CG-OAS: ORDER-CONSISTENCY-GATED ANSWER SELECTION FOR MEDICAL RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical RAG suffers from evidence-order sensitivity — the same set of evidence can produce inconsistent answers simply due to different ordering of the input. While prior work typically treats this as a robustness weakness, we find that this order sensitivity itself can be repurposed as a usable reliability signal. Through evidence-order perturbation (generating variants by permuting the same evidence set), examples fall into two regimes: strict-consistency (all permutations converge to the same answer) and non-strict (disagreement across permutations). The non-strict regime directly pinpoints where the model's evidence comprehension is fragile, providing a principled basis for answer selection. Motivated by this finding, we propose CG-OAS, a consistency-gated answer selection protocol. CG-OAS preserves the predicted answer when all evaluated evidence permutations agree; otherwise, it applies a lightweight, dataset-specific selector trained on non-unanimous validation examples. The selector uses only features derived from hard predictions, requiring neither evidence text nor model probabilities. The protocol returns an answer for every question and operates on outputs from validation-selected LoRA checkpoints without modifying retrieval or requiring additional generator training. On the MedMCQA and MedQA evaluation sets, learned selection improves accuracy over majority voting on non-unanimous questions by 2.90 and 4.98 percentage points, respectively, while leaving unanimous predictions unchanged. Overall, CG-OAS achieves a macro-averaged accuracy of 61.14%, exceeding original-order and majority-vote LoRA RAG by 1.36 and 1.04 percentage points, respectively, with paired bootstrap confidence intervals supporting both macro-averaged gains. These results show that evidence-order perturbations can provide actionable signals for targeted answer selection in medical RAG without updating the generator.

1 INTRODUCTION

Changing the order of retrieved evidence can change a model's answer even when the retrieved passage set is fixed. In medical retrieval-augmented generation (RAG), this sensitivity complicates the interpretation of a single prediction: an answer may depend on how evidence is presented as well as on its content. Prior work studies context position and retrieval permutations as sources of instability (Liu et al., 2024; Zhang et al., 2026; Han et al., 2026). We examine the answer-selection opportunity created by these variations.

Our central question is: *Do prediction patterns induced by different evidence orders contain answer-selection information beyond conventional majority voting?* Multiple evidence orders produce a sequence of discrete answers, which we call a *prediction pool*. Majority voting primarily uses answer frequency. The pool also records which answer occurs under the original order, where candidates appear in the sequence of calls, and how their frequencies relate to the distribution of competing answers. These observations support a supervised selection problem over answers the generator has already produced.

We introduce *Consistency-Gated Order-Aware Answer Selection (CG-OAS)*. The method accepts unanimous predictions directly and applies a candidate-constrained selector to non-unanimous

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pools. The selector learns from labeled validation questions; the generator is frozen throughout aggregation. Its inputs comprise hard-prediction statistics, order-occurrence features, validation priors, and interactions. It neither reads the retrieved passage text nor updates the generator. The unanimity gate makes routing explicit, while candidate-constrained supervision supplies the capacity to depart from the majority answer. Figure 1 summarizes the inference pipeline.

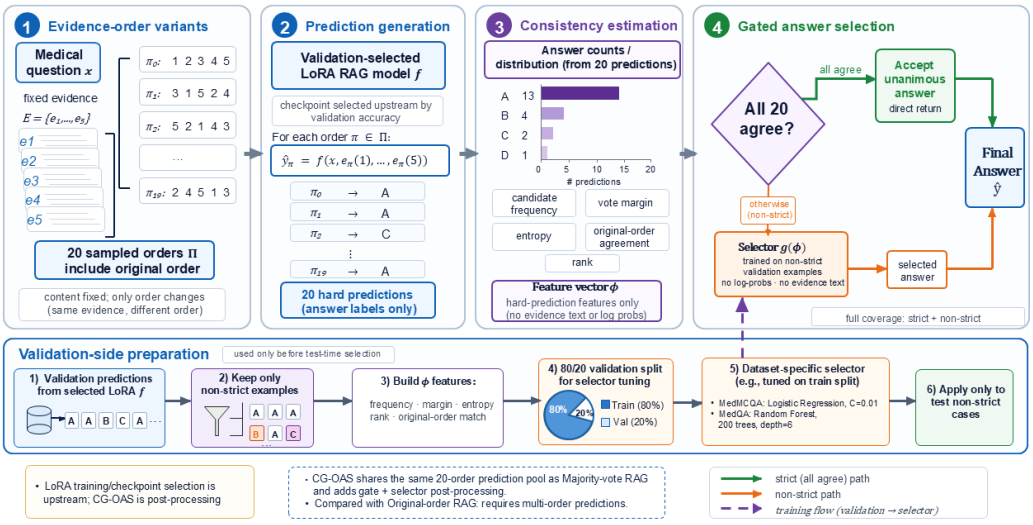


Figure 1: Overview of CG-OAS. The same retrieved evidence set is evaluated under multiple evidence-order variants, producing an order-induced prediction distribution. Strict-consistency examples are accepted directly, while non-strict examples are routed to a validation-trained selector operating on hard-prediction features. CG-OAS is applied after prediction generation and does not modify the retriever, generator, or LoRA training.

Under the canonical 20-order protocol on MedMCQA and MedQA, CG-OAS achieves macro accuracy 0.6108, compared with 0.5979 for Original-order and 0.6010 for Majority-vote. The gains are $+1.29$ and $+0.98$ percentage points (pp), with paired 95% bootstrap intervals $[0.44, 2.17]$ and $[0.21, 1.75]$. These intervals exclude zero for the selected prediction vectors; they are conditional on the selector-family selection described in Section 4. A No-gate selector achieves 0.6076, and its difference from CG-OAS is not statistically resolved. Deduplicating repeated permutations preserves positive point differences over Original-order and Majority-vote, but reverses the small difference from No-gate. Thus the evidence favors supervised aggregation within the evaluated setting, rather than an independent accuracy benefit from unanimity routing.

Our contributions are threefold:

- We formulate and study whether evidence-order-induced discrete prediction patterns carry answer-selection information beyond conventional majority voting.
- We develop CG-OAS, a candidate-constrained supervised aggregation method that operates on a fixed generator's prediction pools and transparently routes unanimous cases.
- We characterize its performance and boundaries on two medical QA datasets using paired bootstrap, coverage analysis, repeated-prompt agreement, passage-count sensitivity, and unique-permutation deduplication.

2 Related Work

Medical RAG and evaluation. RAG combines retrieved information with a generative model (Lewis et al., 2020). MIRAGE and MedRAG study how retrieval corpora, retrieval methods, and

generators affect medical question answering (Xiong et al., 2024). MedMCQA and MedQA provide medical examination questions with discrete answer labels (Pal et al., 2022; Jin et al., 2021). Our evaluation fixes retrieval and generation within each dataset to study aggregation. Benchmark answer accuracy measures success on these questions, rather than clinical reliability or the quality of patient care.

Context position and evidence order. Liu et al. (2024) show that using information in long contexts depends on its position. Lu et al. (2022) study ordering effects in few-shot demonstrations, a different source of prompt sensitivity. More directly, Stable-RAG estimates retrieval-permutation sensitivity, clusters hidden states, and uses representative reasoning outputs for alignment (Zhang et al., 2026). Han et al. (2026) examine uncertainty and order effects under conflicting biomedical evidence. Mixture-of-Intervention (MOI) aggregates observations across retrieved-context permutations to estimate passage utility separately from positional bias and then orders passages for generation (Lee et al., 2025). These studies motivate isolating evidence order while holding the retrieved set fixed. CG-OAS selects an answer after generation from discrete predictions; it neither estimates passage utility nor changes retrieval or the generator. We do not infer performance superiority under unmatched settings.

Multiple predictions, uncertainty, and selection. Self-consistency samples reasoning paths and aggregates their answers (Wang et al., 2022). SelfCheckGPT uses agreement among sampled responses to assess factuality (Manakul et al., 2023), while semantic entropy measures uncertainty over meanings (Farquhar et al., 2024). Confidence-informed self-consistency weights answers by model-derived confidence (Taubenfeld et al., 2025). Ranked voting instead elicits an ordered answer list per response and combines these lists using voting rules (Wang et al., 2025). Both go beyond unweighted majority voting, but require information absent from our single-label records. Evaluating them would require new generation or scoring calls and a matched resource protocol. Hard predictions offer a compact interface without token probabilities or reasoning traces, but discard semantic and confidence information and cannot recover an answer absent from the pool. Unanimity is a routing condition, not calibrated confidence or a correctness guarantee.

Supervised aggregation. Learning from the outputs of other predictors is the classical setting of stacked generalization (Wolpert, 1992). A close modern comparison is Ranking-Improved Self-Consistency (RISC), which learns a LambdaRank model using answer frequency, semantic centrality, and reasoning-trace consistency (Marina et al., 2026). RISC samples chain-of-thought responses and evaluates open-domain QA and mathematics; its features use embeddings of answers and reasoning traces. CG-OAS instead perturbs retrieved-evidence order with greedy answer-label generation and studies medical multiple-choice QA using prediction statistics, occurrence features, and validation priors. Thus our empirical contribution concerns an evidence-order-induced prediction representation and its protocol sensitivity, rather than supervised ranking in general. Different tasks, information interfaces, generation budgets, and selection protocols preclude direct comparison of reported accuracies. Majority voting supplies an unsupervised comparator; No-gate supplies the corresponding supervised comparison using the same feature specification. Neither supervision nor stacking is itself a new learning principle.

3 METHOD

3.1 EVIDENCE ORDERS AND PREDICTION POOLS

For question x_i with gold answer y_i, let $E_i = (e_{i1}, \ldots, e_{ik_i})$ denote its retrieved passages, with $1 \le k_i \le 5$. An evidence order π_{it} is a permutation of passage indices for call t. A frozen generator f_θ and a fixed answer parser produce

$$\hat{y}_{it} = \mathrm{parse}\big(f_\theta(x_i; e_{i,\pi_{it}(1)}, \ldots, e_{i,\pi_{it}(k_i)})\big), \qquad P_i = (\hat{y}_{i0}, \ldots, \hat{y}_{i,B-1}). \tag{1}$$

Here $B = 20$ and P_i is a sequence retaining repeated orders and repeated answers. Let A_i be the set of valid answers observed in P_i. For $a \in A_i$, define its count $c_i(a) = \sum_{t=0}^{B-1} \mathbf{1}[\hat{y}_{it} = a]$ and frequency $q_i(a) = c_i(a)/B$. The unanimity indicator is $u_i = \mathbf{1}[\max_a c_i(a) = B]$. It refers to agreement across the evaluated calls, not across all possible permutations.

The protocol draws 20 evidence orders with replacement, including the original and reversed retrieval orders. Calls 0 and 1 use those two fixed orders; calls 2–19 shuffle passage indices with seeds $42 + 100i + t$, where i is the zero-based position in the matched evaluation list. Repeated permutations are retained. This distinction matters because multiplicity changes both vote frequencies and selector features.

3.2 Candidate scores and supervision

Each candidate is represented by $\phi_i(a) \in \mathbb{R}^{25}$. The representation combines frequency and margin statistics, answer-distribution entropy, frequency rank, first/last/mean occurrence indices, original- and last-order indicators, parsing metadata, validation answer and prediction priors, and five interaction features. Occurrence indices describe positions in the *prediction sequence*, not the positions of individual passages. Appendix A defines the features exactly, including duplicated original-answer indicators and features that become constant on a fitting split.

The interactions allow pool-level statistics to influence candidate ranking. For example, frequency multiplied by normalized entropy distinguishes a given frequency in concentrated and dispersed answer pools. This is a representational capability, rather than a feature-level causal explanation of the observed gains. No evidence text, raw-response length, hidden states, or token probabilities enter the 25-dimensional representation.

For each validation question and observed candidate, the binary training target is $z_{ia} = \mathbf{1}[a = y_i]$. Candidate rows with the gold answer absent have only negative targets. CG-OAS fits its selector on non-unanimous validation questions. An 80/20 split is made by question with seed 42, keeping all candidates of a question together. Features are standardized using the fitting rows, and constant columns are removed for CG-OAS. The classifier returns a positive-class probability $s_\eta(i, a)$ used as a ranking score. The fitted model is retained without refitting on the internal development partition.

The selected classifiers are logistic regression for MedMCQA and a random forest for MedQA. Logistic regression uses $C = 0.01$, balanced class weights, and at most 1,000 iterations. The forest uses 200 trees, maximum depth 6, minimum leaf size 2, and balanced class weights. Both use random state 42. Priors are empirical frequencies computed from the entire 500-question validation pool before the internal split, so that split is not independent of prior estimation. Prior correction enters the selected models through these features; an additive prior-adjusted voting rule is a separate comparator in the implementation, not an additional post-processing step of CG-OAS. Selection details and their statistical scope appear in Section 4.

3.3 Routing and the No-gate comparison

CG-OAS chooses only among observed candidates:

$$\hat{y}_i^{\mathrm{CG}} = \begin{cases} a, & u_i = 1 \text{ and } A_i = \{a\}, \\ \arg\max_{a \in A_i} s_\eta(i, a), & u_i = 0. \end{cases} \tag{2}$$

Score ties prefer greater candidate frequency, then the original-order answer. Empty pools retain an empty prediction and are counted as incorrect; none occurs in the evaluation. A missing selector output routes to majority voting. All evaluation questions have valid predictions, so parser or empty-pool fallbacks do not alter the reported denominators.

Property: unanimity routing. For a selector constrained to choose among answers observed in the prediction pool, a unanimous pool contains only one admissible candidate. Explicit unanimity routing and direct candidate-constrained selection are therefore prediction-equivalent on unanimous examples. The gate provides a transparent routing structure and avoids selector evaluation on these cases; it does not save generation calls.

The No-gate selector uses the same 25-feature specification and the same selected classifier families, but fits on an 80/20 question split of *all* validation questions and scores every test pool. Thus its comparison with CG-OAS changes the fitting distribution as well as explicit routing. It is not an isolated test of an inference-time gate. Holding a candidate-constrained selector fixed and adding the gate would leave every prediction unchanged.

Algorithm 1 summarizes the complete CG-OAS answer-selection procedure.

Algorithm 1 CG-OAS answer selection

Require: Question x; evidence set $E = \{e_1, \ldots, e_5\}$; order set Π containing 20 orders; validation-selected LoRA generator f; selector g trained on validation non-strict predictions
Ensure: Final answer $\hat{y}$
for all $\pi \in \Pi$ **do**
  Generate $\hat{y}_\pi \leftarrow f\big(x, e_{\pi(1)}, \ldots, e_{\pi(5)}\big)$
end for
Compute the answer counts $c(a) = \sum_{\pi \in \Pi} \mathbb{I}[\hat{y}_\pi = a]$
if $|\{\hat{y}_\pi : \pi \in \Pi\}| = 1$ **then**
  return the unanimous answer
else
  Construct hard-prediction features $\phi = \Phi\big(\{\hat{y}_\pi : \pi \in \Pi\}, c\big)$
  return the selector output $g(\phi)$
end if

4 EXPERIMENTS

Datasets and coverage. We evaluate medical multiple-choice answers on MedMCQA and MedQA, with 500 validation questions per dataset. The evaluation contains 4,180 of the 4,183 MedMCQA registry questions and all 1,273 MedQA questions. Three MedMCQA questions were excluded before generation because the fixed BM25/Textbooks retriever returned no passages. No gold labels or model correctness were used to select this subset. All four methods use exactly the same question IDs and denominators: 4,180 and 1,273. Registry coverage is 99.9283% and 100%, respectively, and answer coverage within the evaluation subsets is 100%.

Retrieval and generation. The generator is Qwen2.5-7B-Instruct with a dataset-specific low-rank adaptation (LoRA) checkpoint. Epoch 3 is selected for both datasets by validation majority-vote accuracy among epochs 0, 1, 2, 3, and 5. The generator is held fixed across the four methods. Inference loads the base model in 4-bit NF4 with double quantization and bfloat16 computation, then loads the LoRA adapter. This inference configuration is distinct from the adapter-training configuration. BM25 retrieves up to five passages from Textbooks. Of the 4,180 MedMCQA questions, 4,178 contain five passages, one contains three, and one contains one; all MedQA questions contain five passages.

Each question receives 20 generation calls using the order protocol in Section 3.1. Decoding sets `do_sample=False`, `temperature=0.0`, and `max_new_tokens=16`. No `top_p`, `top_k`, or beam-search arguments are explicitly passed, and the evaluator sets no explicit PyTorch or Transformers generation seed. Greedy decoding therefore describes the requested decoding rule, not a guarantee of identical outputs for repeated prompts. The prompt requests an answer label without an explanation; Appendix B gives the prompt and parser conventions.

Baselines and resource conditions. Original-order uses the prediction from call 0. Majority-vote selects the most frequent answer from the 20-call pool, breaking ties by original-order preference and then descending candidate string. These rules require no aggregation labels. No-gate and CG-OAS use validation supervision and the same 20-call pools. Aggregation never updates the generator. The multi-order methods use 20 times the generation-call budget of Original-order, totaling 83,600 MedMCQA and 25,460 MedQA test calls. These are generation calls, not forward-pass counts.

Selector selection and statistical scope. Candidate classifiers are fitted using validation labels. Logistic-regression hyperparameters are selected by question-level accuracy on the internal development split. The implementation evaluates logistic regression, random forest, and gradient boosting; the reported CG-OAS family for each dataset is the one with the highest *test* accuracy among these three. This selects logistic regression for MedMCQA and random forest for MedQA. Consequently, the results describe selected configurations, and the reported bootstrap intervals do not adjust for selector-family selection. No-gate uses the corresponding families. This distinction separates validation-based parameter fitting and checkpoint selection from test-based family selection, and limits a confirmatory interpretation of the comparisons.

Metrics and uncertainty. Accuracy is the fraction of correct question-level answers. Macro accuracy is the unweighted mean of the two dataset accuracies, rather than accuracy pooled across all questions. Paired bootstrap resamples question IDs jointly across methods within each dataset and averages the two resulting accuracy differences. We use 10,000 replicates, seed 20260703, and percentile 95% confidence intervals. Predictions and fitted selectors remain fixed during resampling. These intervals quantify conditional question-sampling uncertainty, not uncertainty from training, permutation seeds, or model-family selection.

5 RESULTS

5.1 PREDICTION-POOL AGGREGATION IMPROVES ON VOTING

CG-OAS improves answer selection over both unsupervised baselines under the canonical protocol (Table 1). Its macro accuracy is 0.6108, compared with 0.5979 for Original-order and 0.6010 for Majority-vote. The corresponding gains are $+1.29$ and $+0.98$ pp. Both paired intervals exclude zero (Table 2), giving statistically resolved differences for the selected, fixed outputs. The selection scope in Section 4 applies to these intervals.

Table 1: Canonical 20-order results. Accuracy is reported on identical subsets of 4,180 MedMCQA and 1,273 MedQA questions. MV denotes Majority-vote; differences are macro percentage points. Supervision refers only to aggregation, since all rows use the same adapted generator. Original-order uses one call; all other methods use the same 20-call pool. Paired intervals appear in Table 2.

METHOD	SUPERVISION	MEDMCQA	MEDQA	MACRO	Δ VS. MV
Original-order	None	0.5861	0.6096	0.5979	-0.31
Majority-vote	None	0.5861	0.6159	0.6010	$+0.00$
No-gate selector	Validation	0.5907	0.6245	0.6076	$+0.66$
CG-OAS	Validation	0.5947	0.6269	0.6108	$+0.98$

Table 2: Paired macro-accuracy differences under the canonical protocol (10,000 replicates; seed 20260703). Intervals are conditional on the selected prediction vectors and do not adjust for selector-family selection. CG-OAS exceeds both unsupervised baselines; its difference from the No-gate selector is not statistically resolved.

COMPARISON	MACRO DIFFERENCE (PP)	PAIRED 95% CI (PP)
CG-OAS – Original-order	$+1.29$	$[0.44, 2.17]$
CG-OAS – Majority-vote	$+0.98$	$[0.21, 1.75]$
CG-OAS – No-gate selector	$+0.32$	$[-0.37, 0.99]$

The gains are positive on both datasets: CG-OAS exceeds Majority-vote by 0.86 pp on MedMCQA and 1.10 pp on MedQA. Macro significance does not imply significance for each dataset separately: the MedQA interval for this contrast includes zero (Appendix C). No-gate itself exceeds Majority-vote by 0.66 pp in macro accuracy. This point estimate and the close performance of the two learned variants support supervised aggregation as the principal source of improvement.

5.2 COMPARING THE TWO SUPERVISED VARIANTS

CG-OAS has a 0.32 pp higher point estimate than No-gate, but the paired 95% interval is $[-0.37, 0.99]$ pp. The two supervised variants therefore perform comparably at the available statistical precision. Together with their prediction equivalence on unanimous pools, this finding does not identify an independent benefit from inference-time routing.

All improvements over Majority-vote occur on non-unanimous questions. These account for 1,279/4,180 MedMCQA questions and 301/1,273 MedQA questions. Within those groups, accuracy rises from 0.3729 to 0.4011 and from 0.3854 to 0.4319, respectively. On unanimous questions, all candidate-constrained methods retain the same answer. Their accuracies, 0.6801 and 0.6872, also

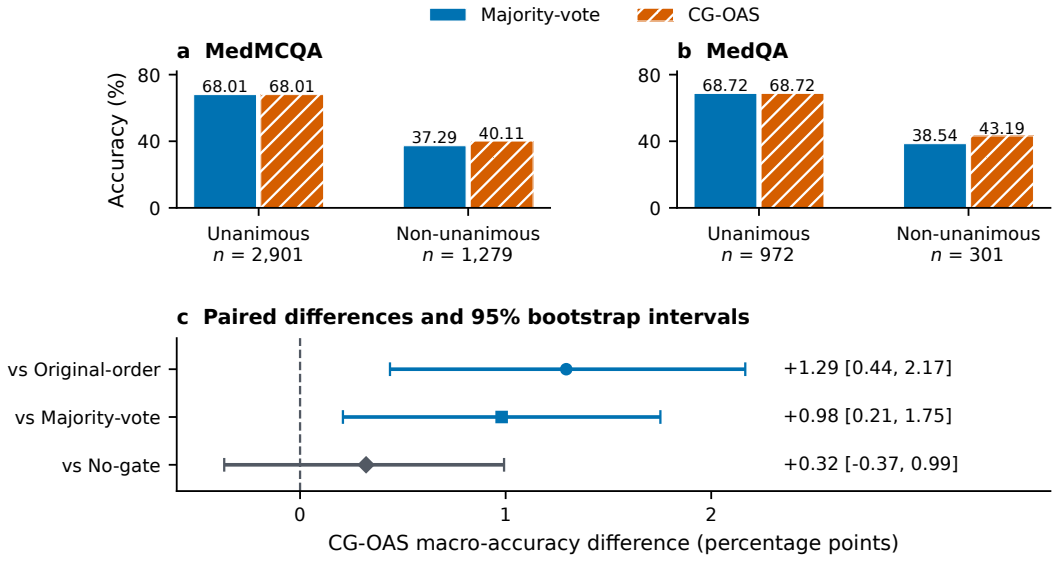


Figure 2: Answer selection under the canonical 20-order protocol. (a,b) Group accuracies and question counts: CG-OAS and Majority-vote agree on unanimous pools, while net gains occur on non-unanimous pools. Bars are descriptive point estimates. (c) Paired macro-accuracy differences with 95% bootstrap intervals (10,000 replicates, seed 20260703) over 4,180 MedMCQA and 1,273 MedQA questions. The dashed line marks zero. The No-gate interval includes zero; the plot does not establish an independent gate benefit. Intervals condition on selected outputs and do not adjust for model-family selection.

show why unanimity is not a guarantee of correctness. Figure 2 relates these group results to the paired macro comparisons. Appendix C gives group denominators and exact correct counts.

6 Robustness and Protocol Boundaries

6.1 Repeated Prompts and Residual Nondeterminism

Repeated permutations provide a direct check on reproducibility within the evaluated pools. Across 8,742 repeated-permutation pairs, 33 parsed answers differ: a mismatch rate of 0.3775% and agreement of 99.62%. The pair-level Clopper–Pearson 95% interval for mismatch is $[0.2600\%, 0.5297\%]$; the Wilson interval is $[0.2689\%, 0.5296\%]$. These are descriptive pair-level intervals; pairs can share a question. High repeated-prompt agreement is consistent with evidence order accounting for most observed variation, while a small generation-level contribution remains.

The mismatches affect 31 questions (26 MedMCQA and five MedQA), all non-unanimous. Excluding these questions as a diagnostic, without refitting, yields macro accuracy 0.6112 for CG-OAS, 0.5986 for Original-order, 0.6021 for Majority-vote, and 0.6092 for No-gate. The point differences remain positive: $+1.26$, $+0.92$, and $+0.20$ pp. This diagnostic exclusion is based on repeated-output instability, not on gold labels, and does not replace the main evaluation subset.

6.2 Unique-Permutation Pools

Repeated orders are expected under sampling with replacement. Among five-passage questions, at least one repetition occurs for 80.8760% of MedMCQA and 79.8115% of MedQA questions. Both datasets average approximately 18.51 distinct permutations per 20 calls. The observed range is 14–20; it excludes the two questions with fewer passages. Repetition preserves the prescribed call budget but increases the weight of some evidence arrangements.

To examine this dependence, we collapse each identical passage-index permutation to one representative, resolve conflicting repeated outputs by majority and then earliest occurrence, and rebuild the

validation and test features using the resulting pool size. Selectors are refitted under this protocol (Appendix D). Table 3 compares the resulting macro accuracies.

Table 3: Order-pool sensitivity on the same 4,180 and 1,273 questions. Entries are macro accuracies. Under deduplication, CG-OAS has positive point differences over Original-order ($+0.87$ pp) and Majority-vote ($+0.40$ pp), but falls below No-gate (-0.15 pp). These deduplicated differences are point estimates, without a macro significance claim.

PROTOCOL	ORIGINAL	MAJORITY	NO-GATE	CG-OAS
20 records, with replacement	0.5979	0.6010	0.6076	0.6108
Unique permutations	0.5979	0.6026	0.6081	0.6065

he sign of the difference from No-gate reverses after deduplication. Thus the small canonical advantage over this supervised comparator is sensitive to order-pool construction. The positive differences from Original-order and Majority-vote remain, but do not establish that every sampling protocol yields the same gains. The canonical and deduplicated results answer related questions about different pool constructions.

6.3 Passage count and registry coverage

Restricting evaluation to questions with exactly five passages leaves 4,178 MedMCQA and 1,273 MedQA questions. CG-OAS reaches macro accuracy 0.6107; its differences from Original-order, Majority-vote, and No-gate are $+1.29$, $+0.98$, and $+0.32$ pp. Counting the three retrieval-empty registry questions as incorrect for every method instead yields macro accuracy 0.6106, with the same differences at two-decimal precision. These checks leave the main point comparisons essentially unchanged. Parser success is 100% across 109,060 generation records: no question has partial or complete parse failure, an empty candidate set, or a sentinel prediction. Full sensitivity tables appear in Appendix D.

7 Limitations

The evidence concerns two medical examination datasets and one main backbone with dataset-specific adapters. The representation combines prediction frequencies, occurrence indices, and validation priors, so improvements over voting do not isolate a particular feature or establish superiority to every equally supervised vote-correction method. Selector families are chosen by test accuracy and the internal development split shares prior estimation with the fitting split. The conditional bootstrap intervals do not account for these selection effects. These results are therefore evidence about the evaluated prediction pools, with a narrower scope than an independently selected, confirmatory model comparison.

The method also depends on how those pools are constructed. It requires 20 generation calls per question, and the unanimity gate does not reduce that cost. Full CG-OAS has no statistically resolved advantage over No-gate under the canonical protocol and a lower point estimate after deduplication. A small amount of generation nondeterminism persists under greedy decoding. Transfer of the fitted selector across models, retrievers, and domains is outside this evaluation. Finally, examination accuracy and agreement do not establish calibrated uncertainty, evidence comprehension, or clinical reliability.

8 Conclusion

Prediction patterns induced by evidence order offer useful answer-selection information beyond conventional majority voting in the evaluated setting. Candidate-constrained supervised aggregation improves the selected canonical macro accuracy over Original-order and Majority-vote, with paired fixed-output intervals excluding zero. The comparison with No-gate places the unanimity gate in its appropriate role: transparent routing rather than an independently established source of accuracy

gains. The deduplication results further show that order-pool construction is part of the method specification and should be reported explicitly.

AI ASSISTANCE STATEMENT

Generative AI tools assisted with literature lookup, manuscript organization and editing, LaTeX preparation, and code for numerical consistency and PDF checks. Numerical claims were checked against saved prediction vectors and statistical outputs, and cited publication records were consulted. The authors retain responsibility for the manuscript, experimental interpretation, and references.

REFERENCES

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Yikun Han, Mengfei Lan, and Halil Kilicoglu. When evidence conflicts: Uncertainty and order effects in retrieval-augmented biomedical question answering. In Dina Demner-Fushman, Sophia Ananiadou, Kirk Roberts, and Junichi Tsujii (eds.), *BioNLP 2026*, pp. 630–643, San Diego, California, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-434-7. doi: 10.18653/v1/2026.bionlp-1.50. URL https://aclanthology.org/2026.bionlp-1.50/.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Youngwon Lee, Seung-won Hwang, Daniel F Campos, Filip Gralinski, Zhewei Yao, and Yuxiong He. Inference scaling for bridging retrieval and augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7339–7354, 2025. doi: 10.18653/v1/2025.findings-naacl.409. URL https://aclanthology.org/2025.findings-naacl.409/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Maria Marina, Daniil Moskovskiy, Sergey Pletenev, Mikhail Salnikov, Alexander Panchenko, and Viktor Moskvoretskii. Boosting self-consistency with ranking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 1017–1032, 2026. doi: 10.18653/v1/2026.acl-srw.89. URL https://aclanthology.org/2026.acl-srw.89/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Weiqin Wang, Yile Wang, and Hui Huang. Ranked voting based self-consistency of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14410–14426, 2025. doi: 10.18653/v1/2025.findings-acl.744. URL https://aclanthology.org/2025.findings-acl.744/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://openreview.net/forum?id=1PL1NIMMrw.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992. doi: 10.1016/s0893-6080(05)80023-1. URL https://doi.org/10.1016/s0893-6080(05)80023-1.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hong-Wei Zheng, and Zhiming Zheng. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25907–25926, 2026. doi: 10.18653/v1/2026.acl-long.1188. URL https://aclanthology.org/2026.acl-long.1188/.

A FEATURE DEFINITIONS AND FITTING PROTOCOL

he selector constructs 25 features per observed candidate. Table 4 specifies the representation. To simplify notation within this appendix, suppress the question index and write $c = c_i(a)$, c_1 and c_2 for the largest and second-largest candidate counts, $K = |A_i|$, and N_v for the number of valid outputs. Set $c_2 = 0$ if $K = 1$. Let $T_a = \{t : \hat{y}_{it} = a\}$, $o = \mathbf{1}[a = \hat{y}_{i0}]$, $v = c_1/20$, and m indicate the top candidate from a frequency count with ties resolved by first occurrence. Define

$$H = -\sum_{b \in A_i} \frac{c_i(b)}{N_v} \log_2\left(\frac{c_i(b)}{N_v} + 10^{-12}\right), \qquad \bar{H} = \begin{cases} H/\log_2 K, & K > 1, \\ 0, & K = 1. \end{cases} \tag{3}$$

The frequency rank is $r = 1 + \#\{b : c_i(b) > c\}$, so tied counts share a rank. This count-based majority indicator is distinct from the final Majority-vote tie rule, which uses original-order preference and descending answer strings.

Table 4: Exact candidate feature specification for 20-order pools. Indices are zero-based call indices. The strict-unanimity flag is routing metadata and is not a learned feature. CG-OAS removes fitting-set constant columns, leaving 24 active dimensions in each selected configuration because feature 16 is constant. No-gate retains the 25 specified columns.

ID	FEATURE	DEFINITION
1	Candidate frequency	$c/20$
2	Majority indicator	m (first-occurring tied maximum)
3	Majority margin	$c_1 - c_2$ (counts)
4	Candidate margin to top	$c - c_1$ (counts)
5	Frequency rank	r
6	Unique answers	K
7–8	Entropy; normalized entropy	H; $\bar{H}$
9–11	Occurrence indices	$\min T_a$; $\max T_a$; mean of T_a
12–14	Original/first-answer indicators	Three identical copies of o
15	Last-answer indicator	$\mathbf{1}[a = \hat{y}_{i,19}]$
16	Candidate parse-success ratio	Successful parses within T_a, divided by $\lvert T_a \rvert$
17	Maximum vote fraction	$v = c_1/20$
18	Low-agreement indicator	$\mathbf{1}[v < 0.6]$
19	Answer prior	Validation gold-label frequency of a
20	Prediction prior	Frequency of a in valid validation outputs
21	Frequency–entropy interaction	$(c/20)\bar{H}$
22	Margin–agreement interaction	$(c - c_1)v$
23	Majority–support interaction	mK
24	Rank–entropy interaction	rH
25	Original–agreement interaction	ov

Both priors use the full validation pool without smoothing; unseen labels receive zero. Standardization subtracts the fitting-set feature mean and divides by its population standard deviation. Missing feature values are replaced by zero before transformation. The three original-answer indicators remain duplicated in the representation. Thus 25 specified dimensions should not be interpreted as 25 independent signals. Class balancing uses inverse candidate-class frequency; there is no additional question-normalized weighting of candidate rows.

A.1 QUESTION SPLITS AND CANDIDATE ROWS

Validation contains 137 non-unanimous MedMCQA questions and 121 non-unanimous MedQA questions. The CG-OAS fit/dev question counts are 109/28 and 96/25, respectively. They expand to 237/60 and 207/54 candidate rows. No-gate fits on 400 questions and develops on 100 per dataset, producing 526/134 MedMCQA and 503/137 MedQA rows. Candidate rows, not questions, are the classifier training units; development accuracy is measured once per question after candidate selection.

A.2 Model configuration and selection

Logistic regression searches $C \in \{0.01, 0.1, 1, 10\}$ and class weighting in $\{\texttt{balanced}, \texttt{none}\}$, retaining the first configuration at a tied best development accuracy. Its default solver is L-BFGS with L2 regularization. Both selected MedMCQA variants use $C = 0.01$ and balanced weights. The random forest has 200 trees, depth 6, minimum leaf size 2, balanced weights, and random state 42. Gradient boosting, included in CG-OAS family selection, has 200 estimators, depth 3, learning rate 0.05, and random state 42. Family selection uses full test accuracy; this choice is reflected in the conditional interpretation of the results. The selector environment uses Python 3.8.10, NumPy 1.24.4, pandas 2.0.3, scikit-learn 1.3.2, and SciPy 1.10.1.

A.3 Ties and exceptional inputs

Selector rows are sorted by positive-class probability, candidate frequency, and original-answer indicator, all descending; residual ties follow the candidate-row ordering of the implementation. The parser first seeks explicit answer prefixes, including "Final answer", "The answer is", and "Answer". If these fail, it searches whole-word labels in label-list order (A–E for these tasks), rather than selecting the leftmost answer mention in the response. An unparseable response becomes an empty prediction. Empty candidate sets retain an empty sentinel row and count as wrong; a missing combined selector prediction falls back to the majority answer. All these counts are zero on the main evaluation subsets.

B Generation, evidence, and coverage

The fixed generator uses the epoch-3 MedMCQA adapter at training step 1,173 and the epoch-3 MedQA adapter at step 432. Inference uses batch size 8, left padding, the tokenizer's chat template with a generation prompt, and truncation to 4,096 tokens. Padding uses the end-of-sequence token when no padding token is defined. The input contains the retrieved documents followed by the question, options, and an answer cue. Passage titles and content are rendered in the order specified by the permutation. The system instruction is:

> You are a medical question answering assistant. Use the retrieved documents below as helpful background information. Return only the final answer label. Do not provide explanation.

The user instruction requests the correct multiple-choice option label, then lists documents, the question, and options, ending with "Final answer:". The model's own chat template formats the system and user messages. The evidence set is fixed before prompt construction. The 4,096-token limit can make retained context depend on order for long prompts; the study concerns the full order-conditioned inference pipeline, rather than an isolated positional mechanism under guaranteed identical retained tokens.

B.1 Coverage and denominators

Table 5 distinguishes registry coverage from answer coverage. The three MedMCQA registry IDs excluded for empty retrieval are:

```
3cf954d4-ea8b-4c75-b70d-2e86af595b9c
8c5568c4-2a9d-4a89-97ba-ce39ee127ceb
cf45037e-e487-43c9-aac6-203387a25b55
```

The exclusions precede generation and do not depend on labels. Normalized-question matching is one-to-one for all included questions, with no mapping collisions. No subsequent parser or candidate-set exclusion changes the evaluation subset for any method.

Table 5: Coverage under the canonical protocol. Every included question contributes 20 successfully parsed outputs, and all methods share its denominator.

QUANTITY	MEDMCQA	MEDQA
Registry questions	4,183	1,273
Evaluation questions (all methods)	4,180	1,273
Registry coverage	99.9283%	100%
Answer coverage in evaluation subset	100%	100%
Generation records	83,600	25,460
Questions with 5 / 3 / 1 passages	4,178 / 1 / 1	1,273 / 0 / 0
Questions with partial / total parse failure	0 / 0	0 / 0
Empty candidate sets / sentinel predictions	0 / 0	0 / 0

Table 6: Accuracy by pool agreement. Each accuracy uses the row's question count as its denominator. Predictions are identical on unanimous questions; all net gains occur in the non-unanimous groups.

DATASET	POOL	QUESTIONS	MAJORITY-VOTE	CG-OAS
MedMCQA	Unanimous	2,901	0.6801	0.6801
MedMCQA	Non-unanimous	1,279	0.3729	0.4011
MedQA	Unanimous	972	0.6872	0.6872
MedQA	Non-unanimous	301	0.3854	0.4319

B.2 Repetition under the sampling model

For $k = 5$, there are $5! = 120$ index permutations. With two distinct fixed orders followed by 18 uniform draws, the probability of at least one repetition is

$$1 - \prod_{j=2}^{19} \frac{120 - j}{120} = 0.81303287. \tag{4}$$

The observed frequencies are 3,379/4,178 (80.8760%) for five-passage MedMCQA questions and 1,016/1,273 (79.8115%) for MedQA. Mean distinct counts are 18.5122 and 18.5082, with range 14–20 in both datasets. The two remaining MedMCQA questions have six and one distinct permutations, respectively, and all 20 predictions are correct and unanimous for each. Repetition is a consequence of the prescribed sampling design.

C Counts and paired uncertainty

Exact correct counts on MedMCQA are 2,450, 2,450, 2,469, and 2,486 for Original-order, Majority-vote, No-gate, and CG-OAS. The corresponding MedQA counts are 776, 784, 795, and 798. Macro accuracy averages the two exact ratios before rounding. Table 6 separates unanimous and non-unanimous questions, while Table 7 reports per-dataset paired intervals.

Resampling preserves each question's paired correctness indicators across methods. If d_{ij} is the correctness difference on question i in dataset j, replicate b computes $\Delta^{(b)} = \frac{1}{2}\sum_{j=1}^{2} n_j^{-1} \sum_{i \in I_j^{(b)}} d_{ij}$, where $I_j^{(b)}$ is a size-n_j sample with replacement. The 2.5th and 97.5th percentiles define the interval. Resampling does not repeat family selection, selector fitting, or generator inference.

D Sensitivity analyses

Table 8 reports macro accuracy under four changes to the canonical evaluation, and Table 9 gives their contrasts. The five-passage subset excludes two MedMCQA questions. The registry sensitivity counts all three retrieval-empty questions as wrong for every method. Excluding repeated-output-unstable questions leaves 4,154 MedMCQA and 1,268 MedQA questions. These three diagnostics retain the fitted predictions and change only the evaluation subset or denominator.

Table 7: Per-dataset paired bootstrap for fixed selected outputs (10,000 replicates; seed 20260703). Differences are reported in percentage points.

DATASET	COMPARISON	DIFFERENCE (PP)	95% CI (PP)
MedMCQA	CG-OAS − Original-order	+0.86	[0.19, 1.53]
MedMCQA	CG-OAS − Majority-vote	+0.86	[0.29, 1.44]
MedMCQA	CG-OAS − No-gate selector	+0.41	[−0.05, 0.86]
MedQA	CG-OAS − Original-order	+1.73	[0.08, 3.30]
MedQA	CG-OAS − Majority-vote	+1.10	[−0.31, 2.51]
MedQA	CG-OAS − No-gate selector	+0.24	[−1.02, 1.49]

Table 8: Macro-accuracy sensitivity under alternative evaluation protocols. Entries are macro accuracies. The canonical protocol uses 20 records with replacement; $k = 5$ evaluates five-order pools; registry sensitivity counts empty retrieval as wrong; unstable-question exclusion removes unstable questions; unique permutations deduplicate order pools.

SETTING	ORIGINAL-ORDER	MAJORITY-VOTE	NO-GATE	CG-OAS
Canonical	0.5979	0.6010	0.6076	0.6108
$k = 5$ only	0.5978	0.6009	0.6075	0.6107
Registry, empty retrieval counted wrong	0.5976	0.6008	0.6074	0.6106
Exclude unstable questions	0.5986	0.6021	0.6092	0.6112
Unique permutations	0.5979	0.6026	0.6081	0.6065

Table 9: CG-OAS macro-accuracy differences under alternative evaluation protocols. Differences are reported in percentage points.

SETTING	VS. ORIGINAL-ORDER	VS. MAJORITY-VOTE	VS. NO-GATE
Canonical	+1.29	+0.98	+0.32
$k = 5$ only	+1.29	+0.98	+0.32
Registry, empty retrieval counted wrong	+1.29	+0.98	+0.32
Exclude unstable questions	+1.26	+0.92	+0.20
Unique permutations	+0.87	+0.40	−0.15

Repeated-output statistics. MedMCQA contributes 27 mismatches among 6,740 repeated-permutation pairs; MedQA contributes six among 2,002. All 31 affected questions are non-unanimous. CG-OAS and Majority-vote differ on six affected questions and answer 13 and 12 of the 31 correctly, respectively. The mismatch intervals summarize pair-level agreement and do not assume that a high agreement rate guarantees correct answers.

Deduplication and refitting. We group records by their passage-index permutation. A group contributes the most frequent parsed label; ties use the earliest occurrence. The original permutation stays first, the reverse stays second when distinct, and remaining permutations follow first appearance. The resulting calls are reindexed. Candidate frequencies, maximum vote fraction, and unanimity use the question-specific distinct-order count. Priors and features are recomputed from deduplicated validation pools, and both selectors are refitted. CG-OAS retains the same family-selection procedure, selecting a random forest for MedMCQA and logistic regression for MedQA; No-gate uses its original family choices. Thus this analysis measures sensitivity of the complete aggregation protocol, not only removing duplicate test votes from a fixed trained selector.

Table 10: Per-dataset accuracy with unique-permutation pools on the original evaluation subsets. Entries are accuracies.

DATASET	ORIGINAL-ORDER	MAJORITY-VOTE	NO-GATE	CG-OAS
MedMCQA	0.5861	0.5861	0.5892	0.5933
MedQA	0.6096	0.6190	0.6269	0.6198