Supplementary Material for The Label Is Not the Answer: Output-Schema Anchoring in Medical Multimodal MCQ Evaluation

Chengyi Zhang[1], Yitong Bai[2], and Ziyang Wang[3⋆]

[1] Department of Mechanical Engineering, Swansea University, UK
[2] The Fu Foundation School of Engineering and Applied Science, Columbia University, USA
[3] School of Computer Science and Digital Technologies, Aston University, UK
977671@swansea.ac.uk; yb2636@columbia.edu; z.wang47@aston.ac.uk

This document provides the reproducibility details that complement the main paper. It contains the verbatim prompt templates, deterministic sampling procedure, statistical specifications, parser precedence, and selected complete model outputs. The configuration names match those used in the paper. Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct are abbreviated as Qwen-3B and Qwen-7B, respectively. Unless noted otherwise, the image, question, option text and order, model weights, decoding settings, and parser were held fixed within each comparison.

1 Complete Prompt Templates

Placeholders enclosed in braces were replaced at run time. Options were rendered in their fixed display order, one candidate per line, from `A. text` to `E. text`. In the cyclic experiments, only the label attached to each option text was reassigned. No preceding assistant response or worked example was supplied outside the templates shown below.

1.1 System message

The Qwen-3B system message was shared by all configurations:

```
You are a medical visual reasoning assistant for the MedReason Challenge.
Use the provided image evidence and the question. Do not invent findings
↪ that
are not supported by the image. If the evidence is insufficient, state
↪ the
limitation clearly.
Return only valid JSON with two keys: reasoning_trace and answer.
```

⋆ Corresponding author.

1.2 Submitted prompt and five-label intervention

The submitted MCQ template was:

```
Task: closed-ended multiple-choice medical visual question answering.

Question:
{question}

Options:
{options}

Instructions:
1. Inspect the image evidence relevant to the question.
2. Choose exactly one option label from the provided options.
3. The answer field must contain only the selected label, for example
↪  "A".
4. The reasoning_trace should be concise and image-grounded.

Return only JSON:
{
  "reasoning_trace": "brief visual rationale",
  "answer": "A"
}
```

For intervention condition $a \in \{A, B, C, D, E\}$, both quoted occurrences of `A` were replaced by a. No other character changed. Each of the five intervention prompts can therefore be recovered from the submitted template by this substitution alone; Table S1 summarises the character-level comparison.

Table S1. Character-level comparison of the completed-example prompts.

Condition	Prose example	JSON answer	Other changed characters
E_A	`"A"`	`"A"`	0
E_B	`"B"`	`"B"`	0
E_C	`"C"`	`"C"`	0
E_D	`"D"`	`"D"`	0
E_E	`"E"`	`"E"`	0

1.3 Comparative prompt

```
Task: closed-ended multiple-choice medical visual question answering.

Question:
{question}

Options:
{options}
```

```
Instructions:
1. Inspect the image evidence relevant to the question before choosing.
2. Compare all option labels one by one. Do not choose the first option
↪  by default.
3. Pay close attention to negation, location, extent, laterality,
↪  layer/depth,
   temporal wording, and therapy-related wording.
4. Choose exactly one option label from the provided options.
5. The answer field must contain only the selected option label, such as
↪  B or D.
6. The reasoning_trace should briefly state why the chosen option matches
↪  the
   image better than the closest alternatives.
7. If uncertain, still choose the best-supported option after comparing
↪  all
   options; do not fall back to A.

Return only valid JSON:
{
  "reasoning_trace": "brief visual rationale comparing the selected
  ↪  option
                     against close alternatives",
  "answer": "selected option label only"
}
```

1.4 Cue-free prompt

Before insertion into this template, a duplicated leading label such as `A)`, `A.`, `A:`, or `A-` is removed from each option text. The renderer then adds the current assignment label once.

```
Task: closed-ended medical visual question answering.

Question:
{question}

Candidate answers:
{options}

Inspect the image, compare the candidate answers, and select the single
↪  candidate
best supported by the visual evidence. Pay attention to negation,
↪  location,
extent, laterality, depth, and temporal wording.

Return only valid JSON using this schema:
{
  "reasoning_trace": "brief image-grounded rationale",
  "answer": "selected candidate label"
}
```

1.5 Cue–prefix factorial control

The cue-present condition with original prefixes used the Submitted template. For its clean-prefix counterpart, duplicated leading labels were removed from the option text. Both cue-absent conditions used the following unfilled template, with either original or cleaned option text:

```
Task: closed-ended multiple-choice medical visual question answering.

Question:
{question}

Options:
{options}

Instructions:
1. Inspect the image evidence relevant to the question.
2. Choose exactly one option label from the provided options.
3. The answer field must contain only the selected label.
4. The reasoning_trace should be concise and image-grounded.

Return only JSON:
{
  "reasoning_trace": "brief visual rationale",
  "answer": "selected option label"
}
```

1.6 InternVL3-9B prompt

InternVL3-9B received `<image>` followed by the system text and task template below. The concrete label a was inserted in the two indicated fields.

```
You are a medical visual reasoning assistant for the MedReason Challenge. Use the
provided image evidence and the question. Do not invent findings that are not
supported by the image. If the evidence is insufficient, state the limitation
clearly. Return only valid JSON with two keys in this order: answer and
reasoning_trace.

Task: closed-ended multiple-choice medical visual question answering.

Question:
{question}

Options:
{options}

Instructions:
1. Inspect the image evidence relevant to the question.
2. Choose exactly one option label from the provided options.
3. Put the answer field first. It must contain only the selected label,
   for example "{a}".
4. The reasoning_trace should be concise and image-grounded.

Return only JSON:
{
  "answer": "{a}",
  "reasoning_trace": "brief visual rationale"
}
```

2 Sampling and Analysis

2.1 Sample construction and study sequence

The participant package contained 20,654 records before exclusion of the 400 public-leaderboard cases and 20,254 afterwards. We retained five-option MCQs and stratified them by reference label. Within each stratum, records were ordered by the hexadecimal SHA-256 digest of `seed:case_id`; the first 120 records were selected, giving 600 cases with equal representation of A–E. The fixed seeds were:

```
primary: perfect-wave-medreason-2026-v1
holdout: perfect-wave-medreason-2026-replication-v1
```

The holdout selection excluded all primary case IDs before applying the same stratified procedure.

Prompt development began after organiser-side pre-evaluation drew attention to the difference between the MCQ and open-ended results. The first 200 primary cases were used to examine the completed-example effect and to develop the Comparative and Cue-free prompts. Their wording and the analysis procedure were then fixed for the remaining 400 primary cases. The non-overlapping 600-case sample was selected afterwards and used as a same-source holdout for the three configurations. Hidden challenge cases were not used for prompt development or analysis.

2.2 Agreement measures

Cyclic agreement was calculated after mapping each emitted label back to the option text denoted under that assignment. All-five agreement is the proportion of cases for which the five reverse-mapped outputs are identical. Pairwise agreement is first averaged over the ten unordered pairs of assignments within each case and then across cases. The distinct-option statistic counts the unique reverse-mapped outputs among the five assignments. Table S2 reports all three measures.

Table S2. Agreement under cyclic label reassignment.

Prompt	All-five (%)	Pairwise (%)	Mean distinct options
Submitted	0.00	0.48	4.96
Comparative	14.33	45.03	2.39
Cue-free	15.00	44.43	2.43

2.3 Statistical procedures

For the completed-example intervention, the paired Monte Carlo null permuted the five observed responses within each case. Pearson's χ^2 was recalculated from

the resulting cue–response table for each of 10,000 permutations. The reported value uses the plus-one correction, $(b + 1)/(B + 1)$, where b is the number of permuted statistics at least as large as the observed statistic and $B = 10{,}000$. No permuted Qwen-3B or InternVL3-9B statistic reached the observed value, giving $p = 1/10001$ in each experiment. Cramér's V was calculated from the observed table.

For a cyclic comparison, each case contributed the difference between the two prompts in its number of correct responses across the five assignments. The exact two-sided sign-flip test was applied to the non-zero case differences. Percentage-point differences were calculated over all 600×5 outputs, and 95% confidence intervals were obtained from 200,000 case-cluster bootstrap samples (random seed 20260807). Canonical-order comparisons used exact two-sided McNemar tests on paired cases. The three prompt comparisons within each sample were adjusted by Holm's procedure. Wilson intervals were used for individual proportions. The three contrasts in the 2×2 cue–prefix control were prespecified as descriptive checks and were not adjusted for multiplicity.

2.4 Completed-example response tables

Table S3 gives the exact counts underlying the cue-following results. Rows denote the label shown in the completed example and columns the reported answer. Each Qwen-3B row contains 600 cases and each InternVL3-9B row 200 cases.

Table S3. Cue–response counts for the completed-example intervention.

Model	Cue	A	B	C	D	E
Qwen-3B	A	597	1	1	0	1
	B	1	598	0	0	1
	C	0	1	595	0	4
	D	2	2	2	588	6
	E	2	0	3	6	589
InternVL3-9B	A	178	6	4	5	7
	B	0	196	2	1	1
	C	1	2	195	1	1
	D	0	1	2	196	1
	E	1	0	1	1	197

For Qwen-3B, 569 cases produced all five labels, 29 produced four, and two produced three; the mean was 4.945 distinct labels per case. InternVL3-9B produced a mean of 4.81 distinct labels, and 175 of 200 cases produced all five labels. Every case changed under at least one completed-example label in both experiments.

3 Output Parsing

For Qwen-3B, the generated continuation was stripped of Markdown fences. The parser then: (1) attempted to decode the full continuation as a JSON object; (2) attempted to decode the substring between the outermost braces; (3) read `answer` or `final_answer`; (4) searched for an explicit "answer", "option", or "choice" expression; (5) accepted a direct valid label; and (6) matched an exact option text. If all routes failed, the submitted code returned the first option. Replacing this final fallback with the Comparative fallback changed none of the 600 canonical outputs, indicating that the last-resort fallback did not account for the observed concentration on A in that sample. The submitted parser did not record the successful route for each Qwen-3B response, so route-specific frequencies cannot be reconstructed.

The InternVL3-9B parser required an explicit JSON answer, an answer-labelled expression, or an unambiguous exact option-text match; it had no first-label fallback. All 1,000 intervention outputs were parsed directly from JSON.

4 Illustrative Model Outputs

Table S4 reproduces all stored fields for two Qwen-3B cases from the completed-example intervention. In the first, the rationale is unchanged while the JSON `answer` follows the example label. In the second, both fields vary with the example. These cases illustrate two observed output patterns; the quantitative analysis is based on the reported answer field.

Table S4. Complete stored outputs for two illustrative Qwen-3B cases.

Case	Cue	Answer	Reasoning trace
MCQ_9905	A	A	The image shows an irregular, non-laminated texture without red thrombus layers, which suggests a non-vascular inflammatory pseudotumor rather than a vascular mechanism.
	B	B	The image shows an irregular, non-laminated texture without red thrombus layers, which suggests a non-vascular inflammatory pseudotumor rather than a vascular mechanism.
	C	C	The image shows an irregular, non-laminated texture without red thrombus layers, which suggests a non-vascular inflammatory pseudotumor rather than a vascular mechanism.
	D	D	The image shows an irregular, non-laminated texture without red thrombus layers, which suggests a non-vascular inflammatory pseudotumor rather than a vascular mechanism.
	E	E	The image shows an irregular, non-laminated texture without red thrombus layers, which suggests a non-vascular inflammatory pseudotumor rather than a vascular mechanism.
MCQ_12306	A	A	The left wrist appears to have more swelling compared to the right wrist.
	B	B	The right thumb appears to have a more pronounced swelling compared to the left thumb.
	C	C	The right hand appears to have more swelling compared to the left hand.
	D	D	The right wrist appears to have more swelling compared to the left wrist.
	E	E	The left hand appears to have more swelling compared to the right hand.