UNIVERSITI MALAYA

EMNLP2026 BUDAPEST

# Between Multiple Choice and Open Response: Evaluating LLMs with Option-Free Deterministic Scoring

Zhong Ken Hew  Sze Jue Yang  Chee Seng Chan

Project Page

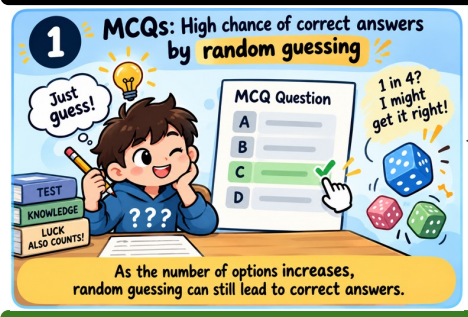
Paper

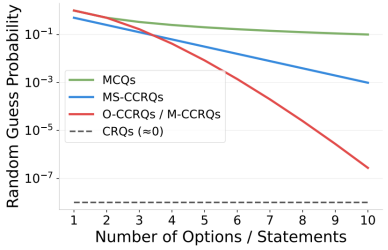
## Model Performance drop when question format change from MCQ to CSRQ

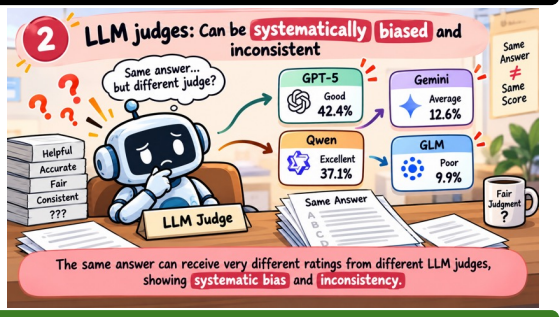
| | CS-Bench (EN) | MalayMMLU | *MyCulture* |
|---|---|---|---|
| GPT- 5 | -13.2% | -22.5% | -31.36% |
| GLM- 4.6 | -29.0% | -26.7% | -31.08% |

**A consistent drop across three datasets:** Across ***MyCulture***, MalayMMLU and CS-Bench (EN), model scores decline substantially when final answer options are removed. The MCQ-CSRQ comparison retains deterministic scoring, revealing sensitivity to answer-option exposure.

## Problem Statement & Evidence




## What is Constrained Structured Response Questions (CSRQ)?

**Multi-Select**

QUESTION
Where is venue of EMNLP 2026?
STATEMENTS
- [ ] London
- [x] Hungary
- [x] Budapest
- [ ] Paris

**Ordering**

QUESTION
What is the correct order to take a photo with your phone?
STATEMENTS
1. Point at the subject
2. Open the camera app
3. Save or view the photo
4. Press the shutter button

ANSWER 2 → 1 → 4 → 3

**Matching**

QUESTION
Match the animal to its sound.

| GROUP 1 | GROUP 2 |
|---|---|
| Dog | Moo |
| Cow | Bark |
| Cat | Meow |
| Lion | Roar |

**Constrained Structured Response Questions (CSRQs)** retain item-specific statements or descriptions but remove candidate final-answer combinations.

**CSRQs** comprise three question types:
i. subset selection (multi-select),
ii. ordering, and
iii. matching

Each format has a *lower chance-level success rate* than standard MCQs, reducing the likelihood of correct guessing while preserving fully deterministic scoring.

## *MyCulture:* Parallel Cultural Evaluation

A **Malaysia-centred** benchmark in Bahasa Malaysia. The same underlying knowledge is tested in MCQ, CSRQ and CRQ formats.

**5,463** *total question instances* | **1,821** *instances in each format* | **6** *cultural dimension*




**Purpose:** Malaysia's cultural knowledge is assessed by ***MyCulture*** in Bahasa Malaysia, and how performance shifts across question and scoring formats is examined.

**Design:** The same cultural knowledge is tested via MCQ, CSRQ, and CRQ formats, allowing comparisons across different answering and scoring methods.

**Grounded Construction:** Questions are built from Malaysian cultural sources such as JKKN and related academic publications.

**Quality Assurance:** LLM-generated questions go through automated format and source checks, with a 5% sample audited by five qualified Malaysians for consistency, relevance, and answerability.

**CSRQ**

**Question:**
Apakah kenyataan yang betul mengenai permainan Lansaran dan budaya suku kaum Murut?

**Statements:**
A. Lansaran dibina secara individu tanpa melibatkan masyarakat.
B. Permainan Lansaran memerlukan ukuran kawasan antara 8 hingga 30 kaki persegi.
C. Kayu "Selangan Batu" digunakan dalam pembinaan lantai Lansaran kerana ketahanannya.
D. Lansaran hanya untuk kanak-kanak dan tidak melibatkan orang dewasa.

**Answer:** B, C

**(Translated version)**
**Question :**
What statements are correct regarding the Lansaran game and the culture of the Murut community?

**Statements:**
A. The Lansaran is built individually without involving the community.
B. The Lansaran game requires an area size of between 8 and 30 square feet.
C. "Selangan Batu" wood is used in constructing the Lansaran floor due to its durability.
D. The Lansaran is only for children and does not involve adults.

**Answer:** B,C

*Example of MyCulture Dataset*

| Family | Model | Reasoning | MCQs | CSRQs | | CRQs | |
|---|---|---|---|---|---|---|---|
| | | | | loose | strict | loose | strict |
| | Random Baseline | | 25.53 | 40.33 | 4.28 | - | - |
| OpenAI | GPT-5 | ✓ | <u>88.14</u> | **78.10** | **56.78** | <u>68.41</u> | <u>22.90</u> |
| | GPT-5-mini | ✓ | 84.73 | 77.50 | 55.46 | 67.25 | **25.65** |
| | GPT-5-nano | ✓ | 80.12 | 73.61 | 48.38 | 63.10 | 20.76 |
| | GPT-4o | | **90.28** | <u>78.00</u> | <u>56.01</u> | **68.61** | 21.14 |
| | GPT-4o-mini | | 78.36 | 74.30 | 49.20 | 65.78 | 20.37 |
| Google | Gemini-2.5-Pro | ✓ | 85.56 | 77.78 | 55.68 | 66.87 | 20.26 |
| | Gemini-2.5-Flash | ✓ | 82.87 | 71.05 | 47.39 | 65.23 | 21.47 |
| | Gemini-2.5-Flash-Lite | ✓ | 79.57 | 72.90 | 48.00 | 62.56 | 18.95 |
| Z.ai | GLM-4.6 | ✓ | 84.46 | 76.37 | 53.38 | 63.20 | 17.74 |
| | GLM-4.6 | | 87.37 | 77.45 | 54.91 | 56.72 | 17.68 |
| | GLM-4-32B-128K | | 75.18 | 71.88 | 47.83 | 59.26 | 16.09 |
| Alibaba | Qwen3-235B-Thinking | ✓ | 61.01 | 62.62 | 42.12 | 50.13 | 18.51 |
| | Qwen3-30B-Thinking | ✓ | 79.57 | 70.49 | 43.82 | 62.48 | 17.35 |
| | Qwen3-4B-Thinking | ✓ | 71.61 | 66.42 | 37.67 | 48.47 | 10.98 |
| | Qwen3-235B-Instruct | | 80.12 | 73.55 | 51.29 | 62.93 | 18.34 |
| | Qwen3-30B-Instruct | | 82.48 | 75.21 | 52.50 | 59.03 | 17.13 |
| | Qwen3-4B-Instruct | | 76.06 | 65.64 | 35.48 | 55.97 | 15.27 |

**Performance of LLMs on the *MyCulture* benchmark across MCQ, CSRQ, and CRQ tasks.** Bold and underlined values indicate the best and second-best results, respectively.