EMNLP 2026

Between Multiple Choice and Open Response

Evaluating LLMs with Option-Free Deterministic Scoring

Zhong Ken Hew* Sze Jue Yang* Chee Seng Chan†

Universiti Malaya, Malaysia

* Equal contribution † Corresponding author

Problem: options and judges shape scores

Format	Answer elicitation	Scoring
MCQ	Visible final options	Deterministic
CRQ	Open-ended response	Human or LLM judge

Answer-option scaffolding

Options can support recognition, elimination and partial knowledge.

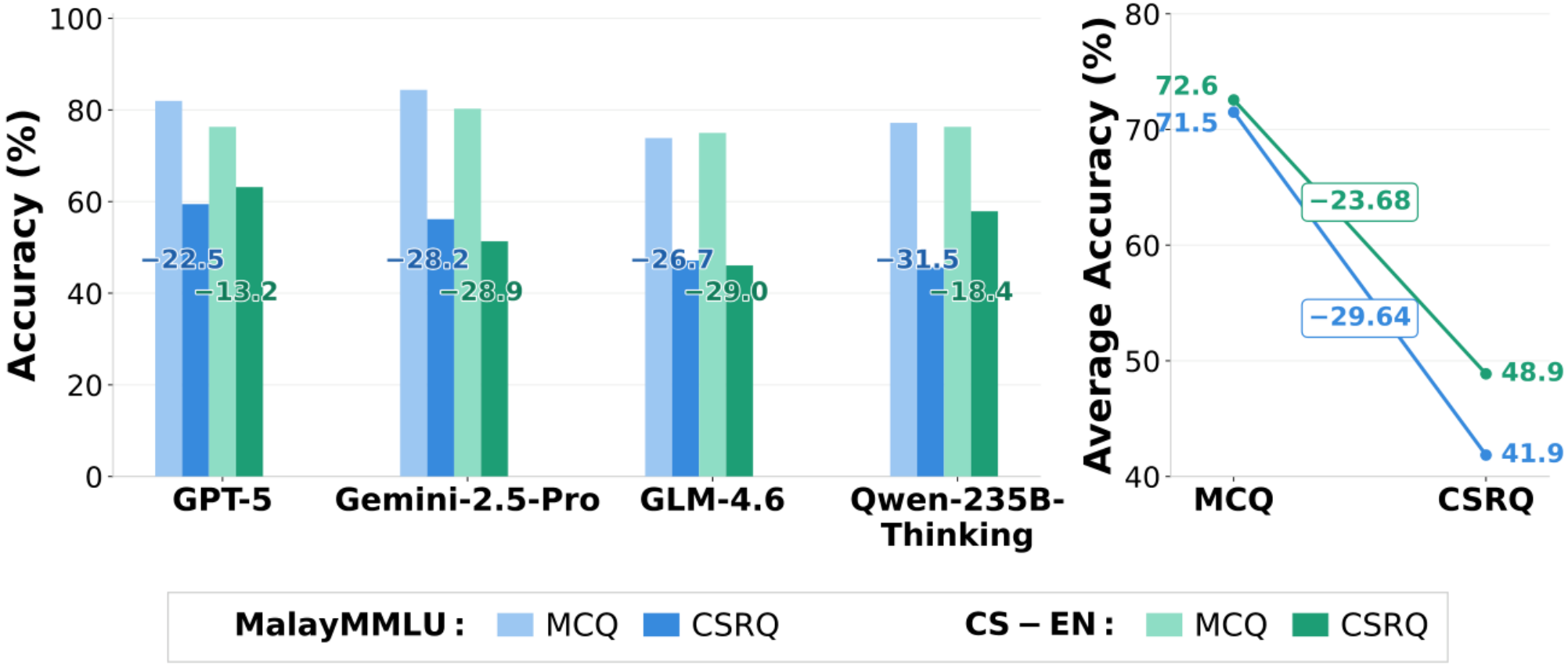
Judge-dependent scoring

Scores can change with the judge and rubric interpretation.

Removing answer options can introduce judge-dependent scoring.

Problem evidence: removing final options

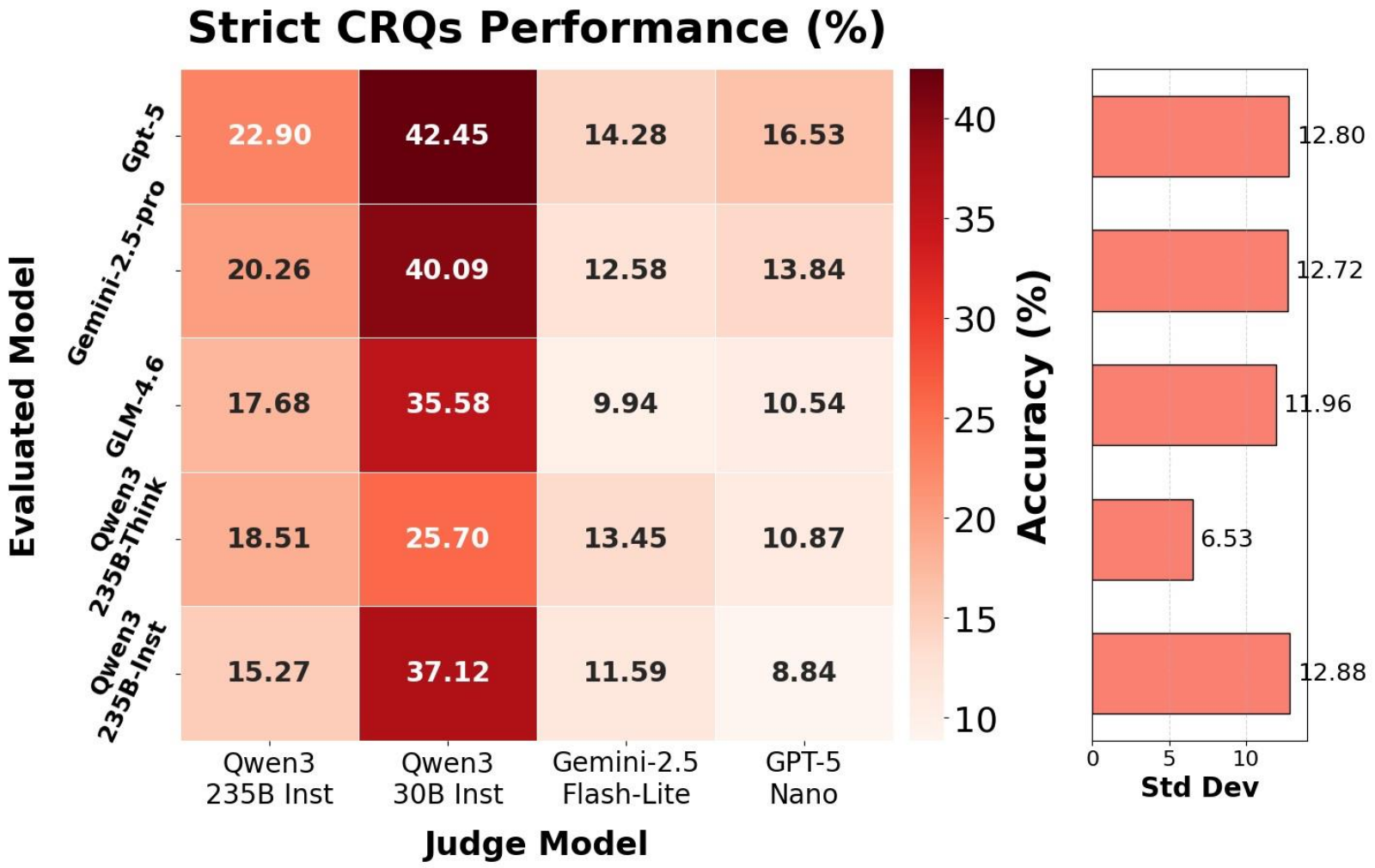
Existing multi-answer items, with deterministic scoring in both formats



Average decrease: MalayMMLU 29.64 pp CS-EN 23.68 pp

Figure 2 and Table 11. CS-EN = English CS-Bench subset. pp = percentage points.

Problem evidence: the judge changes the score



Same CRQ responses
Same rubric
Different judges

28.17 pp

GPT-5 score range
14.28% to 42.45%

Evidence of judge-dependent scoring

Figure 7. Rows: evaluated models. Columns: judges. MyCulture strict CRQ scores.

Research question and contributions

Can we remove final answer options while retaining deterministic scoring?

1	**CSRQ protocol**	Construct structured answers with exact verification
2	**CSRQ converter**	Build suitable structured items from existing benchmarks
3	**MyCulture**	Compare matched formats on Malaysian cultural knowledge

Contribution 1: CSRQ protocol

Variant	Model constructs	Illustrative output
Multi-select	A subset satisfying a condition	{1, 3}
Ordering	A full ordered sequence	(2, 4, 1, 3)
Matching	Pairs between two sets	{(A3, B2), (A1, B3), (A2, B1)}

Provided descriptions define the task.

Candidate final subsets, sequences and pairings remain hidden.

Section 2.2. Illustrative outputs use the paper's formal examples.

CSRQ keeps statements, removes final options

Lansaran, a traditional game of the Murut community

MCQ

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

Statements:
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.

Options: **A.** II, III **B.** I, IV **C.** I, II, IV **D.** II, IV

Answer: A

CSRQ

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

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

Answer: B, C

CRQ

Question:
Huraikan ciri-ciri fizikal, cara pembinaan dan peranan komuniti dalam permainan tradisional Lansaran suku kaum Murut berdasarkan amalan tradisi yang dinyatakan.

Rubric:
1. Tempat permainan Lansaran berukuran antara lapan (8) hingga 30 kaki persegi.
2. Lansaran dapat menampung lebih kurang 30 orang pemain.
3. Lansaran dibina secara bergotong-royong di ruang khas rumah panjang suku kaum Murut.
4. Beberapa batang kayu bulat yang keras dan liat sebesar lengan orang dewasa dikenali sebagai kayu "Selangan Batu".
5. Kayu "Selangan Batu" diikat dengan rotan saga bagi menampung lantai Lansaran.
6. Lantai Lansaran dibina sekaki ke bawah daripada paras lantai rumah secara tergantung.

MCQ selects a final combination. CSRQ constructs the subset: B, C.

Figure 5. Original Bahasa Malaysia example. English versions in Appendix C.

Deterministic scoring and chance success

Strict scoring

1 for an exact match after canonicalization, otherwise 0

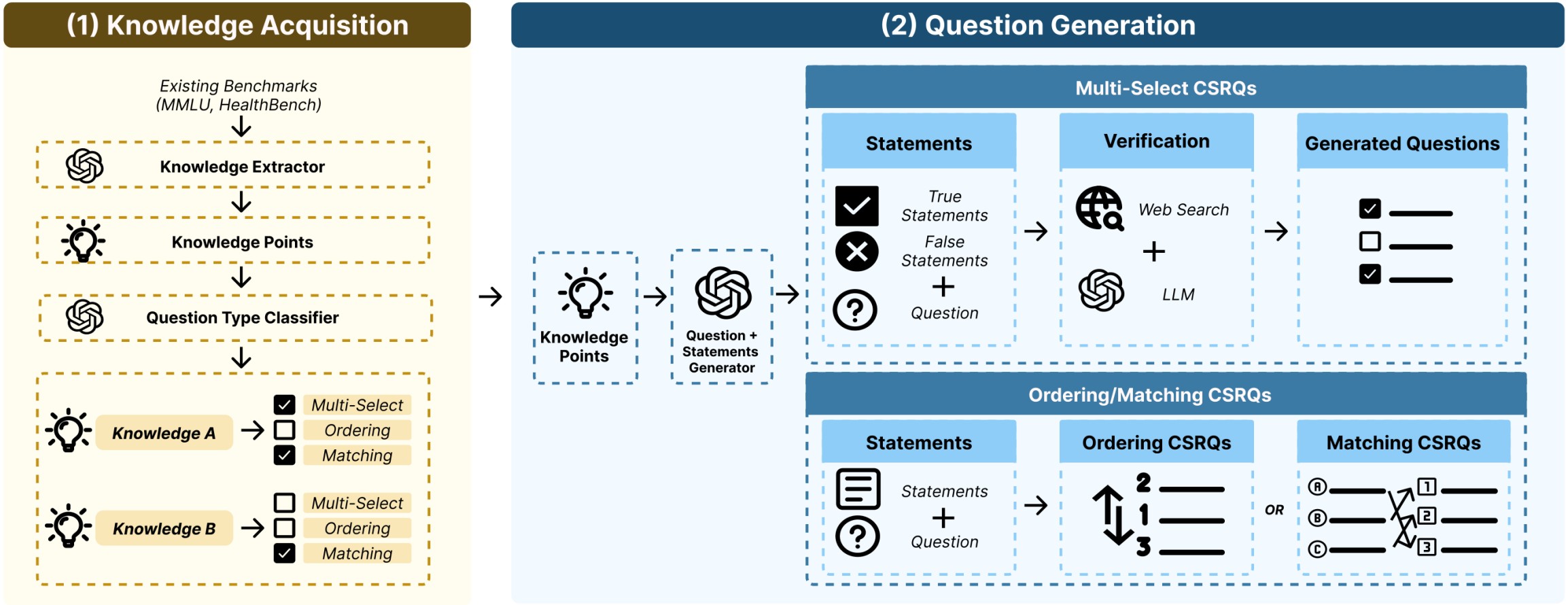
Canonicalization normalizes equivalent surface forms before comparison.

Uniform random guess	Probability	Four-unit example
MCQ with K choices	1 / K	25.00%
Multi-select with M statements	1 / (2^M − 1)	6.67%
Full ordering or bijective matching	1 / M!	4.17%

Chance-success baselines do not measure deeper understanding.

Sections 2.3 and 2.5. Multi-select includes all non-empty subsets.

Contribution 2: CSRQ converter



Suitable items require generation, filtering and independent verification.

Figure 3 and Section 3.2.1

Converted benchmarks show format sensitivity

MMLU accuracy (%)

Model	MCQ	CSRQ
GPT-5	92.50	84.64
Gemini-2.5-Pro	91.06	78.35
GLM-4.6	84.38	80.59
Qwen3-235B	90.39	80.90

HealthBench-Hard score (%)

Model	MCQ	CSRQ	CRQ
GPT-5	96.17	83.33	50.84
Gemini-2.5-Pro	93.99	75.68	31.07
GLM-4.6	95.63	77.87	32.34
Qwen3-235B	95.08	78.42	41.18

Scores change when answer exposure and scoring change.

The ordering does not establish a universally more valid format.

Table 1. Reasoning models. HealthBench CRQ judge: Qwen3-235B-Instruct.

Contribution 3: MyCulture

Matched MCQ, CSRQ and CRQ versions of the same cultural knowledge

5,463

total question instances

1,821 per format
Bahasa Malaysia

Six cultural dimensions

Arts and attire
Customary practices and entertainment
Food and religions

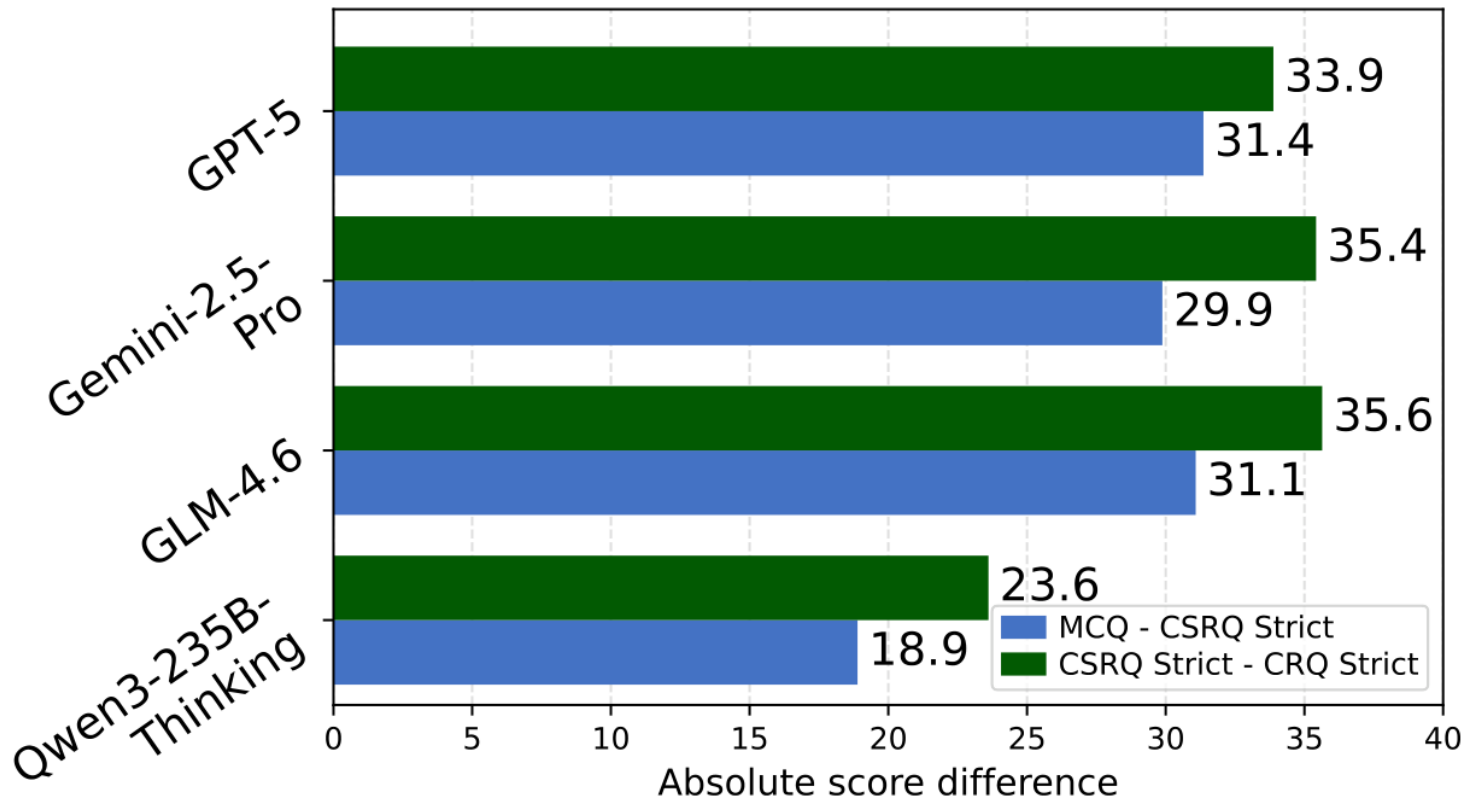
Qualified Malaysian residents
audit 5% of the dataset

Model evaluation: closed-book, tool-free

Section 3.3 and Appendix I. Construction uses reliable cultural and academic sources.

MyCulture scores vary across formats

Absolute score gaps under strict scoring



GPT-5 scores (%)

MCQ	**88.14**
CSRQ strict	**56.78**
CRQ strict	**22.90**

31.36 pp

MCQ-to-CSRQ decrease

Strict scoring requires full correctness. Loose scoring awards partial credit.

Figure 6, Table 4 and Appendix F. CRQ judge: Qwen3-235B-Instruct.

Rankings can remain similar as scores change

Model-rank comparison	Spearman ρ	Kendall τ
MCQ vs CSRQ strict	0.95	0.86
MCQ vs CRQ strict	0.74	0.50
CSRQ strict vs CRQ strict	0.79	0.64

High rank agreement can coexist with large absolute score gaps.

Table 6. Correlations compare model rankings on MyCulture.

Error analysis and construction robustness

Sampled incorrect CSRQ responses

88.75%

Answer-content errors

11.25%

Output-format errors

Construction-model check

Original GLM-4.6 pipeline
Gemini-2.5-Pro
GPT-5.4

GPT-5 outperforms GLM-4.6
on both matched benchmark subsets
under all three pipelines.

120 incorrect responses per model, balanced across MS, ordering and matching

Tables 2 and 3. Sampled errors and matched subsets limit the scope of these checks.

Practical use and limitations

Instrument	Use in practice
MCQ	Low-cost screening and reproducible comparisons
CSRQ	Matched audits of suitable MCQ items
CRQ	Expressive answers with judge sensitivity analysis

CSRQ limitations

Requires decomposable answers and a finite structured response space.

Adds construction and verification costs.

No claim to prove genuine understanding or replace all benchmarks.

A benchmark score depends on its instrument

Model capability and the test both shape measured performance.

CSRQ removes final answer options
while preserving deterministic scoring.

github.com/HewZK2000/MyCulture-CSRQ