31.50/0.26 (31.24 ↓)	34.93/1.62 (33.31 ↓)
mPLUG-Owl-pt	53.24/10.20 (43.04 ↓)	39.91/1.83 (38.08 ↓)	35.57/0.91 (34.66 ↓)	42.57/8.54 (34.03 ↓)
mPLUG-Owl-instr	54.87/11.43 (43.44 ↓)	37.12/2.01 (35.11 ↓)	36.74/2.72 (34.02 ↓)	43.74/6.12 (37.62 ↓)
Emu2-Chat	64.60/44.27 (20.33↓)	81.91/63.67 (18.24↓)	62.11/38.02 (24.09↓)	73.21/52.44 (20.77↓)
Random Chance	Min 20.0	25.0	25.0	25.0

Performance of (V)LLMs before and after adversarial attack. Red shading indicates experiments where the permutation attack reduced performance below chance level.

TL;DR: Answer Set Permutation Breaks (V)LLMs For MCQs

Attack strategy:
- Given a question q and an answer list $A = \{a_1, a_2, \ldots, a_k\}$
- We maximize the loss $\mathcal{L}$ w.r.t. all possible permutation Π

$$\text{Maximize:} \quad \mathcal{L}\left(\text{prompt}(q, A^*)\right)$$
$$\text{s.t.} \quad A^* \in \Pi(A)$$



Summary of MCQA adversarial attack results for LLMs and VLLMs, with average accuracy across all datasets.

No Straightforward Approach Can Mitigate the Vulnerability

Method	Original	Permutation Attack	Majority Vote	C-Calibration	M-Confidence
Llama2-7B	40.91	6.17 (34.74 ↓)	33.64 (7.27 ↓)	5.24 (35.67 ↓)	22.62 (18.29 ↓)
Llama2-13B	52.22	18.33 (33.89 ↓)	48.53 (3.69 ↓)	20.02 (32.20 ↓)	50.83 (1.39 ↓)
Llama2-70B	64.68	33.16 (31.52 ↓)	65.37 (0.69 ↑)	35.77 (28.91 ↓)	64.20 (0.48 ↓)
Vicuna-v1.5-7B	48.57	18.09 (30.48 ↓)	44.10 (4.47 ↓)	11.33 (37.24 ↓)	38.29 (10.28 ↓)
Vicuna-v1.5-13B	54.68	26.27 (28.41 ↓)	52.03 (2.65 ↓)	18.10 (36.58 ↓)	55.58 (0.90 ↑)
WizardLM-13B	48.60	15.87 (32.73 ↓)	30.17 (18.43 ↓)	8.23 (40.37 ↓)	37.81 (11.21 ↓)
InternLM-20B	59.14	29.52 (29.62 ↓)	60.33 (1.19 ↑)	28.94 (30.20 ↓)	64.80 (5.66 ↑)
Falcon-7b	31.66	2.49 (29.17 ↓)	4.38 (27.28 ↓)	3.59 (28.07 ↓)	21.10 (10.56 ↓)
MPT-7B	35.60	3.52 (32.08 ↓)	13.80 (21.80 ↓)	6.24 (29.36 ↓)	21.42 (14.18 ↓)

Impact of post-hoc mitigation strategy against the permutation attack on the MMLU dataset. Contextual calibration fails completely. Majority vote and M-Confidence ameliorate the attack, but do not completely restore performance.

Why Does Majority Vote Fails?

Model	Performance (original/majority vote)	Easy sample Acc. (# samples)	Difficult sample Acc. (# samples)
Llama2-7B	40.91/33.64	100 (4148)	16.56 (9334)
InternLM-20B	59.14/60.33	100 (9957)	10.69 (3512)

Majority vote performance when considering the easy sample w.r.t. a certain model to be the ones where 50% of the permutations are correct. The final performance will be determined by the total number of easy and difficult samples.

Understanding Vulnerability Causes

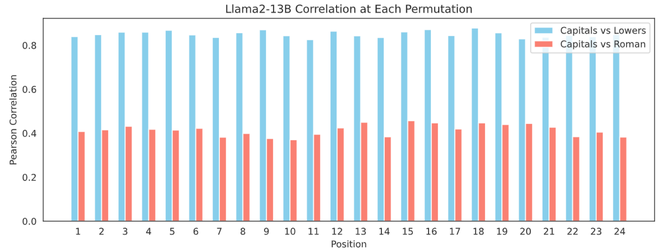
1. Position Bias and Other Attacks

Method	Original	A	B	C	D	CircularEval	Symbol Attack	Permutation Attack
Llama2-7B	40.91	60.02	37.28	30.69	35.43	27.26	25.70	**6.17**
Llama2-13B	52.22	36.15	58.69	59.08	54.91	35.80	30.76	**18.33**
Llama2-70B	64.68	63.63	64.28	67.45	62.43	48.18	47.40	**33.16**
Vicuna-7B	48.57	49.83	63.22	45.46	37.85	20.23	33.85	**18.09**
Vicuna-13B	54.68	47.33	70.00	51.73	52.04	41.42	45.40	**26.27**
WizardLM-13B	48.60	34.75	56.38	45.86	57.56	22.42	29.07	**15.87**
InternLM-7B	45.72	37.23	65.12	41.49	42.33	25.23	29.38	**10.45**
InternLM-20B	59.14	51.05	68.75	53.47	62.35	34.99	47.06	**29.52**
Falcon-7B	31.66	70.86	3.77	10.52	14.85	7.69	14.38	**2.49**
MPT-7B	35.60	0.82	75.35	34.72	2.03	2.44	21.62	**3.52**
GPT-3.5-turbo	64.81	65.84	67.77	73.81	56.55	58.21	63.99	**40.39**

Comparison of positional bias, circular evaluation, symbol attack, and our adversarial permutation on MMLU dataset. Position bias and other attacks have moderate impact compared to our adversarial permutation.

2. Symbol-Content Correlations

Symbol Set	Pearson Correlation	Original Accuracy	Permuted Accuracy
Capital Letters vs. Lowercase Letters	0.76	55.06 vs. 54.87	23.73 vs. 21.68
Capital Letters vs. Roman Numerals	0.36	55.06 vs. 52.49	23.73 vs. 19.33



The correlation of predictions indicates that the model may have learned shortcuts or spurious correlations linking option symbols with answer content.

3. Fine-tuning Helps Yet Not a Universal Solution

Fine-tuning Strategy	ARC-Challenge	MedMCQA
Zero-shot	47.04/ 7.98 (39.06↓)	37.96/ 1.60 (36.36↓)
Regular Fine-tuning	51.42/15.02 (36.40↓)	45.03/14.38 (30.65↓)
Fine-tuning with Permutation	67.64/47.73 (19.91↓)	46.78/26.07 (20.71↓)

Fine-tuning with permutation can enhance the robustness to the permutation attacks compared to the zero-shot and regular fine-tuning baseline.

Takeaways

- ❖ **Do NOT fully rely on MCQs for Evaluations.**
- ❖ **Better training strategies and/or architectures are needed.**