Answer Readout as a Major Confounder in Medical Multiple-Choice Visual Question Answering

Ura Modi[1]* and Mehul S. Raval[1]

Ahmedabad University, Ahmedabad, India
ura.m@ahduni.edu.in, mehul.raval@ahduni.edu.in

Abstract. We study answer readout in a 7B vision-language model adapted with DoRA on the official Med-CMR training split. The submitted system used length-normalised isolated option-text likelihood for multiple-choice questions. In a post-submission analysis on 325 held-out items from the same distribution, switching to joint option presentation and next-token option-symbol scoring increased accuracy from 0.643 to 0.954 without changing weights. This contrast bundles option presentation, prediction target, and scoring; it is not a component ablation. A model-free shortest-option baseline achieved 0.899. The untuned base model achieved approximately 0.33 under symbol scoring; where the shortest baseline failed, the adapted model answered 24/33 correctly. Thus, adaptation and more than option brevity are involved, but visual grounding is not isolated from text cues. A margin-gated hybrid reached 0.960 in-sample but averaged 0.954 under repeated five-fold cross-validation, providing no reliable advantage over symbol scoring. The analysis identifies inference configuration as a material evaluation and explains a limitation of our submitted system; it is not an official final challenge result.

Keywords: Medical VQA · Answer readout · Evaluation methodology · Vision-language models

1 Introduction

The MedReason 2026 challenge [10] evaluates medical multimodal reasoning on Med-CMR [5], a visual question answering (VQA) benchmark spanning 11 organ systems and 12 imaging modalities, and extends it to out-of-distribution (OOD) acquisitions — low-dose X-ray and 7T brain MRI. Submissions are self-contained offline containers scored on three metrics: multiple-choice accuracy (MCQ), open-ended ground-truth correctness (GT), and open-ended visual accuracy (VA). GT and VA are organiser-judged scores on a 0–4 scale. Hidden public and private results contribute equally to each final metric, and placement is the mean of the three per-metric ranks.

* Corresponding author.

Our organiser-side pre-evaluation indicated that MCQ was the weakest of our three metrics. We initially attributed this result to limited model capacity, but a subsequent local audit showed that the inference configuration was a major determinant of measured accuracy. To preserve blind review, we omit the complete public score tuple and team rank; all controlled comparisons below use a local hold-out excluded from fine-tuning.

The observed MCQ score reflected an interaction between the adapted model and its answer readout. On our hold-out, the correct option is shortest in 89.9% of cases, exposing a strong construction regularity. The submitted isolated option-text scorer and a joint option-symbol scorer interact differently with this regularity and with the supervised output format. Symbol scoring over the full option list [14] raised held-out accuracy from 0.643 to 0.954. The two configurations differ simultaneously in option presentation, target, and scoring, so the difference cannot be assigned to one component. We contribute (i) a controlled inference-configuration comparison with fixed cases and model weights (Sec. 5); (ii) base-model, label-distribution, and shortest-baseline residual controls (Sec. 6); and (iii) negative results showing that an in-sample gate advantage does not persist under repeated cross-validation (Sec. 7). These experiments diagnose answer readout; they do not establish visual grounding or robustness to the hidden OOD data. This paper's contribution is the post-submission analysis permitted under the track's guidelines; Section 3 documents the submitted system itself.

2 Related Work

Multiple-choice readout. That the mapping from a language model to a multiple-choice answer is a consequential design choice is established outside medicine. Cloze-style scoring, in which each option's text is scored conditioned on the question alone, was standard in early few-shot evaluation [3,8] and requires a normalisation choice, since raw sequence log-likelihood favours short options. Robinson and Wingate [14] showed that presenting all options together and emitting the associated *symbol* — multiple-choice symbol binding — lets the model compare options explicitly and removes dependence on option tokenisation and length. Leaderboard audits find MCQ scores moving by many points under cosmetic scoring changes [1], and harnesses now expose normalisation as a reported option [4]. Our contribution is a measurement of what the choice costs in a blind medical challenge, with controls on plausible alternative explanations.

Construction shortcuts and medical VQA. Systematic regularities that let a label be predicted without the intended input are a recurring benchmark failure mode [7,13]; our shortest-option statistic is that phenomenon in a medical VQA set. We treat it as a diagnostic rather than a trick: because it is available to *any* readout, sensitivity to it helps diagnose how scoring interacts with the dataset construction. We build on Qwen2.5-VL [2] with DoRA [11], a weight-decomposed variant of LoRA [9].

3 System

Base model and adaptation. A single model answers both question types. The base is Qwen2.5-VL-7B-Instruct [2] (Apache-2.0). We train a DoRA [11] adapter for one epoch on 17,322 examples from the official Med-CMR training split, excluding the local hold-out. Rank is 16, α=32, dropout is 0.05, and the adapted language-model modules are `q/k/v/o_proj` and `gate/up/down_proj`; the vision tower is frozen. Training uses bf16, effective batch 8, AdamW with learning rate 10^{-4}, cosine scheduling, 3% warm-up, maximum sequence length 2,048, seed 42, and an image-area budget of 802,816 pixels, matching inference. The saved final-epoch adapter is used without checkpoint selection on the hold-out. Targets are the dataset-provided answer and `visual_description`; the latter supervises the `reasoning_trace`. This provenance does not guarantee causal faithfulness or visual grounding of generated traces, which we do not directly evaluate. We use no additional training dataset, closed-model distillation, or team-supplied PMC-derived asset. We did not access hidden evaluation images or labels; possible source overlap inherited from the benchmark or the base model's pretraining corpus was not assessed.

Four readouts. Let a case have question q, images I, and options $\{(\ell, t_\ell)\}$ with symbols $\ell \in \{A, \ldots, E\}$ and texts t_ℓ. **len-norm** (used in the submitted container) takes $\arg\max_\ell \frac{1}{|t_\ell|} \log p(t_\ell \mid q, I)$, scoring each option in isolation and dividing by token count. **raw-sum** takes $\arg\max_\ell \log p(t_\ell \mid q, I)$ — the same isolated scoring without normalisation, so shorter options are systematically favoured. **shortest** takes $\arg\min_\ell |t_\ell|$ in characters, a model-free prior using no image and no weights, which we use both as a construction diagnostic and as the container fallback. **direct-label** renders all options jointly and asks for the letter alone. Per label and item, it takes the larger next-token log-probability of the bare or space-prefixed form's first token, then the argmax across labels. This fixed rule never uses correctness to choose a variant; it is symbol binding [14], and also exactly the task the adapter was trained on. **gate** (evaluated, not shipped) keeps the RAW-SUM prediction when its top-two log-probability margin exceeds τ and defers to DIRECT-LABEL otherwise. DIRECT-LABEL needs one forward pass per case against one per option for the isolated scorers, so it is also $\approx 5\times$ cheaper.

Fault-tolerant output completion. Because missing or invalid predictions receive no credit, the runner pre-seeds a schema-valid answer for every case before model loading (MCQ uses the SHORTEST prior), guards setup, writes atomically after every case, and can stop model calls at a configured wall-clock budget while retaining fallbacks. In a 400-case fault-injection test, a simulated setup failure still left 400/400 schema-valid outputs; the 325 MCQ fallbacks scored 0.8985 locally. This mechanism improves output completeness, not clinical safety: the open-ended fallback is a deterministic placeholder and receives no claim of correctness.

Table 1. MCQ accuracy on the same 325 items and weights. LEN-NORM is submitted; post-submission DIRECT-LABEL changes option presentation, target, and scoring jointly. SHORTEST uses no model.

Readout	Uses	Base model	Fine-tuned (v2)
LEN-NORM	option text, isolated, length-normalised	0.431	0.643
SHORTEST	option length only (no model)	0.899	
RAW-SUM	option text, isolated, unnormalised	0.920	0.908
GATE(τ=8)	RAW-SUM margin, else DIRECT-LABEL	—	0.960[†]
direct-label all options jointly, symbol logprob		≈ 0.33[‡]	**0.954**

[†]In-sample only; does not survive cross-validation (Sec. 7.1). Not shipped.
[‡]Approximate run-level aggregate; per-case base predictions were not retained.

4 Experimental Setup

Data and judges. We evaluate on a stratified hold-out of 400 cases (325 MCQ, 75 open-ended) drawn from the official training split with seed 42 and excluded from all fine-tuning. It was excluded from adapter training, but shares the source distribution and construction process of the training data — the central limitation of every number here (Sec. 9). MCQ accuracy is exact and needs no judge. For open-ended metrics we run local proxies for the organiser-side judges: a 4-bit Llama-3.1-70B-Instruct checkpoint for GT [6] and Qwen2.5-VL-7B for VA. These match the published model families but not the undisclosed organiser prompts or infrastructure, so open-ended values are used only for paired local comparisons. A paired bootstrap gives an estimated GT noise scale of $\approx \pm 0.3$ at n=75; changes below this scale are not resolved by the small hold-out. For the large-set comparisons (n=600 open and n=1,200 MCQ), the estimated GT noise scale is $\approx \pm 0.11$, while the MCQ sampling scale is $\approx \pm 1.4$ percentage points.

Statistics and compute. All MCQ statistics are recomputed from a frozen per-case artifact recording, for every case, the gold label, each readout's prediction, and the RAW-SUM top-two margin. Because the margin is stored, GATE(τ) is reconstructable at any τ, so the threshold-selection experiment is exactly reproducible on CPU. We use exact McNemar tests [12] and a 20,000-sample paired bootstrap. Fine-tuning used one H100-class GPU for ≈ 3.7 h and readout experiments ≈ 15 GPU-hours on rented A40 and L4 instances; the analyses reported here are CPU-only.

5 Results: the Readout Ladder

Table 1 is the central result: the 325 held-out cases and model weights are fixed, while prompt option presentation, prediction target, and scoring rule change together.

Inference configuration strongly affects local MCQ measurement. The adapted model under LEN-NORM scores 0.6431 locally; the same weights under DIRECT-LABEL score 0.9538. The observed difference is 31.08 points, but it bundles option presentation, target, and scoring and does not identify their individual effects. It must not be combined with an organiser-side value measured on a different hidden set, and it is not an official final challenge result. If anything, fine-tuning slightly lowers RAW-SUM accuracy, so the large direct-label gain cannot be explained by a general improvement in isolated-option scoring.

Length normalisation was mismatched. LEN-NORM (0.643) loses to RAW-SUM (0.908), the same computation without the division, and to SHORTEST (0.899), which never consults the model. A readout that loses to a character-counting prior provides a weak measure of the capability intended by the task. The mechanism is visible in the data: the gold option is shortest by character count in 89.9% of cases, and token normalisation interacts adversely with this brevity regularity. Character and token length are correlated but not identical, so the comparison does not show removal of precisely the same signal.

Comparison with brevity-driven baselines. DIRECT-LABEL exceeds both brevity-driven readouts by margins not attributable to chance: against RAW-SUM it fixes 21 cases and breaks 6 (net +15, exact McNemar $p = 5.9 \times 10^{-3}$); against SHORTEST it fixes 24 and breaks 6 (net +18, $p = 1.4 \times 10^{-3}$).

6 Controls: Learned Skill or Benchmark Shortcut?

A 31-point difference between inference configurations invites the suspicion that the result is entirely a scoring or construction artifact. Three controls narrow, but do not eliminate, that concern.

Base-model control. Applying DIRECT-LABEL to the *untuned* Qwen2.5-VL-7B, with the same prompt and scoring code, gave an approximate run-level aggregate of 0.33 — far below the fine-tuned 0.954. Per-case predictions were not retained, precluding an exact numerator or paired test. Position or tokenisation alone is therefore insufficient to explain the adapted result, though it may contribute. The complementary asymmetry in Table 1: RAW-SUM is similar for the base and adapted models (0.920 vs. 0.908), whereas the DIRECT-LABEL difference is adaptation-dependent. This control does not distinguish medical-image learning from question- or option-pattern learning.

Distribution control. A degenerate position-exploiting predictor would show a collapsed output distribution. Gold counts are A:75, B:73, C:74, D:50, and E:53; DIRECT-LABEL prediction counts are A:70, B:73, C:77, D:52, and E:53. This is a close match and contrasts with an early 3B baseline that answered "A" on 98% of cases.

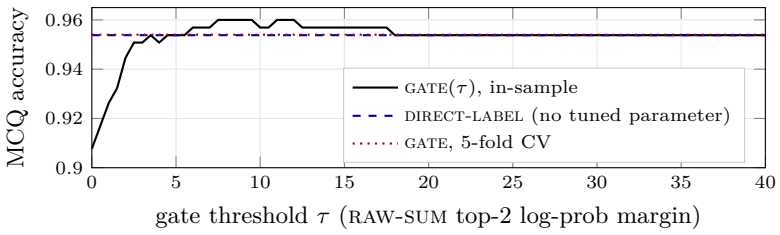

Fig. 1. The gate's in-sample advantage does not persist under repeated five-fold cross-validation. The reported ± 0.0022 is the standard deviation across 50 repetitions.

Shortcut-residual control. If DIRECT-LABEL were only a smoother reading of the brevity shortcut, it should fail wherever the shortcut fails. On the 33 cases where the shortest option is *not* correct, DIRECT-LABEL answers 24 correctly (0.727) against 15 for RAW-SUM. The rescued cases include aortic-arch laterality, venous pathway tracing, and near-identical option texts. This inspection was diagnostic rather than a blinded annotation study and does not establish that the decisions were image-grounded.

The model-free shortest-option rule captures a construction regularity worth 0.899 accuracy on this hold-out. DIRECT-LABEL adds 0.055 absolute accuracy beyond that baseline. The residual gain is not explained solely by choosing the shortest option, but the present experiments do not separate image evidence from question- and option-text cues. Cross-model results reported by Med-CMR [5] use different training and evaluation protocols and are therefore not direct comparators for this controlled readout study.

7 Negative Results

7.1 A Tuned Gate Does Not Survive Cross-Validation

Our best in-sample number was not DIRECT-LABEL. The GATE hybrid — keep the RAW-SUM answer when its top-two margin is confident, defer to DIRECT-LABEL otherwise — reaches 0.9600 at $\tau \in [7.5, 12]$ (312/325), two additional correct predictions relative to DIRECT-LABEL (310/325).

We repeated five-fold cross-validation 50 times, selecting τ on training folds only and scoring the held-out fold. Mean held-out accuracy was 0.9540, with standard deviation 0.0022 across repetitions. A paired bootstrap of the in-sample gate-minus-DIRECT-LABEL difference was $+0.62$ percentage points with a 95% interval of $[-0.62, +1.85]$, which includes zero. We therefore prefer the parameter-free DIRECT-LABEL configuration for subsequent analysis; it is also approximately five times faster because it uses one forward pass rather than one per option.

Table 2. Exploratory local interventions; rows use separate paired comparators and are not cross-row comparable. Small: 75 open cases. Large: 600 open/1,200 MCQ, with GT noise $\approx \pm 0.11$/MCQ sampling scale $\approx \pm 1.4$ points. The 32B check uses 8 selected errors.

Intervention	Set	Effect	Supported conclusion
Metadata prompt	small	GT +0.03	no detectable improvement
Self-RAG	small	GT −0.63	harmful in this configuration
Gold-filtered self-distill.	large	GT +0.03; MCQ −2.8 pt	rejected: MCQ decreased
DoRA rank 32	large	GT +0.07; MCQ +0.5 pt	within local uncertainty
Zero-shot 32B	selected n=8	0/8 corrected	targeted diagnostic only

7.2 Permutation Debiasing Is Exactly Null

A real letter-position bias exists: mean next-token log-probability by option position runs $\{-5.89, -6.06, -5.80, -6.68, -7.24\}$, showing an overall but non-monotonic tendency to penalise later positions. Cyclic option-order rotation with score averaging nevertheless scores 0.9538 over the full 325 cases — identical, case for case, to the single pass (310/325 both times): the bias is real but does not flip the argmax. An early run on the first 100 cases showed +2 points, decaying monotonically ($+2.0 \to +0.66 \to +0.57 \to +0.50 \to 0.00$) as the sample grew; an early stopping decision would therefore have selected a null intervention.

7.3 Open-Ended Interventions

We evaluated five exploratory interventions in local experiments with different sample sizes and, for the larger-hold-out studies, separately trained v2-equivalent comparators. Results are paired only within each row of Table 2 and must not be compared as one common experiment. For scale, the submitted system scored 2.09/4 GT and 2.85/4 VA locally on the 75-case hold-out under the same proxies.

The 32B experiment is deliberately narrow. Qwen2.5-VL-32B in 4-bit required an estimated 68 h for one training epoch on our A40, so we first tested it zero-shot on eight errors selected from the 7B system. It corrected none. Together with a median answer–reference lexical overlap of 0.14 and diagnostic review of GT-zero cases, this suggests that some errors involve incorrect diagnoses, but it does not isolate model knowledge as the sole cause or rule out perception, prompting, and judge effects.

We also constructed directional simulations of the announced low-dose X-ray and 7T MRI shifts. The first 7T implementation incorrectly reduced signal-to-noise ratio; we exclude its quantitative results. A corrected simulator models higher-SNR acquisition together with B_1 bias, susceptibility, and chemical-shift effects. These transformations are not calibrated to the hidden private data, so we make no robustness claim from them and release no derived challenge images.

8 Discussion

Inference configuration should be reported. On our hold-out, switching from the submitted isolated option-text configuration to joint option-symbol scoring changed MCQ accuracy by 31.08 points with weights fixed. Because presentation, target, and scoring change together, this is an aggregate contrast, not the causal effect of readout alone. Though not comparable to cross-model results under different protocols, it shows that inference configuration should be documented with the model and prompt [4,1]. Our submitted scorer evaluated option texts in isolation even though SFT used a joint option prompt and label target. The supported lesson is configuration-specific: evaluation should match the supervised output format, and alternative readouts should be declared.

Report construction baselines and selection uncertainty. A model-free baseline scoring 0.899 changes interpretation of a 0.92 result, as hypothesis-only baselines did in NLI [13]. Gate and debiasing gains disappeared under cross-validated or complete-sample evaluation; tuned readouts need separate selection folds and paired uncertainty.

9 Limitations and Conclusion

The main comparison is not a component ablation. It changes presentation, target, and scoring together; factorial experiments are needed to isolate them. *The hold-out shares the training source and construction process.* Its identifiers were excluded from adapter fine-tuning, but we did not audit source-level near-duplicates or the base model's pretraining exposure. The shortest-option baseline demonstrates that construction artifacts transfer to this split, and no image-removal or image-permutation control establishes that the DIRECT-LABEL gain is visually grounded. *Final hidden results were unavailable at writing*; all readout results are local and in-distribution. *Open-ended judges are local proxies* for the published organiser model families, not replicas of undisclosed prompts or infrastructure. At n=75, the estimated GT noise scale of approximately ± 0.3 means that small deltas support "not detectably better," not equivalence. The base-model DIRECT-LABEL control is an approximate aggregate without per-case outputs; the 32B check is an eight-case diagnostic. Corrected simulations are uncalibrated to the hidden OOD data and support no robustness claim. Finally, generated reasoning traces are supervised by dataset visual descriptions but are not tested for causal faithfulness, hallucination, or clinical utility.

The supported finding is limited: on one in-distribution hold-out, switching from isolated option-text to joint option-symbol scoring substantially changed measured MCQ accuracy. Transfer to hidden domain-shift data and medical visual reasoning remain unverified.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alzahrani, N., Alyahya, H.A., Alnumay, Y., et al.: When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (2024)
2. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
4. Gao, L., Tow, J., Abbasi, B., Biderman, S., et al.: A framework for few-shot language model evaluation. Zenodo (2023). https://doi.org/10.5281/zenodo.10256836
5. Gong, H., Ji, X., Liu, Y., Wu, W., Yan, X., Liu, J., Wu, K., Pan, J., Jian, B., Zhang, J., Hu, X., Li, H.B.: Med-CMR: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. arXiv preprint arXiv:2512.00818 (2025)
6. Grattafiori, A., Dubey, A., Jauhri, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. In: Proceedings of NAACL-HLT, Volume 2 (Short Papers). pp. 107–112 (2018)
8. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: International Conference on Learning Representations (ICLR) (2021)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
10. Li, H.B., Hu, X., Khong, P.L.: MedReason: Benchmarking medical multimodal large language models. Zenodo (2026). https://doi.org/10.5281/zenodo.19847619
11. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: DoRA: Weight-decomposed low-rank adaptation. In: International Conference on Machine Learning (ICML) (2024)
12. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
13. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B.: Hypothesis only baselines in natural language inference. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM). pp. 180–191 (2018)
14. Robinson, J., Wingate, D.: Leveraging large language models for multiple choice question answering. In: International Conference on Learning Representations (ICLR) (2023)