

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

Zhong Ken Hew\* Sze Jue Yang\* Chee Seng Chan<sup>†</sup>  
Universiti Malaya, Malaysia

Project Page: <https://github.com/HewZK2000/MyCulture-CSRQ>

## Abstract

Benchmark scores are treated as evidence of LLM competence, but the same knowledge can yield different scores under different question formats. MCQs provide objective and scalable scoring, but their answer options can turn recall into recognition through answer-option scaffolding. Open-ended constructed response questions (CRQs) remove answer options, but their scores often depend on human or LLM judges and rubric interpretation, introducing judge-dependent scoring. This creates a trade-off between removing answer-option scaffolding and retaining objective scoring. We study whether this trade-off can be reduced through **Constrained Structured Response Questions (CSRQs)**, a structured-response protocol that removes final answer options while preserving deterministic scoring. CSRQs use multi-select, ordering, and matching outputs as auditable answer spaces for exact verification. We evaluate CSRQs using converted knowledge benchmarks and *MyCulture*, which is, to our knowledge, the *first Malaysia-centred cultural benchmark in Bahasa Malaysia*, with parallel MCQ, CSRQ, and CRQ formats. Across benchmarks and model families, scores change when the evaluation instrument changes, *i.e.*, scores drop after answer options are removed, and CRQ scores vary across judges. These findings suggest that benchmark scores should be interpreted as products of both model capability and the test used to measure it.

## 1 Introduction

Benchmark scores are often treated as evidence of model competence, but they are also shaped by how questions are presented and answers are scored. Multiple choice questions (MCQs) are

\*These authors contributed equally to this work; authors are listed alphabetically by last name.

<sup>†</sup>Corresponding author (cs.chan@um.edu.my).

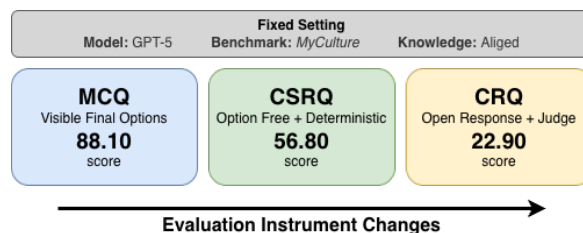


Figure 1: **Does an LLM benchmark score reflect model capability alone, or also the evaluation instrument?** MCQs expose answer choices, while open-ended CRQs introduce judge dependence. We propose CSRQ, which removes candidate final-answer combinations while retaining deterministic scoring. On *MyCulture*, our Malaysia-centred Bahasa Malaysia benchmark, the same GPT-5 and underlying knowledge yield markedly different scores across formats.

scalable and deterministically scored, yet visible answer candidates enable recognition, elimination, and chance success (Burton, 2001; Balepur et al., 2025; Raman et al., 2025; Chandak et al., 2025). Open-ended constructed response questions (CRQs) remove these candidates but introduce dependence on human or LLM judges and rubric interpretation (Livingston, 2009; Zheng et al., 2023; Bavaresco et al., 2025; Shi et al., 2025; Watts et al., 2024). Thus, benchmark scores reflect both model capability and evaluation procedure.

We investigate whether models can answer without final options while retaining deterministic scoring through **Constrained Structured Response Questions (CSRQs)**. Rather than selecting among composite candidate answers, a model constructs a verifiable, structured response, such as a subset, an ordering, or a matching. For example, instead of choosing between “**A. Statements 1 and 2**” and “**B. Statements 1 and 3**”, CSRQ presents the individual statements and requires the model to return the correct subset, that is “*1 and 2.*”

To study this under controlled conditions, we introduce *MyCulture*, a Malaysia-centred cultural benchmark in Bahasa Malaysia created as a

parallel-format testbed. Each underlying knowledge item is instantiated as an MCQ, a CSRQ, and a CRQ, holding the tested knowledge constant while varying the evaluation instrument (Figure 1). CSRQ builds on established human assessment formats, including multi-selection, ordering, and matching, using their structured output spaces as canonical answer spaces for LLM evaluation. As shown in Appendix Figure 8, CSRQs lie between MCQs and CRQs, requiring responses to be constructed rather than selected from enumerated final-answer combinations while remaining deterministically scorable.

CSRQ is not intended to prove genuine understanding, measure higher-order comprehension, or replace existing benchmarks. Instead, it isolates how scores change when final answer options are removed while deterministic scoring is retained. The MCQ-CSRQ gap measures sensitivity to answer-option removal, while variation across CRQ judges measures sensitivity to judge choice. Together, these analyses test whether benchmark scores reflect both model capability and the evaluation instrument rather than the model alone.

We evaluate this claim in two settings. First, we convert existing knowledge benchmarks into CSRQs to test whether answer-option sensitivity extends beyond the cultural testbed. Second, we evaluate *MyCulture* across parallel MCQ, CSRQ, and CRQ formats to examine how evaluation instruments shape measured cultural competence.

Our contributions are as follows:

- (i) We frame knowledge oriented LLM evaluation, with cultural competence as a central testbed, as an instrument validity problem and identify two evaluation induced illusions from answer options in MCQs and judge dependent scoring in CRQs.
- (ii) We study whether the useful functions of MCQs and CRQs can be recomposed through **Constrained Structured Response Questions** (CSRQs), an auditable protocol that elicits answers without final options and uses deterministic scoring, and provides a structured workflow for converting suitable benchmark items into CSRQs.
- (iii) We instantiate and evaluate CSRQs through converted knowledge benchmarks and *MyCulture*, a Malaysia-centred cultural testbed with parallel MCQ, CSRQ, and CRQ formats,

showing that competence varies across evaluation instruments.

## 2 Constrained Structured Response Questions (CSRQs)

### 2.1 Motivation

We formalize CSRQ around the evaluation-format tradeoff introduced above. MCQs provide objective scoring, but expose final answer options. CRQs remove final answer options, but require judge-dependent scoring. CSRQ studies whether these useful functions can be recomposed into an intermediate evaluation instrument.

**Answer-option scaffoldin.** MCQs combine the question with visible answer candidates. This design is useful for objective scoring, but it can also turn recall into recognition. A model may use answer options as hints, rely on elimination, or succeed through partial knowledge rather than independently producing the target answer. Existing MCQ benchmarks also have non-trivial random baselines, often around 25% for four-option settings and higher for some structured formats (Li et al., 2023; Hendrycks et al., 2020; Poh et al., 2024). We refer to this confound as *answer-option scaffoldin*, where the measured score partly reflects answer-option scaffolding rather than model competence alone.

**Judge-dependent scoring.** Open-ended constructed response questions (CRQs) remove final answer options and therefore support option-free elicitation. Final-answer-option-free (denoted as *option-free* hereafter) elicitation means that no candidate final answers are explicitly enumerated. CSRQs provide item-specific statements or descriptions as response primitives, but do not expose candidate subsets, orderings, or matchings from which the model selects. However, their scores depend on how responses are judged. Human grading is costly and difficult to scale, while LLM judges can vary with prompt wording, rubric interpretation, judge model, and model capability (Shi et al., 2025; Bavaresco et al., 2025). We define this as *judge-dependent scoring*, where the measured score reflects not only the evaluated response but also the scoring instrument.

These two confounds suggest that the key question is not whether MCQs or CRQs are better. The key question is whether their useful functions can be recomposed. *Can evaluation remove final an-*

swer options like CRQs, while retaining deterministic scoring like MCQs?

## 2.2 Formulation

We study this recomposition through **Constrained Structured Response Questions (CSRQs)**, an auditable structured-response protocol that combines option-free elicitation with deterministic scoring. CSRQ is not proposed as a new human-test item archetype. Multi-selection, ordering, and matching have long existed in educational and psychological assessment. Our contribution is to use these structured output spaces as canonical answer spaces for LLM evaluation.

Formally, a CSRQ is a tuple  $(Q, S, A)$ , where  $Q$  is the question prompt,  $S = \{s_1, \dots, s_M\}$  is a set of  $M$  provided statements or descriptions, and  $A$  is the ground-truth answer expressed in a canonical structure. Unlike an MCQ with a visible option set  $C = \{c_1, \dots, c_K\}$ , a CSRQ does not expose final answer candidates. The model must construct the answer  $\hat{A}$  directly from the task specification and the provided descriptions.

We instantiate CSRQs through 3 structured response variants.

- (i) **Multi Select (MS)** requires the model to output all statements in  $S$  that satisfy the condition specified in  $Q$ . Given  $S = \{s_1, s_2, s_3, s_4\}$ , a valid prediction may be  $\hat{A} = \{s_1, s_3\}$ .
- (ii) **Ordering (O)** requires the model to output an ordered sequence over  $S$  according to the relation specified in  $Q$ , such as chronology, procedure, or causal progression. Given  $S = \{s_1, s_2, s_3, s_4\}$ , a valid prediction may be  $\hat{A} = (s_2, s_4, s_1, s_3)$ .
- (iii) **Matching (M)** requires the model to construct pairwise associations between two statement sets. Given  $S_A = \{s_{a_1}, s_{a_2}, s_{a_3}\}$  and  $S_B = \{s_{b_1}, s_{b_2}, s_{b_3}\}$ , a valid prediction may be  $\hat{A} = \{(s_{a_3}, s_{b_2}), (s_{a_1}, s_{b_3}), (s_{a_2}, s_{b_1})\}$ .

## 2.3 Functional Properties

CSRQs are designed around two functional properties, which we describe as follows:

**Option-free elicitation.** CSRQs do not provide final answer candidates. The statements in  $S$  are descriptive units used to define the task, while MCQ options are candidate outputs that may directly contain the correct answer. In an MCQ, the

model selects from visible answer candidates. In a CSRQ, the model constructs a structured answer such as a subset, an ordering, or a matching.

**Deterministic scoring.** CSRQs preserve the main scoring function of MCQs without relying on visible answer options. Since the admissible output is a structured object, a model response can be canonicalized and compared against a predefined key. Let  $\hat{A}$  denote the model prediction and  $A$  denote the ground-truth answer. We define a canonicalization function  $f(\cdot)$  that normalizes equivalent surface forms. Strict scoring is defined as

$$\text{Score}(\hat{A}) = \mathbb{I} \left[ f(\hat{A}) = f(A) \right],$$

where  $\mathbb{I}[\cdot]$  returns 1 if the canonicalized prediction exactly matches the canonicalized answer, and 0 otherwise. The score is determined by the response structure, not by human preference, rubric interpretation, or judge-model sensitivity.

## 2.4 Validity Considerations

From a psychometric perspective, evaluation instruments shape the constructs and response processes measured (Messick, 1995; Kane, 2013). CSRQs strengthen construct validity by reducing construct-irrelevant variance from option order, distractor quality, and rater-dependent scoring (Rodriguez, 2003; Jonsson and Svingby, 2007). Their constrained answer spaces enable systematic control of content and difficulty, yielding more standardized and interpretable measures than CRQs. However, adequate content coverage and robustness to format variation remain necessary for CSRQs to represent the intended capability (Kane, 2013; Sireci and Faulkner-Bond, 2014).

## 2.5 Chance-Success Analysis

We characterize random-success probabilities under exact structural scoring. This analysis is deliberately narrow. It does not measure semantic difficulty, cognitive complexity, or genuine understanding. It only describes how likely a random answer is to exactly match the canonical answer under each format.

For a  $K$ -choice MCQ with one correct answer, random guessing succeeds with probability

$$P_{\text{MCQ}}(\hat{A} = A \mid Q, C) = \frac{1}{K}.$$

For a Multi Select CSRQ over  $M$  statements, the admissible answer space contains all non-

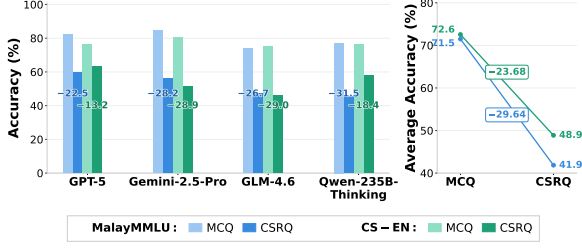


Figure 2: **Option-removal sensitivity in existing benchmarks.** **Left:** MCQ and CSRQ scores for representative models on MalayMMLU and CS-EN. **Right:** average MCQ and CSRQ scores across all evaluated models. The consistent drop after removing final answer options shows option-removal sensitivity under deterministic scoring, rather than evidence that CSRQ measures deeper understanding.

empty subsets of  $S$ . The random-success probability is

$$P_{MS}(\hat{A} = A \mid Q, S) = \frac{1}{2^M - 1}.$$

For an Ordering CSRQ, the admissible answer space contains all permutations of  $S$ . The random-success probability is

$$P_O(\hat{A} = A \mid Q, S) = \frac{1}{M!}.$$

For a Matching CSRQ with two equal-sized sets  $S_A$  and  $S_B$ , the admissible answer space contains all bijections from  $S_A$  to  $S_B$ . The random-success probability is

$$P_M(\hat{A} = A \mid Q, S_A, S_B) = \frac{1}{M!}.$$

These baselines quantify one source of option-removal sensitivity under exact scoring. They should not be interpreted as evidence that CSRQs measure deeper understanding or competency.

### 3 Empirical Validation

We evaluate whether CSRQ can serve as a feasible intermediate instrument between MCQ and CRQ in a closed-book, tool-free setting, without web search, retrieval, or use of external tools. The goal is not to prove that CSRQ measures genuine understanding, nor to claim that CSRQ replaces existing benchmarks. Instead, we test whether measured scores change when final answer options are removed while deterministic scoring is preserved, and whether open-ended CRQ scores vary under judge-dependent grading.

**Evaluated Models.** We evaluate a diverse set of closed- and open-source LLMs spanning multiple model families, scales, and reasoning modes, including OpenAI (OpenAI, 2025), Gemini (Comanici et al., 2025), GLM (Zeng et al., 2025), and Qwen3 (Yang et al., 2025). Where available, both reasoning and instruction-tuned variants are included. Models range from compact models with fewer than 10B parameters to large-scale models with more than 100B parameters. Full list is provided in Appendix D.

#### 3.1 Instrument Sensitivity in Existing Benchmarks

We first examine whether option-removal sensitivity appears beyond *MyCulture*. We use two complementary strategies. First, we extract existing multi-answer MCQ items and remove final answer options while preserving the question intent. Second, we convert representative MCQ- and CRQ-based benchmarks into CSRQ format to study how measured scores change under option-free elicitation and deterministic scoring.

##### 3.1.1 Extracting Existing Multi-Answer MCQs

We select MalayMMLU (Poh et al., 2024) and CS-Bench (Song et al., 2025) because they contain multi-answer MCQ items that can be directly reformulated as CSRQs. This setting allows a controlled comparison between MCQ and CSRQ evaluation because the underlying knowledge remains fixed while the final answer options are removed.

**Insights.** Figure 2 compares model performance under MCQ and CSRQ evaluation on the same underlying items. Across models and benchmark subsets, scores consistently decline after final answer options are removed. This pattern appears across model families and parameter scales, including models with more than 100B parameters.

This drop should not be interpreted as proof that models lack knowledge. It indicates that measured performance changes when the evaluation instrument changes. In our work, this is option-removal sensitivity. MCQs expose final answer candidates, while CSRQs require models to construct the answer from a larger structured answer space. The observed gap therefore provides supporting evidence that MCQ scores may partly benefit from answer-option scaffolding. Full results are reported in Appendix E.

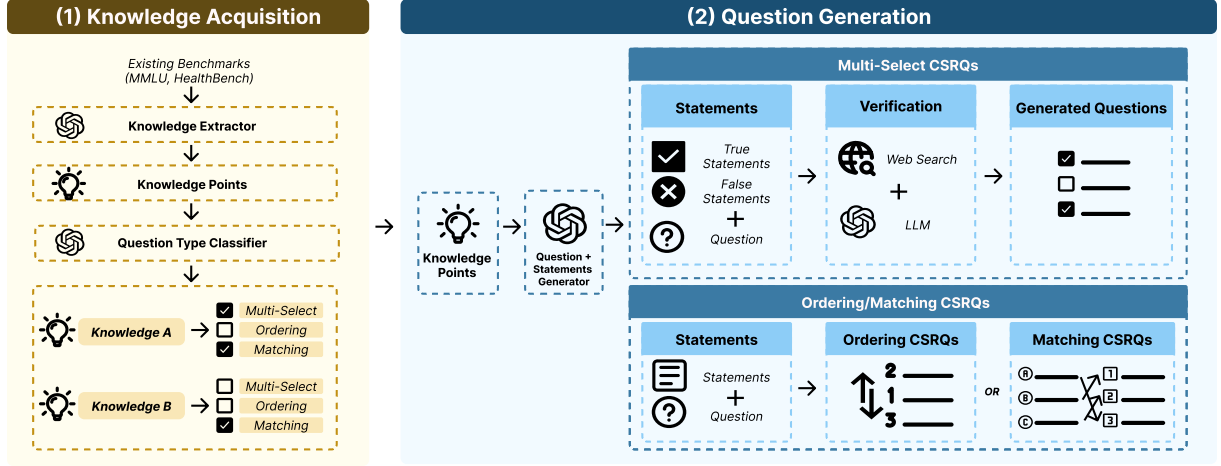


Figure 3: **CSRQ conversion workflow.** It converts suitable MCQ or CRQ items into CSRQ through knowledge acquisition, CSRQ-type selection, structured item generation, and verification. The workflow supports controlled analysis of instrument sensitivity under option-free elicitation and deterministic scoring.

### 3.2 Converting Existing Benchmarks

Direct extraction requires benchmarks with suitable multi-answer structures. To broaden our diagnostic, we convert existing benchmarks into matched formats. We use MMLU (Hendrycks et al., 2020) to examine score changes after removing final answer options, and HealthBench-Hard (Arora et al., 2025), originally evaluated with open-ended CRQs, to construct matched CSRQ and MCQ versions. This enables direct comparison across MCQ, CSRQ, and CRQ in a non-cultural domain while varying option exposure and scoring.

#### 3.2.1 Conversion workflow

Converted benchmarks provide diagnostic support for instrument sensitivity beyond *MyCulture*. We document a controlled workflow that converts suitable MCQ- and CRQ-based items into CSRQs, including item generation, filtering, and verification. We present it for reproducibility, not to imply that all benchmark items are convertible. Figure 3 summarizes the workflow.

**Knowledge acquisition.** Existing benchmark items are single-answer MCQs or CRQs and cannot be directly converted to CSRQs. We first extract the knowledge needed to answer each item. An LLM with web-search augmentation identifies relevant factual, relational, or procedural knowledge while preserving the intended content for structured CSRQ construction.

**CSRQ-type selection and generation.** An LLM-based classifier selects the most suitable CSRQ variant among multi-select, ordering, and match-

| Model          | MMLU  |       |          | HealthBench-Hard |       |       |
|----------------|-------|-------|----------|------------------|-------|-------|
|                | MCQ   | CSRQ  | $\Delta$ | MCQ              | CSRQ  | CRQ   |
| GPT-5          | 92.50 | 84.64 | 7.86     | 96.17            | 83.33 | 50.84 |
| Gemini-2.5-Pro | 91.06 | 78.35 | 12.71    | 93.99            | 75.68 | 31.07 |
| GLM-4.6        | 84.38 | 80.59 | 3.79     | 95.63            | 77.87 | 32.34 |
| Qwen3-235B     | 90.39 | 80.90 | 8.61     | 95.08            | 78.42 | 41.18 |

Table 1: **Supporting diagnostic on converted benchmarks.** All evaluated models are reasoning models. For MMLU,  $\Delta$  denotes the absolute difference between MCQ and CSRQ scores. For HealthBench-Hard, matched MCQ, CSRQ, and CRQ versions are constructed from the same underlying items, with CRQs judged by Qwen3-235B-Instruct.

ing based on the extracted knowledge and provides a brief rationale. It then generates the question and structured statements, grounding multi-select candidates in the extracted knowledge.

**Statement verification.** Generated statements are independently verified using a web-enabled LLM to reduce construction errors. For multi-select CSRQs, both valid and distractor statements are checked, with failed statements are discarded.

**Reproducibility.** The prompts, converted items, filtering logs, and conversion code are available [here](#) to support reproducible analysis of instrument sensitivity in CSRQ evaluation.

#### 3.2.2 Insights

Table 1 provides two complementary diagnostics. On MMLU, scores decrease after conversion from MCQ to CSRQ, consistent with sensitivity to removing final answer options. On HealthBench-Hard, all models score highest on MCQ, followed

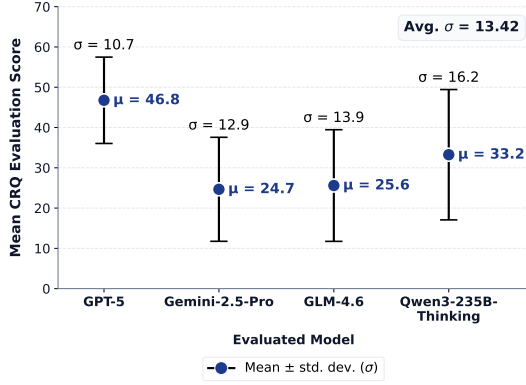


Figure 4: **CRQ evaluation scores under different LLM judges.** A higher standard deviation ( $\sigma$ ) indicates greater sensitivity to judge selection. All evaluated models are reasoning models. Full results are provided in Appendix J.

| Model               | OFE (%)      | ACE (%)      |
|---------------------|--------------|--------------|
| GPT-5               | 2.5          | 97.5         |
| Gemini-2.5-Pro      | 4.2          | 95.8         |
| GLM-4.6             | 8.3          | 91.7         |
| Qwen3-235B-Thinking | 30.0         | 70.0         |
| <b>Average</b>      | <b>11.25</b> | <b>88.75</b> |

Table 2: **Error analysis on a subsample of incorrectly answered questions from each model.** OFE denotes output-format error and ACE denotes answer-content error.

by CSRQ and CRQ, reproducing the three-format pattern observed on *MyCulture* in a non-cultural benchmark. We do not interpret this ordering as evidence that any format is universally easier, harder, or more valid, but rather that measured performance changes with answer-option exposure and scoring procedure.

**Judge sensitivity.** To examine judge-dependent sensitivity, Figure 4 evaluates the same CRQ outputs using different LLM judges under a fixed rubric. Scores vary substantially across judges, with an average standard deviation of 13.42%, indicating that measured performance depends in part on the scoring instrument. Full results are provided in Appendix J.

**Error analysis.** Table 2 classifies incorrect CSRQ responses into output-format errors (OFE) or answer-content errors (ACE). Most are content-related, with variation across models.

**Construction-model robustness.** Table 3 tests whether CSRQ results depend on the LLM used for item construction. We repeat conversion on

| Benchmark | Model   | Ori   | G25P  | G54   |
|-----------|---------|-------|-------|-------|
| MMLU      | GLM-4.6 | 84.80 | 83.63 | 87.72 |
|           | GPT-5   | 90.06 | 87.72 | 93.57 |
| HBH       | GLM-4.6 | 75.00 | 80.00 | 84.00 |
|           | GPT-5   | 77.00 | 88.00 | 86.00 |

Table 3: **Robustness of CSRQ evaluation to the LLM used for item construction.** Results are reported on matched subsets of MMLU and HealthBench-Hard (HBH). Ori. denotes the original GLM-4.6 construction pipeline; G25P denotes Gemini-2.5-Pro; and G54 denotes GPT-5.4.

| Model                            | MCQ          | CSRQ         |              | CRQ          |              |
|----------------------------------|--------------|--------------|--------------|--------------|--------------|
|                                  |              | L            | S            | L            | S            |
| Random                           | 25.53        | 40.33        | 4.28         | —            | —            |
| GPT-5 <sup>†</sup>               | <u>88.14</u> | <b>78.10</b> | <b>56.78</b> | <u>68.41</u> | <u>22.90</u> |
| GPT-4o                           | <b>90.28</b> | <u>78.00</u> | <u>56.01</u> | <b>68.61</b> | 21.14        |
| Gemini-2.5-Pro <sup>†</sup>      | 85.56        | 77.78        | 55.68        | 66.87        | 20.26        |
| GLM-4.6 <sup>†</sup>             | 84.46        | 76.37        | 53.38        | 63.20        | 17.74        |
| Qwen3-235B-Thinking <sup>†</sup> | 61.01        | 62.62        | 42.12        | 50.13        | 18.51        |
| Qwen3-235B-Instruct              | 80.12        | 73.55        | 51.29        | 62.93        | 18.34        |
| SeaLLMs-v3-7B                    | 29.98        | 52.96        | 26.30        | 47.62        | 7.86         |
| SEA-LION-v4-32B                  | 83.25        | 76.16        | 53.54        | 62.55        | 18.23        |

Table 4: **Models evaluated on *MyCulture*.** L denotes loose scoring, and S denotes strict scoring; scoring details and full results are provided in Appendix F. **Bold** indicates best performance and underline indicates second-best. <sup>†</sup> denotes reasoning models, and by default the reasoning effort is medium where available.

matched subsets of MMLU and HealthBench-Hard using Gemini-2.5-Pro and GPT-5.4 alongside the original GLM-4.6 pipeline. Although absolute scores vary, GPT-5 consistently outperforms GLM-4.6 on both benchmarks across all pipelines, suggesting that the comparisons are robust to construction-model choice.

### 3.3 *MyCulture* as a Cultural Testbed

We next instantiate the same instrument validity question in cultural evaluation. Cultural competence is a high-stakes testbed because benchmark scores are often interpreted as evidence that models understand local norms, practices, and values. We do not claim that one benchmark covers all cultural knowledge. Instead, we built *MyCulture* as a controlled, Malaysia-centred setting to examine whether measured cultural competence varies across the MCQ, CSRQ, and CRQ instruments.

To our knowledge, *MyCulture* is the *first Malaysia-centred cultural benchmark in Bahasa Malaysia*. It contains 5,463 questions across six

| MCQ                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | CSRQ                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | CRQ                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Question:</b><br/>Apakah kenyataan yang betul mengenai permainan Lansaran dan budaya suku kaum Murut?</p> <p><b>Statements:</b><br/>I. Lansaran dibina secara individu tanpa melibatkan masyarakat.<br/>II. Permainan Lansaran memerlukan ukuran kawasan antara 8 hingga 30 kaki persegi.<br/>III. Kayu "Selangan Batu" digunakan dalam pembinaan lantai Lansaran kerana ketahanannya.<br/>IV. Lansaran hanya untuk kanak-kanak dan tidak melibatkan orang dewasa.</p> <p><b>Options:</b> A. II, III B. I, IV C. I, II, IV D. II, IV</p> <p><b>Answer:</b> A</p> | <p><b>Question:</b><br/>Apakah kenyataan yang betul mengenai permainan Lansaran dan budaya suku kaum Murut?</p> <p><b>Statements:</b><br/>A. Lansaran dibina secara individu tanpa melibatkan masyarakat.<br/>B. Permainan Lansaran memerlukan ukuran kawasan antara 8 hingga 30 kaki persegi.<br/>C. Kayu "Selangan Batu" digunakan dalam pembinaan lantai Lansaran kerana ketahanannya.<br/>D. Lansaran hanya untuk kanak-kanak dan tidak melibatkan orang dewasa.</p> <p><b>Answer:</b> B, C</p> | <p><b>Question:</b><br/>Huraikan ciri-ciri fizikal, cara pembinaan dan peranan komuniti dalam permainan tradisional Lansaran suku kaum Murut berdasarkan amalan tradisi yang dinyatakan.</p> <p><b>Rubric:</b><br/>1. Tempat permainan Lansaran berukuran antara lapan (8) hingga 30 kaki persegi.<br/>2. Lansaran dapat menampung lebih kurang 30 orang pemain.<br/>3. Lansaran dibina secara bergotong-royong di ruang khas rumah panjang suku kaum Murut.<br/>4. Beberapa batang kayu bulat yang keras dan liat sebesar lengan orang dewasa dikenali sebagai kayu "Selangan Batu".<br/>5. Kayu "Selangan Batu" diikat dengan rotan saga bagi menampung lantai Lansaran.<br/>6. Lantai Lansaran dibina sekaki ke bawah daripada paras lantai rumah secara tergantung.</p> |

Figure 5: Example from the *MyCulture* dataset showing MCQ, CSRQ, and CRQ formats for assessing factual and descriptive knowledge of the Murut traditional game *Lansaran*, including its characteristics, construction, and community role. An English version is provided in Appendix C.

| Model              | Performance Diff. ( $\Delta$ = MCQ – CSRQ) |              |              |
|--------------------|--------------------------------------------|--------------|--------------|
|                    | Multi Select                               | Ordering     | Matching     |
| GPT-5              | 20.43                                      | <b>38.53</b> | <u>37.80</u> |
| Gemini-2.5-Pro     | 22.29                                      | <u>33.90</u> | 35.38        |
| GLM-4.6            | 24.86                                      | 30.48        | <b>39.86</b> |
| Qwen-235B-Thinking | <u>26.58</u>                               | 17.64        | 10.24        |
| Average $\Delta$   | 23.54                                      | <b>30.14</b> | <u>30.82</u> |

Table 5: **Performance difference ( $\Delta$ ) between MCQ and CSRQ** across Multi-Select, Ordering, and Matching types on *MyCulture*. **Bold** indicates the largest difference and underline indicates the second-largest.

| Ranking pair           | Spearman's $\rho$ | Kendall's $\tau$ |
|------------------------|-------------------|------------------|
| MCQ-CSRQ-Strict        | 0.95              | 0.86             |
| MCQ-CRQ-Strict         | 0.74              | 0.50             |
| CSRQ-Strict-CRQ-Strict | 0.79              | 0.64             |

Table 6: **Rank consistency across *MyCulture* evaluation instruments.** Correlations are computed over model rankings derived from strict-scoring results.

dimensions, including arts, attire, customary practices, entertainment, food, and religions. Each question is represented in MCQ, CSRQ, and CRQ formats, preserving the underlying knowledge while varying the evaluation instrument.

*MyCulture* is constructed from reliable Malaysian cultural and academic sources. LLMs generate and verify items across the three formats, check source-supported answers, and ensure that CSRQs are objectively parseable. Strong LLMs further flag ambiguous items for manual review, while qualified Malaysian residents audit 5% of the dataset. Construction details are provided in Appendix I, and an example is shown in Figure 5.

**Insights.** Table 4 shows that model scores differ substantially across MCQ, CSRQ, and CRQ

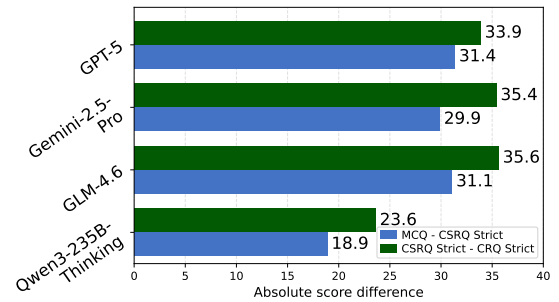


Figure 6: **Instrument gaps on *MyCulture* across models.** Each bar shows the absolute score change between two evaluation instruments. The MCQ-CSRQ gap reflects option-removal sensitivity, while the CSRQ-CRQ gap reflects the shift from deterministic structured scoring to open-ended judge-based scoring.

formats. MCQ scores are generally higher than strict CSRQ scores, showing that measured cultural competence changes when final answer options are removed. CRQ strict scores also differ substantially under judge-based grading, showing a different source of measurement sensitivity. The purpose of this result is not to rank formats by difficulty, but to show that the evaluation instrument changes the measured score.

To make the instrument effect more explicit, Figure 6 visualizes the absolute score gaps between MCQ, CSRQ-strict, and CRQ-strict evaluation across models. The MCQ-CSRQ gaps are substantial across visualized models, indicating that measured cultural competence decreases when the final answer-option layer is removed while deterministic scoring is preserved. Besides that, we provide a model breakdown of performance degradation in Table 5. The performance degradation is consistent across all CSRQ variants. Full details

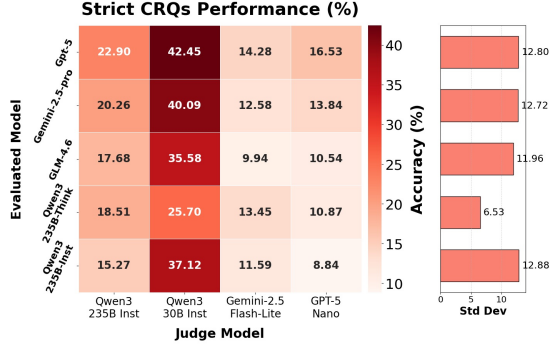


Figure 7: **Judge-induced sensitivity in *MyCulture* CRQ evaluation.** The same CRQ outputs are evaluated by different LLM judges under the same rubric. The observed variation shows that open-ended cultural evaluation can be sensitive to judge selection.

are available in Table 14 (Appendix G).

Additionally, Table 2 presents a manual error analysis of 120 incorrect responses per model, evenly sampled across the three question formats. On average, most errors are attributable to answer content issues (88.8%), whereas output-format errors account for only 11.3%. The latter mainly arise from response incompleteness, excessively long reasoning traces, or failure to follow the required answer format. Full details are provided in Table 8 (Appendix H).

The CSRQ-CRQ gaps are also large, indicating that scores change again when evaluation shifts from deterministic, structured verification to open-ended, judge-based scoring. This pattern supports our central claim that benchmark scores reflect not only model capability, but also the evaluation instrument used to measure it. Although absolute scores change across instruments, MCQ and CSRQ-Strict yield highly consistent rankings ( $\rho = 0.95$ ,  $\tau = 0.86$ ; Table 6), while CRQ-Strict shows lower agreement with MCQ ( $\rho = 0.74$ ,  $\tau = 0.50$ ) and CSRQ-Strict ( $\rho = 0.79$ ,  $\tau = 0.64$ ), showing greater rank shifts under judge-based scoring.

We also evaluate the same CRQ outputs under different LLM judges. Figure 7 shows substantial variation across judges even under the same rubric and evaluation setting, providing evidence of judge-dependent scoring in the cultural testbed. Open-ended cultural scores can therefore reflect both model response quality and the choice of judge. Pairwise judge agreement ranges from 0.40 to 1.00 for both Spearman’s  $\rho$  and Kendall’s  $\tau$  (Table 7). Judges agree on the top model but differ elsewhere, showing that CRQ judging affects both relative rankings and absolute scores.

| Judge pair | Spearman’s $\rho$ | Kendall’s $\tau$ |
|------------|-------------------|------------------|
| Q235-Q30   | 0.60              | 0.40             |
| Q235-GFL   | 0.80              | 0.60             |
| Q235-G5N   | 1.00              | 1.00             |
| Q30-GFL    | 0.40              | 0.40             |
| Q30-G5N    | 0.60              | 0.40             |
| GFL-G5N    | 0.80              | 0.60             |

Table 7: **Agreement between model rankings produced by different CRQ judges.** Q235: Qwen3-235B-Instruct; Q30: Qwen3-30B-Instruct; GFL: Gemini-2.5-Flash-Lite; G5N: GPT-5-Nano.

Together, both converted benchmarks and *MyCulture* support the same conclusion. The measured competence is not only a property of the model, but it also depends on the evaluation instrument used to measure it.

### 3.4 Practical Instrument Selection

MCQs are well-suited to low-cost screening and leaderboard comparisons because of their reproducible scoring. CSRQs provide a middle ground by removing final-answer candidates while retaining deterministic, judge-free scoring. CRQs support more expressive responses when reliable human or multi-judge evaluation is available. We recommend CSRQs as a matched audit of MCQ results when option scaffolding may affect absolute scores, especially in low-resource settings, whereas CRQ-based comparisons should account for judge variation through multi-judge comparisons or sensitivity analyses.

## 4 Related Work

This work relates to four research lines, which are (a) MCQ-based LLM evaluation, (b) option-free and answer-matching evaluation, (c) open-ended and (d) judge-based evaluation, and cultural LLM evaluation. We position CSRQ not as a new item archetype, but as an instrument validity protocol that recomposes two useful functions, which are option-free elicitation and deterministic scoring.

**MCQ-based LLM evaluation.** MCQs are widely used because they are scalable, reproducible, and objectively scored. Representative benchmarks include MMLU and its variants (Hendrycks et al., 2020; Li et al., 2023; Poh et al., 2024), HellaSwag (Zellers et al., 2019), CommonsenseQA (Talmor et al., 2019), and OpenBookQA (Mihaylov et al., 2018). However, visible final-answer options introduce

guessing (Burton, 2001) and can enable shortcuts, elimination, and option-conditioned reasoning, particularly in reasoning models (Chandak et al., 2025; Raman et al., 2025; Balepur et al., 2025). We call this answer-option scaffolding, where measured scores partly reflect the provided options rather than model competence alone.

#### **Option-free and answer-matching evaluation.**

Recent work (Chandak et al., 2025; Balepur et al., 2025; Hutchinson, 2026; Zhang and Nguyen, 2025) evaluates LLMs without exposing final-answer candidates. Chandak et al. (2025) use free-form responses with a reference-guided LLM matcher, whereas BiGGen Bench (Kim et al., 2025) uses instance-specific rubrics and evaluator LMs for graded scoring. These approaches reduce visible-option scaffolding but rely on unconstrained generation followed by reference-guided matching or rubric-based evaluation. CSRQ instead constrains responses to canonical structures, such as subsets, permutations, and matchings, enabling deterministic parsing and exact verification without an evaluation-time judge. Thus, CSRQ is a middle-ground instrument that combines option-free elicitation with deterministic scoring and supports matched MCQ–CSRQ–CRQ comparisons of final-answer exposure and judge dependence.

#### **Open-ended and judge-based evaluation.**

Open-ended CRQs remove answer options and therefore better support option-free elicitation. Benchmarks such as HealthBench (Arora et al., 2025), SimpleQA (Wei et al., 2024), and GSM8K (Cobbe et al., 2021) rely on reference answers, rubrics, or judge-based scoring procedures. Recent work also argues that constructed-answer formats can better reflect model knowledge and user needs than conventional MCQA in some settings (Balepur et al., 2025). However, open-ended evaluation introduces another measurement problem, that is, the score depends on the evaluator. LLM-as-a-judge can exhibit position bias, self-bias, prompt sensitivity, task-dependent variability, and disagreement across judge models (Ye et al., 2024; Shi et al., 2025; Watts et al., 2024; Bavaresco et al., 2025). We study this as judge-dependent scoring, where the measured score reflects both the evaluated response and the scoring instrument. CSRQ addresses a narrower but complementary question: *whether some option-free evaluation settings can retain deterministic scoring without requiring*

*human or LLM judges.*

#### **Structured constructed-response assessment.**

Multi-select, ordering, and matching tasks have long existed in educational and psychological assessment (Haladyna, 2004; Nitko and Brookhart, 2006). CSRQ does not claim novelty in these item archetypes. Its contribution is to use such structured output spaces as canonical answer spaces for LLM evaluation. This allows the model to produce an answer without seeing final answer options, while correctness is determined by deterministic parsing and exact structural verification. Importantly, CSRQ removes the final answer-option layer, not all task-provided textual units. The statements or descriptions in a CSRQ define the task space, whereas MCQ options provide pre-composed candidate answers that may directly reveal the correct subset, order, or association. CSRQ therefore recomposes the useful scoring function of MCQs with the option-free elicitation function of CRQs.

**Cultural LLM evaluation.** Cultural benchmarks aim to evaluate models ability to capture culturally situated knowledge, including local norms, practices, languages, and values. Prior work has introduced culturally grounded benchmark datasets across diverse regional and linguistic contexts (Koto et al., 2024; Wang et al., 2024). Herein, we proposed *MyCulture* as the testbed.

## **5 Conclusion**

This paper frames LLM evaluation as an instrument validity problem. MCQs offer deterministic scoring but may inflate competence through recognition, elimination, and partial-knowledge cues, while CRQs remove answer options but introduce grader-dependent variation. We introduce Constrained Structured Response Questions (CSRQs), which remove the final answer-option layer while preserving deterministic scoring. Across converted knowledge benchmarks and *MyCulture*, we show that performance changes with the evaluation format. Scores generally decrease when MCQ options are removed, while CRQ scores vary across judges. CSRQs, therefore, provide a middle ground between MCQ objectivity and CRQ option-free elicitation. Rather than replacing benchmarks, CSRQs offer an auditable instrument for studying how benchmark scores depend on protocols.

## **Limitation**

CSRQs have several limitations. First, CSRQs are not universally applicable, as they require that answers be decomposable into at least two valid response units (e.g., selection, ordering, or matching), which makes some tasks difficult to construct or convert without reformulation. Second, CSRQs are more suitable for knowledge that can be represented through structured statements and may be less applicable to highly abstract, creative, or unconstrained reasoning tasks. Third, although CSRQs remove explicit answer options and reduce judge dependence, the response space remains finite and constrained by the provided statements, and therefore does not fully replicate the open-ended generation found in CRQs. Finally, constructing CSRQs introduces additional benchmark development costs, including statement generation, distractor creation, and verification procedures to ensure correctness and consistency.

## References

- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *Preprint*, arXiv:2505.08775.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced B\) flawed C\) fixable D\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Richard F Burton. 2001. Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1):41–50.
- Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. 2025. [Answer matching outperforms multiple choice for language model evaluation](#). *Preprint*, arXiv:2507.02856.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Thomas M Haladyna. 2004. *Developing and validating multiple-choice test items*. Routledge.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Daniel Hutchinson. 2026. Benchmarking as source criticism: From recognition to reasoning in llm assessment. *Journal of Open Humanities Data*, 12(1).
- Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2):130–144.
- Michael T Kane. 2013. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73.
- Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, and 13 others. 2025. [The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models](#). *Preprint*, arXiv:2406.05761.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. Indoculture: Exploring geographically influenced cultural commonsense reasoning across eleven indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.
- Samuel A Livingston. 2009. Constructed-response test questions: Why we use them; how we score them. r&d connections. number 11. *Educational Testing Service*.
- Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2381–2391.
- Anthony J Nitko and Susan M Brookhart. 2006. *Educational assessment of students*. Prentice-Hall, Inc.
- OpenAI. 2025. [Gpt-5: Large language model](#). Accessed: 2026-01-03.
- Soon Poh, Sze Jue Yang, Jeraelyn Tan, Lawrence Chieng, Jia Tan, Zhenyu Yu, Foong Mun, and Chee Seng Chan. 2024. Malaymmlu: A multitask benchmark for the low-resource malay language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 650–669.

- Narun Raman, Taylor Lundy, and Kevin Leyton-Brown. 2025. Reasoning models are test exploiters: Rethinking multiple-choice. *arXiv preprint arXiv:2507.15337*.
- Michael C Rodriguez. 2003. Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2):163–184.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Stephen Sireci and Molly Faulkner-Bond. 2014. Validity evidence based on test content. *Psicothema*, 26(1):100–107.
- Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, and 1 others. 2025. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. In *International Conference on Learning Representations*, volume 2025, pages 55244–55288.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.
- Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024. Kulture bench: A benchmark for assessing language model in korean cultural context. *arXiv preprint arXiv:2412.07251*.
- Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. [PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *Preprint*, arXiv:2411.04368.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. [Justice or prejudice? quantifying biases in llm-as-a-judge](#). *Preprint*, arXiv:2410.02736.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.
- Yue Zhang and Nhan Nguyen. 2025. [Rethinking MCQ benchmarks: Mandatory reasoning evaluation reveals significant performance drops in large language models](#). In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.
- 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.

## Appendix

This appendix provides supplementary materials for the experiments and analyses presented in the main paper. We first present the chance-success analysis in Appendix A, followed by the evaluation prompts in Appendix B, representative *MyCulture* examples in Appendix C, and model artifacts in Appendix D. We then report the full results for benchmark conversion, *MyCulture* evaluation, CSRQ subtype sensitivity, and error analysis in Appendices E, F, G, and H, respectively.

### A Chance-Success Analysis

Figure 8 visualizes chance-success baselines for MCQ, CSRQ variants, and an approximate CRQ reference baseline under exact structural scoring. The CRQ baseline is shown as near-zero because open-ended responses have an effectively unconstrained answer space under exact-match scoring, making random exact success negligible. This comparison is intended only as a structural baseline, such that it characterizes exact random success, not semantic difficulty, cognitive complexity, or genuine model understanding.

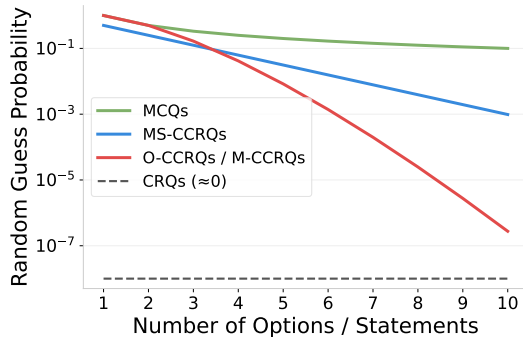


Figure 8: **Chance-success baselines under exact structural scoring.** The figure compares random-success probabilities for MCQs, CSRQ variants, and an approximate CRQ reference baseline as the number of options or statements increases. The CRQ baseline is shown as near-zero to reflect its open-ended answer space under exact-match scoring. The analysis characterizes chance success under exact scoring only, not semantic difficulty or cognitive complexity.

### B Prompt Used in Experiments

Figure 9 and 10 illustrates the prompt used for MCQ evaluation in all experiment, Figure 11 present the prompts used for CRQ evaluation on *MyCulture* only. For HealthBench- Hard (Arora

et al., 2025) experiment, we follow their original prompt and judging method.

### C Example of *MyCulture* dataset

In this section, we present representative examples from the *MyCulture* dataset, encompassing MCQs, CSRQs, and CRQs across three question structures, which are multi-select, ordering, and matching. Figures 12, 13, 14 illustrate example MCQs for the three structures, namely multi-select, ordering, and matching. Figures 15, 16, 17 present corresponding examples of CSRQs across the same three structures. Figure 18 provides an example of a CRQ.

### D Model Artifacts

Table 10 summarizes the model version artifacts used during the experiment.

### E Full performance for conversion of existing benchmark

This section shows the full performance of the model for the experiment in Section 3.1 and Section 3.2. Table 11 shows the performance of the model on converting the existing multi-answer MCQs benchmark to CSRQ, while Table 12 shows the performance of the model’s benchmark of CSRQ generated from the proposed conversion pipeline.

### F Full performance for *MyCulture*

This section shows the full performance of the model for the experiment in Section 3.3. Table 13 shows the performance of the model evaluated on *MyCulture*, which consists of 3 types of formats.

**Scoring Strategies for CSRQs and CRQs.** Since both CSRQs and CRQs are option-free, we evaluate model responses using two scoring strategies, which are partial correctness and full correctness. The *loose* metric captures partial correctness, assigning credit when part of the predicted answer overlaps with the ground truth. The *strict* metric captures full correctness, requiring the predicted answer to exactly match the complete ground-truth sequence.

### G CSRQ subtype sensitivity

Table 14 reports the performance degradation across the three question types Multi-Answer, Ordering, and Matching when transitioning from

| Model               | Category                | Count     | %           | MS       | O         | M         |
|---------------------|-------------------------|-----------|-------------|----------|-----------|-----------|
| GPT-5               | Response Incompleteness | 3         | 2.5         | –        | –         | 3         |
|                     | <b>Total</b>            | <b>3</b>  | <b>2.5</b>  | <b>–</b> | <b>–</b>  | <b>3</b>  |
| Gemini-2.5-Pro      | Response Incompleteness | 5         | 4.2         | –        | 3         | 2         |
|                     | <b>Total</b>            | <b>5</b>  | <b>4.2</b>  | <b>–</b> | <b>3</b>  | <b>2</b>  |
| GLM-4.6             | Response Incompleteness | 10        | 8.3         | –        | –         | 10        |
|                     | <b>Total</b>            | <b>10</b> | <b>8.3</b>  | <b>–</b> | <b>–</b>  | <b>10</b> |
| Qwen3-235B-Thinking | Reasoning Truncation    | 30        | 25.0        | 6        | 13        | 11        |
|                     | Response Incompleteness | 5         | 4.2         | –        | –         | 5         |
|                     | Instruction Following   | 1         | 0.8         | –        | –         | 1         |
|                     | <b>Total</b>            | <b>36</b> | <b>30.0</b> | <b>6</b> | <b>13</b> | <b>17</b> |

Table 8: **Breakdown of output-format errors by remark category.** MS, O, and M denote Multi-Select, Ordering, and Matching question formats, respectively.

MCQ to CSRQ evaluation on *MyCulture*. A response incompleteness occurs when models do not generate all required matches, resulting in fewer matching pairs than the total number of statement pairs. In contrast, an instruction following error occurs when models fail to produce outputs in the required formats. Reasoning truncation refers to cases where the response is truncated because the reasoning process exceeds the available output length.

## H Error analysis on *MyCulture*

Table 8 breaks down output format errors by model and question format. Most parse errors come from Qwen3-235B-Thinking, mainly due to overly long reasoning outputs, while the other models mostly produce incomplete answers. The errors are concentrated in matching and ordering questions, with fewer issues observed in multi-select questions.

## I Construction details of *MyCulture*

We describe the construction pipeline of *MyCulture* in Algorithm 1. We first scrape and curate textual materials from reliable sources such as Malaysia’s National Department for Culture and Arts (JKKN)<sup>1</sup> and relevant academic publications. We then use LLMs to draft question-answer pairs across MCQ, CSRQ and CRQ formats. Additionally, we leverage LLMs to verify the format compliance with the target question type and evidence consistency, namely that the proposed answer is supported by the retrieved source content. For CSRQs, we additionally require the answer to be expressible in a canonical form that is deterministically parseable for objective scoring.

As an auxiliary sanity check, we run a small

### Algorithm 1 Construction Pipeline for *MyCulture*

---

**Require:** Reliable cultural sources  $\mathcal{S}$  (e.g., JKKN, academic journals), initial system prompt  $P_{\text{sys}}$   
**Ensure:** High-quality validated cultural dataset  $\mathcal{D}_{\text{final}}$

- 1: **Step 1: Data Collection**
- 2:  $\mathcal{C} \leftarrow \text{ScrapeContent}(\mathcal{S})$
- 3: **while true do**
- 4:   **Step 2: Question-Answer Generation**
- 5:    $\mathcal{D}_{\text{raw}} \leftarrow \text{GenerateQA}(\mathcal{C}, P_{\text{sys}})$
- 6:   **Step 3: Automated Quality Filtering**
- 7:    $\mathcal{D}_{\text{filtered}} \leftarrow \emptyset$
- 8:   **for all**  $(q, a) \in \mathcal{D}_{\text{raw}}$  **do**
- 9:      $v_{\text{struct}} \leftarrow \text{JudgeStructure}(q)$
- 10:     $v_{\text{fact}} \leftarrow \text{CheckAnswerability}(q, a)$
- 11:    **if**  $v_{\text{struct}} = \text{Valid}$  **and**  $v_{\text{fact}} = \text{Correct}$  **then**
- 12:      $\mathcal{D}_{\text{filtered}} \leftarrow \mathcal{D}_{\text{filtered}} \cup \{(q, a)\}$
- 13:    **end if**
- 14:   **end for**
- 15:   **Step 4: Human Validation**
- 16:    $\mathcal{D}_{\text{sample}} \leftarrow \text{Subsample}(\mathcal{D}_{\text{filtered}}, 5\%)$
- 17:    $V_{\text{human}} \leftarrow \text{HumanValidate}(\mathcal{D}_{\text{sample}})$
- 18:   **if**  $V_{\text{human}} = \text{Pass}$  **then**
- 19:      $\mathcal{D}_{\text{final}} \leftarrow \mathcal{D}_{\text{filtered}}$
- 20:     **return**  $\mathcal{D}_{\text{final}}$
- 21:   **else**
- 22:      $P_{\text{sys}} \leftarrow \text{RewritePrompt}(P_{\text{sys}})$
- 23:   **end if**
- 24: **end while**

---

set of strong LLMs (e.g., Gemini-2.5-Pro, GPT-5, GLM-4.6, Qwen3-235B) on the filtered items to surface potentially ambiguous or ill-posed questions. We do not treat model failure as evidence of invalidity. Instead, flagged items are routed to manual inspection or discarded if they lack clear source support.

Finally, we conduct human auditing on a 5% subsample of the filtered dataset. Auditors are Malaysian residents with at least 10 years of lived experience in Malaysia, hold at least a bachelor’s degree, and have completed the national Malay examination (SPM)<sup>2</sup>. We hire five auditors who assess whether each audited item is logical, culturally relevant, and answerable based on the provided source evidence. If the audit reveals recurring failure modes such as unsupported answers, ambiguous phrasing, or format violations, we refine the generation prompt and filtering rules and repeat the process. The hiring and audit processes are reviewed by the Ethics Review Board (ERB), and all auditors are hired in compliance with Malaysia’s minimum wage requirements. In total, we have constructed 5,463 questions, where each type of question, MCQ, CSRQ and CRQ, gets 1,821 questions.

<sup>1</sup><https://pemetaanbudaya.jkkn.gov.my/>

<sup>2</sup>[https://ms.wikipedia.org/wiki/Sijil\\_Pelajaran\\_Malaysia](https://ms.wikipedia.org/wiki/Sijil_Pelajaran_Malaysia)

## **J Performance of Models on various Judge**

Table 9 show the detailed performance across different judges. The experiment is discussed in Section 3.2.

| Evaluated Model     | Judge Model        |         |                | $\sigma$     |
|---------------------|--------------------|---------|----------------|--------------|
|                     | Qwen-235B-Instruct | GPT-5.2 | Gemini-2.5-Pro |              |
| GPT-5               | 50.84              | 34.60   | 54.85          | 10.72        |
| Gemini-2.5-Pro      | 31.07              | 9.80    | 33.12          | 12.91        |
| GLM-4.6             | 32.34              | 9.66    | 34.79          | 13.86        |
| Qwen3-235B-Thinking | 41.18              | 14.63   | 43.94          | 16.18        |
| <b>Average</b>      | –                  | –       | –              | <b>13.42</b> |

Table 9: **CRQ evaluation scores under different LLM judges.** Higher standard deviation ( $\sigma$ ) indicates greater sensitivity to judge selection. All models are evaluated with the reasoning mode.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Malay Prompt</b></p> <p>Berikut adalah soalan. Format jawapan mengikut jenis soalan adalah seperti berikut: bagi soalan jenis Gabungan, gunakan format jawapan \boxed{A}, \boxed{B}, \boxed{C}, atau \boxed{D}; bagi soalan jenis Berurutan, gunakan format jawapan \boxed{A}, \boxed{B}, \boxed{C}, atau \boxed{D}; dan bagi soalan jenis Padanan, gunakan format jawapan \boxed{A}, \boxed{B}, \boxed{C}, atau \boxed{D}. Berikut adalah soalan: {soalan}. Jenis soalan: {jenissoalan}. Jawapan:</p> |
| <p><b>English Prompt</b></p> <p>Below is a question. The answer format according to the question type is as follows: for Combination questions, use the answer format \boxed{A}, \boxed{B}, \boxed{C}, or \boxed{D}; for Ordering questions, use the answer format \boxed{A}, \boxed{B}, \boxed{C}, or \boxed{D}; and for Matching questions, use the answer format \boxed{A}, \boxed{B}, \boxed{C}, or \boxed{D}. The question is as follows: {question}. Question type: {questiontype}. Answer:</p>        |

Figure 9: **Prompt used for the all MCQ experiment.** The upper prompt (in Malay) shows the Malay prompt, while the lower prompt presents its English version.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Malay Prompt</b></p> <p>Berikut adalah soalan. Jawapan perlu diletakkan dalam format \boxed{}. Format jawapan mengikut jenis soalan adalah seperti berikut: bagi soalan jenis Gabungan, gunakan format \boxed{A,B,C,D}; bagi soalan jenis Berurutan, gunakan format \boxed{A,B,C,D}; dan bagi soalan jenis Padanan, gunakan format \boxed{[[A,i],[B,ii],[C,iii],...]}. Berikut adalah soalan: {soalan}. Jenis soalan: {jenissoalan}. Jawapan:</p>           |
| <p><b>English Prompt</b></p> <p>Below is a question. The answer must be provided in the \boxed{} format. According to the question type, the answer format is as follows: For Combination questions, use the format \boxed{A,B,C,D}; For Ordering questions, use the format \boxed{A,B,C,D}; For Matching questions, use the format \boxed{[[A,i],[B,ii],[C,iii],...]}.</p> <p>The question is as follows: {question}. Question type: {questiontype}. Answer:</p> |

Figure 10: **Prompt used for the all CSRQs experiment.** The upper prompt (in Malay) shows the Malay prompt, while the lower prompt presents its English version.

**Prompt**  
Sila jawab soalan berikut: {soalan} Jawapan:

**Translated Prompt**  
Please answer the following question: {question} Answer:

Figure 11: **Prompt used for the CRQs experiments *MyCulture*.** The upper prompt (in Malay) shows the original prompt, while the lower prompt presents its translated version.

Multi Select - MCQs

**Original Question**  

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

**Pernyataan:**

- I. Lansaran dibina secara individu tanpa melibatkan masyarakat.
- II. Permainan Lansaran memerlukan ukuran kawasan antara 8 hingga 30 kaki persegi.
- III. Kayu 'Selangan Batu' digunakan dalam pembinaan lantai Lansaran kerana ketahanannya.
- IV. Lansaran hanya untuk kanak-kanak dan tidak melibatkan orang dewasa.

**Pilihan:**

- A. II, III
- B. I, IV
- C. I, II, IV
- D. II, IV

**Jawapan:** A

**Translated Question**  

What statements are correct regarding the Lansaran game and the culture of the Murut community?

**Statements:**

- I. The Lansaran is built individually without involving the community.
- II. The Lansaran game requires an area size of between 8 and 30 square feet.
- III. "Selangan Batu" wood is used in constructing the Lansaran floor due to its durability.
- IV. The Lansaran is only for children and does not involve adults.

**Options:**

- A. II, III
- B. I, IV
- C. I, II, IV
- D. II, IV

**Answer:** A

Figure 12: **Example of a multi-select MCQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

Ordering - MCQs

**Original Question**  

Pilih urutan yang betul bagi proses berikut berkaitan dengan penghasilan kuih tradisional:

**Pernyataan:**

- I. Uli tepung gandum dan basuh tepung untuk membuang gluten.
- II. Proses membasuh tepung bertujuan menghasilkan kanji.
- III. Air tepung diperap dan buang air setiap 6 jam selama 3 hari.
- IV. Masukkan tepung yang mendap ke dalam air gula dan kacau.

**Pilihan:**

- A. I, II, III, IV
- B. II, I, IV, III
- C. III, I, IV, II
- D. IV, III, II, I

**Jawapan:** A

**Translated Question**  

Select the correct order for the following process related to the production of traditional delicacy:

**Statements:**

- I. Knead the wheat flour and wash it to remove gluten.
- II. The process of washing the flour aims to produce starch.
- III. The flour water is left to ferment, and the water is discarded every 6 hours for 3 days.
- IV. Add the settled flour into sugar water and stir.

**Options:**

- A. I, II, III, IV
- B. II, I, IV, III
- C. III, I, IV, II
- D. IV, III, II, I

**Answer:** A

Figure 13: **Example of an ordering MCQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

## Matching - MCQs

| Original Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Translated Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Padankan alat muzik dengan peranannya.</p> <p><b>Kumpulan Satu:</b><br/> A. Kertuk Ulu<br/> B. Tamborin<br/> C. Rebana<br/> D. Canang</p> <p><b>Kumpulan Dua:</b><br/> I. Menambah ritma dan variasi bunyi.<br/> II. Memberikan bunyi gabungan gitar bass.<br/> III. Sebagai penyeri irama di bahagian klimaks.<br/> IV. Sebagai lapisan ritma tambahan.</p> <p><b>Pilihan:</b><br/> A. A-II, B-I, C-IV, D-III<br/> B. A-I, B-IV, C-II, D-III<br/> C. A-IV, B-I, C-III, D-II<br/> D. A-II, B-III, C-IV, D-I</p> <p><b>Jawapan:</b> C</p> | <p>Match the musical instruments with their roles.</p> <p><b>Group One:</b><br/> A. Kertuk Ulu<br/> B. Tambourine<br/> C. Rebana<br/> D. Canang</p> <p><b>Group Two:</b><br/> I. Adds rhythm and sound variation.<br/> II. Produces a sound combination similar to a bass guitar.<br/> III. Serves as a rhythmic embellishment during the climax section.<br/> IV. Acts as an additional rhythmic layer.</p> <p><b>Options:</b><br/> A. A-II, B-I, C-IV, D-III<br/> B. A-I, B-IV, C-II, D-III<br/> C. A-IV, B-I, C-III, D-II<br/> D. A-II, B-III, C-IV, D-I</p> <p><b>Answer:</b> C</p> |

Figure 14: **Example of a matching MCQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

## Multi Select - CSRQ

| Original Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Translated Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Apakah kenyataan yang betul mengenai permainan Lansaran dan budaya suku kaum Murut?</p> <p><b>Pernyataan:</b><br/> A. Lansaran dibina secara individu tanpa melibatkan masyarakat.<br/> B. Permainan Lansaran memerlukan ukuran kawasan antara 8 hingga 30 kaki persegi.<br/> C. Kayu 'Selangan Batu' digunakan dalam pembinaan lantai Lansaran kerana ketahanannya.<br/> D. Lansaran hanya untuk kanak-kanak dan tidak melibatkan orang dewasa.</p> <p><b>Jawapan:</b> B,C</p> | <p>What statements are correct regarding the Lansaran game and the culture of the Murut community?</p> <p><b>Statements:</b><br/> A. The Lansaran is built individually without involving the community.<br/> B. The Lansaran game requires an area size of between 8 and 30 square feet.<br/> C. "Selangan Batu" wood is used in constructing the Lansaran floor due to its durability.<br/> D. The Lansaran is only for children and does not involve adults.</p> <p><b>Answer:</b> B,C</p> |

Figure 15: **Example of a multi-select CSRQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

## Ordering - CSRQ

| Original Question                                                                                                                                                                                                                                                                                                                                                                                                                  | Translated Question                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Pilih urutan yang betul bagi proses berikut berkaitan dengan penghasilan kuih tradisional:</p> <p><b>Pernyataan:</b><br/> A. Uli tepung gandum dan basuh tepung untuk membuang gluten.<br/> B. Proses membasuh tepung bertujuan menghasilkan kanji.<br/> C. Air tepung diperap dan buang air setiap 6 jam selama 3 hari.<br/> D. Masukkan tepung yang mendap ke dalam air gula dan kacau.</p> <p><b>Jawapan:</b> A, B, C, D</p> | <p>Select the correct order for the following process related to the production of traditional delicacy:</p> <p><b>Description:</b><br/> A. Knead the wheat flour and wash it to remove gluten.<br/> B. The process of washing the flour aims to produce starch.<br/> C. The flour water is left to ferment, and the water is discarded every 6 hours for 3 days.<br/> D. Add the settled flour into sugar water and stir.</p> <p><b>Answer:</b> A, B, C, D</p> |

Figure 16: **Example of an ordering CSRQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

## Matching - CSRQ

| Original Question                                                                                                                                                                                                                                                                                                                                                                                     | Translated Question                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Padankan alat muzik dengan peranannya.</p> <p><b>Kumpulan Satu:</b><br/>A. Kertuk Ulu<br/>B. Tamborin<br/>C. Rebana<br/>D. Canang</p> <p><b>Kumpulan Dua:</b><br/>I. Menambah ritma dan variasi bunyi.<br/>II. Memberikan bunyi gabungan gitar bass.<br/>III. Sebagai penyeri irama di bahagian klimaks.<br/>IV. Sebagai lapisan ritma tambahan.</p> <p><b>Jawapan:</b> A-IV, B-I, C-III, D-II</p> | <p>Match the musical instruments with their roles.</p> <p><b>Group One:</b><br/>A. Kertuk Ulu<br/>B. Tambourine<br/>C. Rebana<br/>D. Canang</p> <p><b>Group Two:</b><br/>I. Adds rhythm and sound variation.<br/>II. Produces a sound combination similar to a bass guitar.<br/>III. Serves as a rhythmic embellishment during the climax section.<br/>IV. Acts as an additional rhythmic layer.</p> <p><b>Answer:</b> A-IV, B-I, C-III, D-II</p> |

Figure 17: **Example of a matching CSRQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

## CRQ

| Original Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Translated Question                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Huraikan ciri-ciri fizikal, cara pembinaan dan peranan komuniti dalam permainan tradisional Lansaran suku kaum Murut berdasarkan amalan tradisi yang dinyatakan.</p> <p><b>Titik jawapan:</b></p> <ul style="list-style-type: none"><li>• Tempat permainan Lansaran berukuran antara 8 hingga 30 kaki persegi.</li><li>• Lansaran dapat menampung lebih kurang 30 orang pemain.</li><li>• Lansaran dibina secara bergotong-royong di ruang khas rumah panjang suku kaum Murut.</li><li>• Kayu 'Selangan Batu' digunakan kerana keras dan liat.</li><li>• Kayu diikat dengan rotan saga untuk menampung lantai.</li><li>• Lantai dibina sekaki ke bawah secara tergantung.</li></ul> | <p>Describe the physical characteristics, construction methods, and the role of the community in the traditional Lansaran game of the Murut ethnic group based on the stated traditional practices.</p> <p><b>Answer points:</b></p> <ul style="list-style-type: none"><li>• The Lansaran playing area measures between 8 and 30 square feet.</li><li>• The Lansaran can accommodate approximately 30 players.</li><li>• The Lansaran is constructed through communal cooperation (gotong-royong) in a designated space within the Murut longhouse.</li><li>• Selangan Batu wood is used because it is hard and resilient.</li><li>• The wood is bound with rotan saga to support the floor structure.</li><li>• The floor is built suspended, approximately one foot below the ground.</li></ul> |

Figure 18: **Example of a CRQ.** The left panel presents the original question, while the right panel shows its translated counterpart.

| Model                   | Artifact / Version                         |
|-------------------------|--------------------------------------------|
| GPT-5.2                 | gpt-5.2-2025-12-11                         |
| GPT-5                   | gpt-5-2025-08-07                           |
| GPT-5-mini              | gpt-5-mini-2025-08-07                      |
| GPT-5-nano              | gpt-5-nano-2025-08-07                      |
| Gemini-2.5-Pro*         | latest                                     |
| Gemini-2.5-Flash*       | latest                                     |
| Gemini-2.5-Flash-Lite*  | latest                                     |
| GLM-4.6*                | latest                                     |
| GLM-4-32B-128K          | 0414                                       |
| Qwen3-235B-Thinking     | Qwen/Qwen3-235B-A22B-Thinking-2507         |
| Qwen3-30B-Thinking      | Qwen/Qwen3-30B-A3B-Thinking-2507           |
| Qwen3-4B-Thinking       | Qwen/Qwen3-4B-Thinking-2507                |
| Qwen3-235B-Instruct     | Qwen/Qwen3-235B-A22B-Instruct-2507         |
| Qwen3-30B-Instruct      | Qwen/Qwen3-30B-A3B-Instruct-2507           |
| Qwen3-4B-Instruct       | Qwen/Qwen3-4B-Instruct-2507                |
| SeaLLM-v3-7B            | SeaLLMs/SeaLLMs-v3-7B-Chat                 |
| SeaLLM-v3-1.5B          | SeaLLMs/SeaLLMs-v3-1.5B-Chat               |
| Qwen-SEA-LION-V4-32B-IT | aisingapore/Qwen-SEA-LION-v4-32B-IT        |
| MaLLaM-1.1B             | mesolitica/mallam-1.1b-20k-instructions-v2 |
| MaLLaM-5B               | mesolitica/mallam-5b-20k-instructions-v2   |

Table 10: **Model variants and their associated artifact versions used in the experiments are reported.** For models without explicit release identifiers, the most recent publicly available version at the time of evaluation was adopted. An asterisk (\*) denotes models that do not disclose artifact details; in such cases, the version corresponds to the latest available release as of May 2026.

| Family               | Model                              | Reasoning | MalayMMLU    |       |              | CS-EN        |       |              | Avg. $\Delta$ |
|----------------------|------------------------------------|-----------|--------------|-------|--------------|--------------|-------|--------------|---------------|
|                      |                                    |           | MCQ          | CSRQ  | $\Delta$     | MCQ          | CSRQ  | $\Delta$     |               |
| Closed-source Models |                                    |           |              |       |              |              |       |              |               |
| OpenAI               | GPT-5 <sup>†</sup>                 | ✓         | 81.98        | 59.46 | 22.52        | 76.32        | 63.16 | 13.16        | 17.84         |
|                      | GPT-5-mini <sup>†</sup>            | ✓         | 78.08        | 50.45 | 27.63        | 69.74        | 47.37 | 22.37        | 25.00         |
|                      | GPT-5-nano <sup>†</sup>            | ✓         | 67.87        | 38.44 | 29.43        | 64.47        | 40.79 | 23.68        | 26.56         |
| Google               | Gemini-2.5-Pro <sup>†</sup>        | ✓         | 84.38        | 56.16 | 28.22        | 80.26        | 51.32 | 28.94        | 28.58         |
|                      | Gemini-2.5-Flash <sup>†</sup>      | ✓         | 63.06        | 48.05 | 15.01        | 69.74        | 51.32 | 18.42        | 16.72         |
|                      | Gemini-2.5-Flash-Lite <sup>†</sup> | ✓         | 64.86        | 37.84 | 27.02        | 60.53        | 43.42 | 17.11        | 22.07         |
| Open-source Models   |                                    |           |              |       |              |              |       |              |               |
| Z.ai                 | GLM-4.6 <sup>†</sup>               | ✓         | 73.87        | 47.15 | 26.72        | 75.00        | 46.05 | 28.95        | 27.84         |
|                      |                                    |           | 67.57        | 38.44 | 29.13        | 68.42        | 44.74 | 23.68        | 26.41         |
| Alibaba              | Qwen3-235B-Thinking <sup>†</sup>   | ✓         | 77.18        | 45.65 | 31.53        | 76.32        | 57.89 | 18.43        | 24.98         |
|                      | Qwen3-30B-Thinking <sup>†</sup>    | ✓         | 71.77        | 34.23 | <u>37.54</u> | 73.68        | 44.74 | <u>28.94</u> | <u>33.24</u>  |
|                      | Qwen3-4B-Thinking <sup>†</sup>     | ✓         | 63.96        | 25.53 | <b>38.43</b> | 65.79        | 46.05 | 19.74        | 29.09         |
|                      | Qwen3-235B-Instruct                |           | 76.88        | 45.05 | 31.83        | 86.84        | 63.13 | 23.71        | 27.77         |
|                      | Qwen3-30B-Instruct                 |           | 70.57        | 35.14 | <b>35.43</b> | 82.89        | 44.74 | <b>38.15</b> | <b>36.79</b>  |
|                      | Qwen3-4B-Instruct                  |           | 58.86        | 24.32 | 34.54        | 65.75        | 39.47 | 26.28        | 30.41         |
| Average $\Delta$     |                                    |           | <b>29.64</b> |       |              | <b>23.68</b> |       |              | <b>26.66</b>  |

Table 11: **Option-removal effect on existing benchmarks.** For each benchmark, CSRQ removes final answer options while preserving objective exact-match scoring. We report  $\Delta = \text{MCQ} - \text{CSRQ}$  to make the effect of removing answer candidates explicit. Across MalayMMLU and CS-EN, models consistently score lower under CSRQ, with an average drop of 26.66 points. Bold indicates the largest degradation, and underlined values indicate the second-largest degradation. <sup>†</sup> denotes reasoning models, and by default, we set the reasoning effort to **medium** where available.

| Family               | Model                              | Reasoning | MMLU         |              |              | HealthBench-Hard |              |              |
|----------------------|------------------------------------|-----------|--------------|--------------|--------------|------------------|--------------|--------------|
|                      |                                    |           | MCQ          | CSRQ         | $\Delta$     | CRQ              | CSRQ         | $\Delta$     |
| Closed-source Models |                                    |           |              |              |              |                  |              |              |
| OpenAI               | GPT-5 <sup>†</sup>                 | ✓         | <b>92.50</b> | <b>84.64</b> | 7.86         | <b>50.84</b>     | <b>83.33</b> | 32.49        |
|                      | GPT-5-mini <sup>†</sup>            | ✓         | <u>90.51</u> | <u>82.30</u> | 8.21         | <u>48.89</u>     | <u>82.51</u> | 33.62        |
|                      | GPT-5-nano <sup>†</sup>            | ✓         | 87.01        | 72.02        | 14.99        | <u>42.44</u>     | 72.95        | 30.51        |
| Google               | Gemini-2.5-Pro <sup>†</sup>        | ✓         | <u>91.06</u> | 78.35        | 12.71        | 31.07            | 75.68        | <u>44.61</u> |
|                      | Gemini-2.5-Flash <sup>†</sup>      | ✓         | 87.61        | 80.02        | 7.59         | 29.82            | <u>78.42</u> | <b>48.60</b> |
|                      | Gemini-2.5-Flash-Lite <sup>†</sup> | ✓         | 86.80        | 74.49        | 12.31        | 22.96            | 59.84        | 36.88        |
| Open-source Models   |                                    |           |              |              |              |                  |              |              |
| Z.ai                 | GLM-4.6 <sup>†</sup>               | ✓         | 84.38        | 80.59        | 3.79         | <b>32.34</b>     | 77.87        | <u>45.53</u> |
|                      |                                    |           | 81.30        | 72.16        | 9.14         | 32.21            | 69.95        | 37.74        |
| Alibaba              | Qwen3-235B-Thinking <sup>†</sup>   | ✓         | 90.39        | 80.90        | 8.61         | 41.18            | <u>78.42</u> | 37.24        |
|                      | Qwen3-4B-Thinking <sup>†</sup>     | ✓         | 81.58        | 64.15        | <b>17.43</b> | 23.06            | 58.20        | 35.14        |
|                      | Qwen3-30B-Thinking <sup>†</sup>    | ✓         | 87.72        | 73.05        | 14.67        | 31.59            | 71.31        | 39.72        |
|                      | Qwen3-235B-Instruct                |           | 89.51        | 81.66        | 7.85         | 31.87            | 77.60        | <b>45.73</b> |
|                      | Qwen3-30B-Instruct                 |           | 85.55        | 73.91        | 11.64        | 27.95            | 71.86        | 43.91        |
|                      | Qwen3-4B-Instruct                  |           | 79.41        | 62.32        | <u>17.09</u> | 22.52            | 59.02        | 36.50        |
| Average $\Delta$     |                                    |           |              |              | <b>10.88</b> | <b>39.25</b>     |              |              |

Table 12: **Performance comparison between original multiple-choice evaluation and CSRQ evaluation across MMLU and HealthBench-Hard.**  $\Delta$  denotes the absolute performance degradation between evaluation settings. Bold indicates the largest degradation, and underlined values indicate the second-largest degradation within each  $\Delta$  column. Qwen3-235B-A22B-Instruct is used as a judge for CRQ. <sup>†</sup> denotes reasoning models, and by default, we set the reasoning effort to **medium** where available.

| Family     | Model                              | Reasoning | MCQs         | CSRQs        |              | CRQs         |              |
|------------|------------------------------------|-----------|--------------|--------------|--------------|--------------|--------------|
|            |                                    |           |              | loose        | strict       | loose        | strict       |
|            | Random Baseline                    |           | 25.53        | 40.33        | 4.28         | -            | -            |
| OpenAI     | GPT-5 <sup>†</sup>                 | ✓         | 88.14        | <b>78.10</b> | <b>56.78</b> | 68.41        | 22.90        |
|            | GPT-5-mini <sup>†</sup>            | ✓         | 84.73        | 77.50        | 55.46        | 67.25        | <b>25.65</b> |
|            | GPT-5-nano <sup>†</sup>            | ✓         | 80.12        | 73.61        | 48.38        | 63.10        | 20.76        |
|            | GPT-4o                             |           | <b>90.28</b> | <u>78.00</u> | <u>56.01</u> | <b>68.61</b> | 21.14        |
|            | GPT-4o-mini                        |           | 78.36        | 74.30        | 49.20        | 65.78        | 20.37        |
| Google     | Gemini-2.5-Pro <sup>†</sup>        | ✓         | 85.56        | 77.78        | 55.68        | 66.87        | 20.26        |
|            | Gemini-2.5-Flash <sup>†</sup>      | ✓         | 82.87        | 71.05        | 47.39        | 65.23        | 21.47        |
|            | Gemini-2.5-Flash-Lite <sup>†</sup> | ✓         | 79.57        | 72.90        | 48.00        | 62.56        | 18.95        |
| Z.ai       | GLM-4.6 <sup>†</sup>               | ✓         | 84.46        | 76.37        | 53.38        | 63.20        | 17.74        |
|            | GLM-4.6 <sup>†</sup>               |           | 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 <sup>†</sup>   | ✓         | 61.01        | 62.62        | 42.12        | 50.13        | 18.51        |
|            | Qwen3-30B-Thinking <sup>†</sup>    | ✓         | 79.57        | 70.49        | 43.82        | 62.48        | 17.35        |
|            | Qwen3-4B-Thinking <sup>†</sup>     | ✓         | 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        |
| SeaLLM     | SeaLLMs-v3-7B                      |           | 29.98        | 52.96        | 26.30        | 47.62        | 7.86         |
|            | SeaLLMs-v3-1.5B*                   |           | 0.00         | 0.16         | 0.00         | 39.54        | 6.45         |
| AI SG      | Qwen-SEA-LION-v4-32B-IT            |           | 83.25        | 76.16        | 53.54        | 62.55        | 18.23        |
| Mesolitica | MaLLaM-1.1B*                       |           | 0.00         | 0.00         | 0.00         | 39.56        | 5.44         |
|            | MaLLaM-5B*                         |           | 0.00         | 0.00         | 0.00         | 46.10        | 6.64         |

Table 13: **Performance of various LLMs on the *MyCulture* benchmark across MCQ, CSRQ, and CRQ tasks.** **Bold** values indicate the best performance, while underlined values denote the second-best. An asterisk (\*) marks instances where the model failed to follow instructions to output the answer in a parseable format, hence unable to be extracted via regular expressions. † denotes reasoning models, and by default, we set the reasoning effort to **medium** where available.

| Family           | Model                              | Multi Select |       |              | Ordering     |       |              | Matching     |       |              | Overall |       |              |
|------------------|------------------------------------|--------------|-------|--------------|--------------|-------|--------------|--------------|-------|--------------|---------|-------|--------------|
|                  |                                    | MCQ          | CSRQ  | $\Delta$     | MCQ          | CSRQ  | $\Delta$     | MCQ          | CSRQ  | $\Delta$     | MCQ     | CSRQ  | $\Delta$     |
| OpenAI           | GPT-5 <sup>†</sup>                 | 89.00        | 68.57 | 20.43        | 89.04        | 50.51 | <u>38.53</u> | 86.03        | 48.23 | 37.80        | 88.14   | 56.78 | 31.36        |
|                  | GPT-5-mini <sup>†</sup>            | 89.71        | 66.71 | 23.00        | 82.53        | 48.63 | 33.90        | 80.63        | 48.23 | 32.40        | 84.73   | 55.46 | 29.27        |
|                  | GPT-5-nano <sup>†</sup>            | 85.57        | 50.57 | 35.00        | 79.97        | 47.43 | 32.54        | 73.18        | 46.55 | 26.63        | 80.12   | 48.38 | 31.74        |
|                  | GPT-4o                             | 92.14        | 71.00 | 21.14        | 91.27        | 47.95 | <b>43.32</b> | 86.78        | 45.25 | <b>41.53</b> | 90.28   | 56.01 | 34.27        |
|                  | GPT-4o-mini                        | 84.71        | 59.14 | 25.57        | 77.57        | 46.40 | 31.17        | 70.95        | 39.29 | 31.66        | 78.36   | 49.20 | 29.16        |
| Google           | Gemini-2.5-Pro <sup>†</sup>        | 90.00        | 67.71 | 22.29        | 82.53        | 48.63 | 33.90        | 83.05        | 47.67 | 35.38        | 85.56   | 55.68 | 29.88        |
|                  | Gemini-2.5-Flash <sup>†</sup>      | 86.71        | 52.43 | 34.28        | 78.94        | 46.40 | 32.54        | 82.12        | 41.90 | <u>40.22</u> | 82.87   | 47.39 | 35.48        |
|                  | Gemini-2.5-Flash-Lite <sup>†</sup> | 82.71        | 52.00 | 30.71        | 76.88        | 47.09 | 29.79        | 78.40        | 43.76 | 34.64        | 79.57   | 48.00 | 31.57        |
| Z.ai             | GLM-4.6                            | 90.43        | 65.57 | 24.86        | 79.11        | 48.63 | 30.48        | 82.50        | 42.64 | 39.86        | 84.46   | 53.38 | 31.08        |
|                  | GLM-4.6 <sup>†</sup>               | 94.14        | 65.43 | 28.71        | 88.53        | 52.40 | 36.13        | 77.28        | 43.95 | 33.33        | 87.37   | 54.91 | 32.46        |
|                  | GLM-4-32B-128K                     | 87.71        | 63.14 | 24.57        | 75.17        | 44.52 | 30.65        | 58.85        | 31.47 | 27.38        | 75.18   | 47.83 | 27.35        |
| Alibaba          | Qwen3-235B-Thinking <sup>†</sup>   | 72.29        | 45.71 | 26.58        | 60.62        | 42.98 | 17.64        | 46.74        | 36.50 | 10.24        | 61.01   | 42.12 | 18.89        |
|                  | Qwen3-30B-Thinking <sup>†</sup>    | 82.14        | 42.00 | 40.14        | 80.82        | 48.63 | 32.19        | 74.86        | 40.97 | 33.89        | 79.57   | 43.82 | <u>35.75</u> |
|                  | Qwen3-4B-Thinking <sup>†</sup>     | 74.29        | 31.43 | 42.86        | 76.20        | 46.40 | 29.80        | 63.13        | 36.31 | 26.82        | 71.61   | 37.67 | 33.94        |
|                  | Qwen3-235B-Instruct                | 88.43        | 62.57 | 25.86        | 81.16        | 46.75 | 34.41        | 68.16        | 41.53 | 26.63        | 80.12   | 51.29 | 28.83        |
|                  | Qwen3-30B-Instruct                 | 88.71        | 63.86 | 24.85        | 80.99        | 48.63 | 32.36        | 75.98        | 41.90 | 34.08        | 82.48   | 52.50 | 29.98        |
|                  | Qwen3-4B-Instruct                  | 81.43        | 38.00 | <b>43.43</b> | 77.57        | 39.55 | 38.02        | 67.41        | 27.75 | 39.66        | 76.06   | 35.48 | <b>40.58</b> |
| SeaLLM           | SeaLLMs-v3-7B                      | 33.43        | 33.86 | 0.43         | 50.51        | 29.97 | 20.54        | 3.17         | 12.48 | 9.31         | 29.98   | 26.30 | 3.68         |
|                  | SeaLLMs-v3-1.5B                    | 0.00         | 0.00  | 0.00         | 0.00         | 0.00  | 0.00         | 0.00         | 0.00  | 0.00         | 0.00    | 0.00  | 0.00         |
| AI SG            | Qwen-SEA-LION-v4-32B               | 90.00        | 66.43 | 23.57        | 80.14        | 49.32 | 30.82        | 77.84        | 41.34 | 36.50        | 83.25   | 53.54 | 29.71        |
| Mesolitica       | MaLLaM-5B                          | 0            | 0     | 0            | 0            | 0     | 0            | 0            | 0     | 0            | 0       | 0     | 0            |
|                  | MaLLaM-1.1B                        | 0            | 0     | 0            | 0            | 0     | 0            | 0            | 0     | 0            | 0       | 0     | 0            |
| Average $\Delta$ |                                    |              |       | 26.41        | <b>32.16</b> |       |              | <u>31.31</u> |       |              | 30.39   |       |              |

Table 14: **Performance comparison between MCQ and CSRQ settings across Multi-Select, Ordering, Matching, and overall evaluation on MyCulture.**  $\Delta$  denotes absolute performance degradation from MCQ to CSRQ. Within each  $\Delta$  column, **bold** indicates the largest degradation and underlined values indicate the second-largest. <sup>†</sup> denotes reasoning models, and by default, we set the reasoning effort to **medium** where available.