Candidate-Expanding Routing with Permutation-Stabilized Experts for Mixed-Format Medical VQA

Hai-Dang Nguyen[1,2] and Huy-Hieu Pham[1,2,3⋆]

[1] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[2] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[3] Center for Innovations in Health Sciences, VinUniversity, Hanoi, Vietnam
dang.nh3@vinuni.edu.vn, hieu.ph@vinuni.edu.vn

Abstract. Mixed-format medical visual question answering (VQA) requires stable option selection and machine-readable free-text output. The two formats fail differently: multiple-choice predictions can change with option symbols or positions, while clinically plausible open answers can fail automated evaluation when serialization is malformed. We address both challenges with an answer-text memory, a permutation-stabilized vision–language expert, and a sparse candidate-expanding router. The cyclic schedule follows prior work; our contribution is to make expert top-2 a routable candidate alongside memory and expert top-1. On a 1,403-case retrospective internal analysis, this expansion improves a matched binary router from 88.95% to **91.73%** (+2.78 percentage points; 95% CI 1.57–3.99), with 56 rescued errors and 17 regressions. Oracle coverage rises from 90.31% to 96.15%, and the final submitted configuration reaches **92.23%** on the same retrospective split. For open questions, strict generation and deterministic guards produce 475/475 schema-valid participant-facing outputs without repair, retry, or hard-gate failure. Visual ablations reveal substantial textual dependence. Candidate expansion supplies the principal controlled routing gain; open-path evidence establishes output-contract validity rather than clinical correctness in medical use or deployment.

Keywords: Medical VQA · vision–language models · option-order sensitivity · expert routing · output validation

1 Introduction

Medical visual question answering (VQA) connects heterogeneous clinical images with questions ranging from recognition to multi-step reasoning [11, 12]. Recent public MLLMs include Qwen2.5-VL, MedGemma, and Lingshu [2, 9, 19]. Med-CMR spans 11 organ systems, 12 modalities, and seven visual or clinical-reasoning tasks [5]. Its mixed format creates two interface problems: MCQs require stable selection from case-defined options, while open questions require

⋆ Corresponding author.

useful content in a machine-readable form. An MCQ can fail through unstable option identifiers, whereas a clinically plausible open answer can become unscorable through malformed fields. A single submission must control both semantic selection and the observable output contract.

MCQ selection is sensitive to option order and symbols [17, 21]. Prior cyclic evaluation exposes each semantic option to every identifier before back-mapping and aggregation [25]; permutation sensitivity also affects VLMs [26], motivating recent LVLM bias correction [1]. Calibration estimates label priors [24], and self-consistency aggregates repeated reasoning paths [20]. Rather than collapse repeated scores to one answer, we retain ranked semantic alternatives so that a correct top-2 option remains available to the selector.

Retrieval-augmented and medical memory models usually condition one generator on retrieved evidence [10, 22]; learning-to-defer and selective-prediction methods instead choose which predictor should answer [4, 13, 16]. We keep memory and VLM as independent candidate generators and merge them only at selection. Routing memory with expert top-1 and top-2 expands reachable answers when the first two candidates are wrong. It also separates *candidate coverage*, whether any candidate is correct, from *router regret*, errors made after a correct candidate is available.

For open VQA, hallucination and rationale faithfulness concern clinical content and causal support [6, 8]; valid JSON establishes neither. Parsing, deterministic guards, and one retry address malformed fields, hidden text, and repetition only as an output contract, without claiming clinical verification.

To address these challenges, we separate candidate formation from selection. The MCQ path constructs one memory candidate and two ranked VLM candidates; the open path places one structured generation behind an explicit serialization contract. The main contributions are:

- We formulate MCQ inference as heterogeneous candidate expansion over answer-text memory and expert top-1/top-2. The third candidate raises matched accuracy by 2.78 percentage points and oracle coverage by 5.84 points on retrospective internal analysis.
- We adapt prior cyclic scheduling to case-defined medical-VQA labels through token-aware scoring, semantic back-mapping, and option-space aggregation while retaining ranked expert outputs for selection.
- We separate coverage from routing error through matched ablations, rescue and regression counts, and router regret. Shortcut controls and the open-output audit bound the resulting claims.

The rest of the paper is organized as follows. Section 2 presents the permutation-stabilized expert, answer-text memory, candidate-expanding router, and open-output contract. Section 3 describes the experimental protocol, baselines, controlled ablations, mechanistic analyses, qualitative evidence, and robustness studies. Section 4 discusses findings, limitations, and conclusions.

2 Method

2.1 System Overview

The common input is a medical image I, question q, task type, and, for MCQ, case-defined label–text pairs $O = \{(\ell_i, o_i)\}_{i=1}^{K}$. A *slot* joins a displayed symbol and its position. The MCQ path outputs one $\hat{\ell} \in \{\ell_i\}$; the open path outputs a validated JSON object containing `answer` and `reasoning_trace`. Figure 1 separates these paths and organizes each as input, method, and output. MCQ retrieval and VLM scoring share no prediction state and meet only at the router; the open branch uses its own retrieval and generation path.

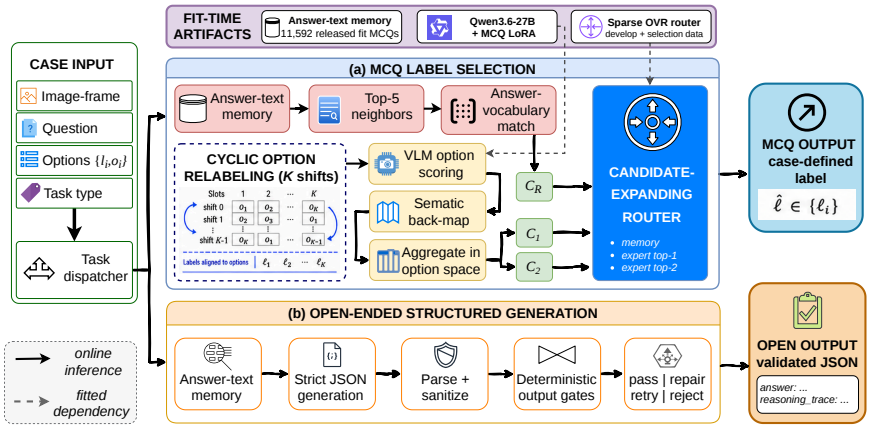

Fig. 1. Mixed-format input-to-output pipeline. A shared image and question are dispatched by task type. (a) For MCQs, answer-text memory independently yields c_R, while cyclic relabeling, VLM scoring, and semantic back-mapping yield c_1 and c_2; our router returns one valid case-defined label. (b) For open questions, retrieved context conditions strict JSON generation, followed by parsing and deterministic gates that return validated `answer` and `reasoning_trace` fields or trigger repair, retry, or rejection. Dashed arrows are fitted dependencies; solid arrows are online inference.

2.2 Heterogeneous MCQ Candidate Generation

Answer-text memory. The memory contains 11,592 released fit MCQs. Fit documents concatenate question, options, organizer visual description, metadata, and correct answer text; queries contain only question and options. Case-folded regex unigrams use smoothed IDF, L2 normalization, and cosine similarity (min_df = 1). We rank 50 neighbors, retain five, match their answer text to current options, and similarity-vote for c_R. Historical labels and evaluated references are never transferred. Because fit descriptions contain a normalized answer substring in 422 cases, we call this an *answer-aware answer-vocabulary memory* and audit the confound in Section 3.6.

Permutation-stabilized expert. Following the cyclic selector evaluation of Zheng et al. [25], we use a balanced K-shift schedule so that every semantic option occupies every case-defined symbol–position slot once. For $s \in \{0, \ldots, K-1\}$,

$$\pi_s(i) = 1 + ((i - s - 1) \bmod K). \tag{1}$$

Only the slot assignment changes; image, question, and option content remain fixed. Option o_i occupies $\ell_{\pi_s(i)}$; scores are mapped back before the submitted log-score average,

$$\bar{z}_i = K^{-1} \sum_{s=0}^{K-1} z_s(\ell_{\pi_s(i)}), \qquad \hat{i} = \arg\max_i \bar{z}_i, \tag{2}$$

This ranks the geometric mean after within-shift normalization; its two highest scores define c_1, c_2. Under the additive slot model $z_s(\ell_{\pi_s(i)}) = u_i + b_{\pi_s(i)} + \epsilon_{is}$, balanced exposure contributes the same mean slot term to each option, but cannot remove content–slot interactions. Benchmark labels are single-token A–E; scoring tests six whitespace/punctuation variants. Parser/router tests also cover numeric, Roman, and arbitrary labels; ties follow case order.

2.3 Candidate-Expanding Router

Candidates are memory c_R and expert ranks c_1, c_2. A one-vs-rest L1 logistic model uses 58 inference features covering agreement, retrieval, confidence, input length, metadata, keywords, and shift consistency. References, correctness, IDs, and oracle switches are excluded. Targets choose the first correct candidate in priority c_R, c_1, c_2 and omit unreachable rows. Five-fold record-level out-of-fold predictions select class multipliers. The final `liblinear` fit ($C = 1$, balanced classes, seed 26) uses development plus a permitted 100-case selection set. Inference applies factors $(1.0, 1.0, 0.7)$, masks duplicate/invalid candidates, and returns the exact case label.

2.4 Open-Ended Output Contract

Open retrieval searches 14,308 fit cases; up to three contexts enter one greedy JSON prompt (256-token cap; penalty 1.05; no repeated 8-grams). Both fields are generated once. Parsing accepts JSON or labeled fields, removes hidden text and Markdown, and caps answer/rationale at 35/80 words. Schema, length, repetition, symbol, meta-text, type, and consistency rules allow serialization-only repair and one retry; persistent failure aborts. Each object thus contains a nonempty answer and brief rationale. These checks verify output-contract properties, not clinical correctness or faithfulness, and are not hallucination detection.

3 Experiments

3.1 Experimental Setup

Data and protocol. Med-CMR reports 20,653 VQA pairs across 11 organ systems, 12 modalities, and seven tasks [5]. We stratify 17,722 labeled challenge cases [15] by question type, task, organ, and modality: fit has 14,308 cases (11,592 MCQ/2,716 open), while development and internal analysis each have 1,707 (1,403/304). Development supports selection. The later-inspected internal split is retrospective development evidence, not an independent evaluation. The unscored participant set has 2,057/475 cases and no final hidden result. MCQs use exact-label accuracy, 10,000 pHash-cluster bootstrap replicates [3], and Holm-corrected exact McNemar tests [14]. Patient/study IDs are unavailable. Med-CMR is CC BY-NC 4.0; no new patient data were collected, and use is research-only.

Models and baselines. Qwen3.6-27B revision `6a9e13b` [18] runs bfloat16 with a frozen vision stack. LoRA ($r = 8$, $\alpha = 16$, dropout 0.05) uses AdamW (10^{-5}), batch 1, accumulation 16/8, and seed 42 on one H100 80GB [7]. A 300-step fit uses 11,592 MCQs plus 2,000 PMC-VQA cases with two relabelings [23]; 40-step continuation on a permitted 100-case selection set chooses step 30.

3.2 Main Baselines

Table 1 reports the reviewed same-split baselines. Candidate expansion raises binary routing from 88.95% to 91.73% (95% CI [1.57, 3.99]). The final retrospective result is 92.23%; the development-tuned margin switch reaches 80.40%.

Table 1. Same-split baselines on 1,403 retrospective internal MCQs. The margin switch is selected on development and frozen; *(Ours)* identifies the proposed and final systems.

Method	Acc. (%)	Method	Acc. (%)
Text / retrieval baselines			
Global answer-frequency prior	23.16	Question-only retrieval	35.71
Answer-free lexical memory	60.73	Answer-aware lexical memory	61.15
VLM expert baselines			
Base Qwen3.6 + context, direct	58.95	Label-shuffle LoRA, direct	72.70
Cyclic probability mean	80.04	Continued adapted expert	81.25
Routing systems			
Dev-tuned margin switch	80.40	Memory + expert top-1	88.95
Top-2 router (Ours)	**91.73**	**Final system (Ours)**	**92.23**

3.3 Controlled Ablation

Table 2 separates expert scoring from candidate expansion. Top-2 rescues 56 binary-router errors and causes 17 regressions: 39 net correct cases (+2.78 percentage points; 95% CI $[1.57, 3.99]$; Holm-adjusted $p = 2.1 \times 10^{-5}$). Permutation diagnostics then change accuracy by -0.14 points (95% CI $[-0.50, 0.21]$).

Table 2. Matched internal-MCQ changes. Each row states its comparison; CIs use pHash-cluster bootstrap sampling. *(Ours)* marks candidate expansion. The base-to-adapted contrast confounds domain supervision with label-shuffle augmentation.

Controlled change	From	To	Δ pp	95% CI
A. Expert stabilization				
Add cyclic stabilization to base	58.95	63.36	4.42	[+2.36, +6.48]
Add label-shuffle adaptation	58.95	72.70	13.76	[+11.91, +15.67]
Add cyclic plurality to adapted	72.70	79.83	7.13	[+5.41, +8.92]
Use probability-mean aggregation	79.83	80.04	0.21	[-0.78, +1.21]
Switch to submitted log mean	80.04	79.90	−0.14	[-0.79, +0.50]
Continue adapted expert	79.90	81.25	1.35	[+0.71, +2.06]
B. Controlled router ablation				
Add expert top-2 (Ours)	88.95	**91.73**	**2.78**	**[+1.57, +3.99]**
Add permutation diagnostics	91.73	91.59	−0.14	[-0.50, +0.21]

3.4 Mechanistic Analysis

Figure 2 distinguishes agreement-based diagnosis from candidate reachability. Adding c_2 raises oracle coverage from 90.31% to 96.15% and matched routing from 88.95% to 91.73%.

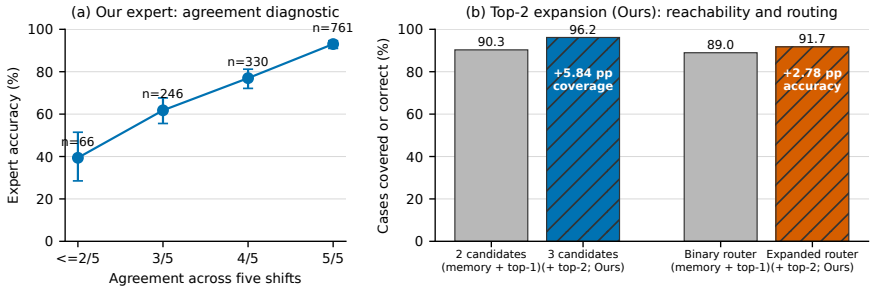

Fig. 2. Mechanistic evidence. (a) Expert accuracy versus five-shift agreement, with Wilson intervals and bin counts; agreement is diagnostic only. (b) Top-2 expands oracle coverage from 90.31% to 96.15% and matched routing from 88.95% to 91.73%; hatching marks expanded systems.

Cyclic scoring improves both checkpoints. Probability and log averaging are equivalent on identical logits (80.04% vs. 79.90%, $p = 0.83$), while the adaptation contrast remains confounded by a missing same-data standard-LoRA checkpoint. Candidate expansion makes 82 answers newly reachable. The router selects $c_R/c_1/c_2$ for 861/427/115 cases, with conditional accuracies 96.86/93.68/52.17%; c_2 cases are harder because earlier candidates conflict. The remaining 55 reachable errors yield 3.92 points of router regret. Agreement predicts correctness (AUROC 0.750), but margin is stronger (0.858), matching the neutral feature ablation in routing.

3.5 Qualitative Evidence Across Formats

Figure 3 shows two CT cases selected by fixed, reproducible criteria. In the MCQ, memory and top-1 fail; top-2, our router, and the reference agree, while visual removal produces another error. The open example maximizes a fixed answer/rationale overlap score among fit-disjoint radiology outputs and preserves both reference findings. Together they expose the two mechanisms without replacing aggregate or clinical validation.

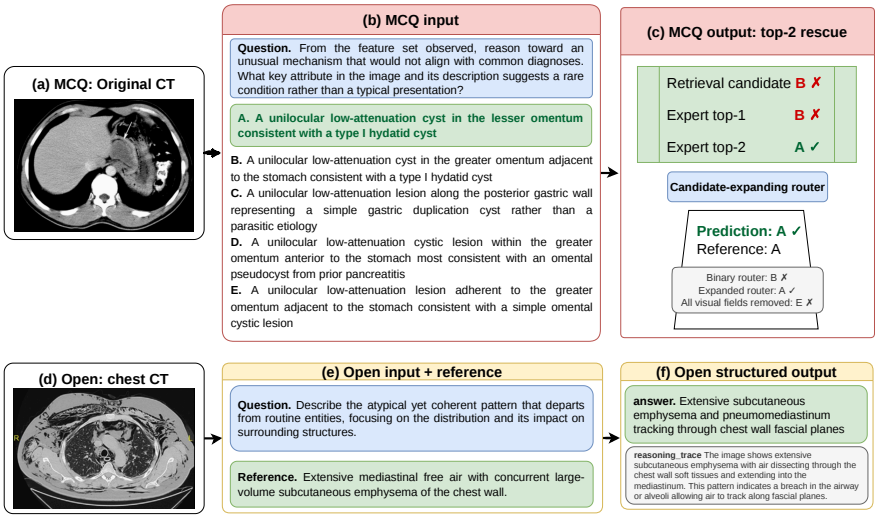

Fig. 3. Real Med-CMR input-to-output examples [5, 15]. (a–c) For abdominal CT, memory/top-1 predict B, top-2 exposes reference A, and our router selects A; visual removal predicts E. (d–f) For chest CT, the open output preserves both reference findings in valid JSON. Images are only resized; both are SHA-, pHash-, and template-disjoint from fit. These cases illustrate mechanism, not aggregate validity.

3.6 Robustness and Sensitivity

Options-only and question-plus-options retrieval score 61.30% and 61.65%, exposing answer-vocabulary regularity. Answer-free memory scores 60.73% versus 61.15% answer-aware; their disjoint scores are 60.35/61.04%. We find 78 exact-image overlaps but no patient/study IDs. A pHash-grouped refit reaches 92.37%, close to 92.23%; this limits grouping sensitivity but is not independent validation. Each control in Table 3 recomputes expert features and routing with the trained system fixed.

Table 3. Visual sensitivity on retrospective internal MCQs. *(Ours)* is matched input; other rows are independent controls, not a cumulative sequence.

Input condition	Expert	Router	Δ (pp)	95% CI	Flip
Full input (Ours)	**81.25**	**92.23**	0.00	reference	0.00
Raw image omitted	76.27	88.95	−3.28	[-4.49, -2.14]	6.91
All visual fields omitted	75.98	89.17	−3.06	[-4.34, -1.85]	7.34
Modality shuffle (3 seeds)	79.66	90.16	−2.07	[-3.09, -1.07]	6.27

Open-output contract. All 475 participant-facing outputs are schema-valid, with no repair, retry, or hard-gate failure; seven review warnings remain. Organizer pre-evaluation on 425 open cases reported GT 1.859/4 and VA 2.774/4. No gate intervened and the evaluator is unavailable; this establishes output-contract validity only, not clinical benefit.

4 Discussion and Conclusion

Expert top-2 provides the principal controlled routing gain: 39 net correct cases (+2.78 percentage points) and oracle coverage of 96.15% rather than 90.31%. This headroom enables the final 92.23% retrospective result; Figure 3 illustrates the mechanism, not aggregate validity.

Cyclic relabeling is prior work; probability and log aggregation are equivalent here, and agreement is diagnostic but neutral for routing. Evidence remains retrospective, with answer regularity, image/template overlap, record-level OOF, and no patient/study IDs. Raw-image omission retains 88.95%, so image dependence cannot establish faithful grounding. Hidden results and clinician review are unavailable.

Candidate expansion, not extra uncertainty features, supplies the main gain; open guards enforce serialization only. Future work should freeze the system for blind patient-disjoint evaluation, use set-valued targets, and compare guarded and raw answers with clinician-supported metrics.

Acknowledgments. This research was funded by the National Foundation for Science and Technology Development (NAFOSTED) through Project No. IZVSZ2_229539 (2025–2027).
Disclosure of Interests. The authors have no competing interests to declare.

References

1. Atabuzzaman, M., Asgarov, A., Thomas, C.: Benchmarking and mitigating MCQA selection bias of large vision-language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 33548–33562. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.emnlp-main.1703
2. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025). https://doi.org/10.48550/arXiv.2502.13923
3. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability, Chapman and Hall/CRC (1993)
4. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
5. Gong, H., Ji, X., Liu, Y., et al.: Med-CMR: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41224–41234 (2026)
6. Gu, Z., Yin, C., Liu, F., Zhang, P.: MedVH: Towards systematic evaluation of hallucination for large vision language models in the medical context. arXiv preprint arXiv:2407.02730 (2024). https://doi.org/10.48550/arXiv.2407.02730
7. 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 (2022)
8. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.386
9. LASA Team, Xu, W., Chan, H.P., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025). https://doi.org/10.48550/arXiv.2506.07044
10. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track (2023). https://doi.org/10.52202/075280-1240
12. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023). https://doi.org/10.1016/j.artmed.2023.102611
13. Mao, A., Mohri, M., Zhong, Y.: Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In: Proceedings of the 42nd

International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 43035–43066. PMLR (2025)
14. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
15. MedReason Challenge Organizing Team: MedReason participation guide (2026), https://medreason26.github.io/challenge.html
16. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 7076–7087. PMLR (2020)
17. Pezeshkpour, P., Hruschka, E.: Large language models sensitivity to the order of options in Multiple-Choice questions. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 2006–2017. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-naacl.130
18. Qwen Team: Qwen3.6-27B (Apr 2026), https://huggingface.co/Qwen/Qwen3.6-27B
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025). https://doi.org/10.48550/arXiv.2507.05201
20. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations (2023)
21. Yang, Z., Jian, P., Li, C.: Option symbol matters: Investigating and mitigating Multiple-Choice option symbol bias of large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1902–1917. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.naacl-long.95
22. Yuan, Z., Jin, Q., Tan, C., Zhao, Z., Yuan, H., Huang, F., Huang, S.: RAMM: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 547–556. ACM (2023). https://doi.org/10.1145/3581783.3611830
23. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023). https://doi.org/10.48550/arXiv.2305.10415
24. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 12697–12706. PMLR (2021)
25. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. In: International Conference on Learning Representations (2024)
26. Zong, Y., Yu, T., Chavhan, R., Zhao, B., Hospedales, T.: Fool your (Vision and) language model with embarrassingly simple permutations. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 62892–62913. PMLR (2024)