CITATION-CONSISTENT VOTING FOR PERMUTATION-ROBUST RETRIEVAL-AUGMENTED GENERATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Retrieval-augmented generation (RAG) systems are sensitive to the ordering of retrieved documents, with language models exhibiting position bias that causes inconsistent outputs across different document permutations. We propose Citation-Consistent Voting (CCV), a training-free method that improves RAG robustness by aggregating answers based on document-ID agreement rather than answer frequency. CCV prompts the generator to cite supporting documents, maps citations to permutation-invariant document identifiers, and selects answers with the highest citation consistency across multiple permutations. On NaturalQuestions with Qwen3-8B and Contriever retrieval, CCV achieves 46.37% SubEM, outperforming majority voting by +0.19 points at $K = 20$ permutations. The improvement scales monotonically with the number of permutations, and diagnostic analysis confirms that citation agreement correlates significantly with answer correctness ($p = 1.14 \times 10^{-5}$). CCV requires no additional training and is compatible with any RAG system that can produce structured citations.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Retrieval-augmented generation (RAG) has emerged as a powerful paradigm for improving the factual accuracy of large language models by grounding responses in retrieved documents (Gao et al., 2023). However, RAG systems exhibit a critical vulnerability: sensitivity to the ordering of retrieved documents. Liu et al. (2023) demonstrated the "lost in the middle" phenomenon, where models struggle to utilize information positioned in the middle of the context. More recently, Zhang et al. (2026) showed that even in short contexts with only five documents, merely permuting the document order can induce qualitatively different answers, including confident but incorrect responses.

Existing approaches to address this permutation sensitivity fall into two categories. Training-based methods like Stable-RAG (Zhang et al., 2026) use preference optimization to make generators invariant to document ordering, but require access to model weights and training data. Inference-time methods like Mixture-of-Agents (Wang et al., 2024) generate multiple responses under different document orderings and aggregate via majority voting, but this approach ignores the rich citation structure that modern RAG systems produce.

We observe that when language models cite evidence for their answers, the document they cite provides a more stable signal than the answer text itself. Our key insight is that correct answers tend to cite the same underlying document across permutations, while incorrect answers show more citation variance. This motivates a new aggregation strategy: rather than voting on answer frequency, we vote on document-ID agreement among evidence-valid responses.

We propose Citation-Consistent Voting (CCV), a training-free method that aggregates answers based on document-ID agreement rather than answer frequency. CCV filters responses for evidence validity (ensuring the answer is grounded in the cited document), maps citations to permutation-invariant document IDs, and selects the answer with the highest citation-consistency score. Our contributions are:

[1] https://gitlab.com/fars-a/latent-mode-voting-rag

- We propose Citation-Consistent Voting, a training-free inference-time method for permutation-robust RAG that exploits citation structure for answer aggregation.
- We demonstrate that CCV improves SubEM by +0.19 over majority voting on NaturalQuestions at $K = 20$ permutations, with gains scaling monotonically with K.
- We validate that document-ID agreement correlates significantly with answer correctness ($p = 1.14 \times 10^{-5}$), confirming that citation consistency is a meaningful quality signal.

2 Related Work

2.1 Position Bias in Language Models

Large language models exhibit systematic biases in how they process information based on its position within the input context. Liu et al. (2023) demonstrated the "lost in the middle" phenomenon, where models perform best when relevant information appears at the beginning or end of the context, with significant performance degradation for information in the middle. This position sensitivity is particularly problematic for retrieval-augmented generation, where the ordering of retrieved documents can substantially affect answer quality. Our work addresses this challenge through inference-time aggregation rather than architectural modifications.

2.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) enhances language model outputs by grounding responses in retrieved documents (Gao et al., 2023). Dense retrieval methods such as DPR (Karpukhin et al., 2020) and Contriever (Izacard et al., 2021) have enabled effective passage retrieval for open-domain question answering. However, RAG systems inherit the position biases of their underlying language models, making them sensitive to the ordering of retrieved documents. Recent work has explored various strategies to improve RAG robustness, including training-based approaches (Zhang et al., 2026) and inference-time methods (Lee et al., 2024).

2.3 Self-Consistency and Ensemble Methods

Self-consistency (Wang et al., 2022) improves chain-of-thought reasoning by sampling multiple reasoning paths and selecting the most consistent answer through majority voting. This principle has been extended to multi-agent systems, where Mixture-of-Agents (Wang et al., 2024) aggregates outputs from multiple language models to improve response quality. Tang et al. (2023) applied permutation self-consistency to listwise ranking, demonstrating that aggregating predictions across different input orderings can mitigate position bias. Our citation-consistent voting extends this paradigm by using document citations rather than answer frequency as the aggregation signal.

2.4 Permutation-Robust RAG

Several approaches have been proposed to address permutation sensitivity in RAG systems. StableRAG (Zhang et al., 2026) uses preference optimization to train models that produce consistent outputs regardless of document ordering, but requires access to model weights and training data. Tang et al. (2023) proposed permutation self-consistency for ranking tasks, aggregating predictions across multiple orderings. Our approach differs by exploiting the citation structure in model outputs, using document-ID agreement rather than answer frequency for aggregation, and requiring no training.

3 Method

We propose Citation-Consistent Voting (CCV), a training-free method for improving RAG robustness by aggregating answers based on document-ID agreement rather than answer frequency. Figure 1 illustrates the three-stage pipeline.

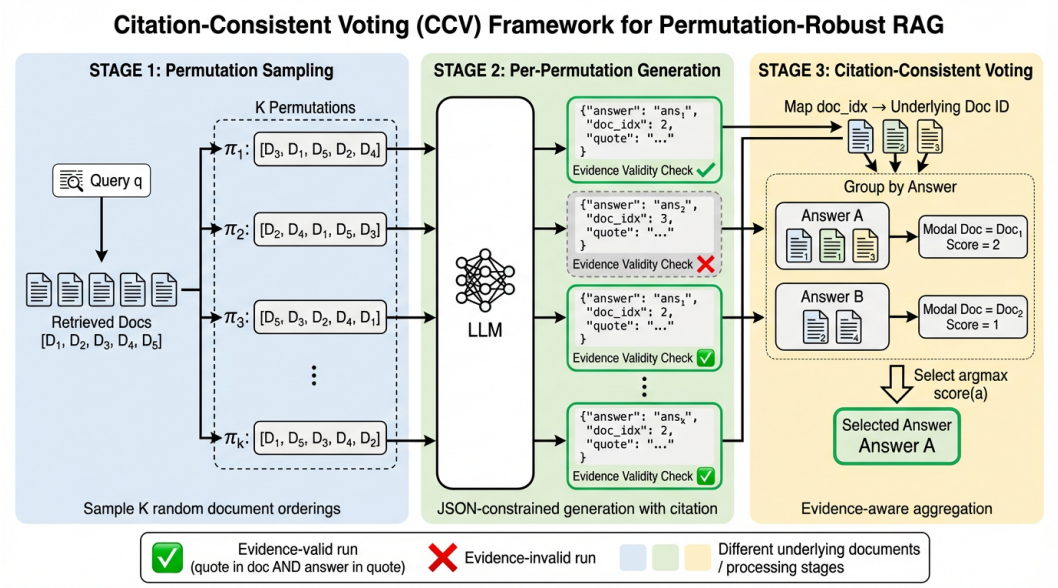

Figure 1: Overview of Citation-Consistent Voting (CCV) for permutation-robust RAG. Given a query and retrieved documents, we generate K responses under different document orderings, filter for evidence-valid outputs (where the answer is grounded in a cited document), and aggregate by document-ID agreement rather than answer frequency.

3.1 Problem Setup

Consider a query q and a set of N retrieved documents $\mathcal{D} = \{d_1, \ldots, d_N\}$. Standard RAG generates a single response by conditioning on a fixed ordering of $\mathcal{D}$. However, the generator's output can vary significantly depending on document order due to position bias (Liu et al., 2023). Multi-permutation RAG addresses this by sampling K random permutations $\pi_1, \ldots, \pi_K$ of the document indices and generating K responses, one for each ordering. The challenge is then to aggregate these K responses into a single, more reliable answer.

3.2 Evidence Filtering

We prompt the generator to produce structured outputs containing an answer, a document index, and a supporting quote. Specifically, each response r_k for permutation π_k contains: (1) an answer string a_k, (2) a document index $j_k \in \{1, \ldots, N\}$ indicating which document in the permuted order supports the answer, and (3) a quote e_k extracted from that document.

A response is considered *evidence-valid* if it satisfies two criteria. First, the cited document index must be valid ($j_k \in \{1, \ldots, N\}$). Second, the answer must be grounded in the cited document, which we verify by checking that the quote e_k appears in document $d_{\pi_k(j_k)}$ and that the answer a_k appears in the quote e_k. Responses failing these checks are filtered out before aggregation.

3.3 Citation-Consistent Aggregation

The key insight of CCV is that correct answers tend to cite the same underlying document across permutations, while incorrect answers show more citation variance. We exploit this by aggregating based on document-ID agreement rather than answer frequency.

For each permutation π_k, we map the cited document index j_k to an underlying document ID that is invariant to permutation: $\text{docID}(r_k) = \pi_k(j_k)$. This mapping ensures that citations to the same document are recognized as such regardless of where that document appeared in the permuted order.

Table 1: Main results on NaturalQuestions (3,610 test queries) with Qwen3-8B and Contriever Top-5 retrieval. Citation-consistent voting achieves the highest SubEM and F1 among multi-permutation methods at $K = 20$.

Method	K	SubEM (%)	F1 (%)
Vanilla RAG (Standard)	1	48.98	41.10
Vanilla RAG (JSON Citation)	1	45.10	45.18
MoA Majority Vote	5	45.82±0.28	46.38±0.21
MoA Majority Vote	20	46.18	46.66
Citation-Consistent Vote (Ours)	20	**46.37**	**46.83**

For each unique answer a, let $R(a)$ denote the set of evidence-valid responses producing answer a. We compute the modal document ID among these responses:

$$\text{doc}^*(a) = \arg\max_d |\{r \in R(a) : \text{docID}(r) = d\}| \tag{1}$$

The citation-consistency score for answer a is then defined as the number of evidence-valid responses that cite this modal document:

$$\text{score}(a) = |\{r \in R(a) : \text{docID}(r) = \text{doc}^*(a)\}| \tag{2}$$

The final answer is selected as $\arg\max_a \text{score}(a)$, with ties broken by raw answer frequency across all K runs.

3.4 Practical Considerations

Quote Constraint Relaxation. The strict quote verification (requiring exact substring match) can reduce citation path coverage when models produce paraphrased or slightly modified quotes. We find that relaxing this constraint to only require valid document index (without quote verification) improves coverage while maintaining the benefit of document-ID agreement. We evaluate both variants in our experiments.

Computational Cost. CCV requires K forward passes through the generator, the same as standard multi-permutation approaches like Mixture-of-Agents (Wang et al., 2024). The additional overhead of evidence filtering and citation-consistent aggregation is negligible compared to generation cost. The method is compatible with batched inference and can leverage KV-cache sharing across permutations that share document prefixes.

4 Experiments

4.1 Experimental Setup

We evaluate Citation-Consistent Voting on the NaturalQuestions (Kwiatkowski et al., 2019) open-domain QA benchmark, using 3,610 test queries. For retrieval, we use Contriever (Izacard et al., 2021) to retrieve the top-5 passages from Wikipedia. The generator is Qwen3-8B with greedy decoding (temperature=0). We report Substring Exact Match (SubEM), where a prediction is correct if the gold answer appears as a substring, and token-level F1 score.

We compare against two baselines. Vanilla RAG generates a single response without permutation ensembling. MoA Majority Vote (Wang et al., 2024) generates K responses under different document orderings and selects the most frequent answer. For multi-permutation methods, we evaluate at $K \in \{5, 10, 15, 20\}$ permutations.

4.2 Main Results

Table 1 presents the main results. Citation-consistent voting achieves 46.37% SubEM and 46.83% F1 at $K = 20$, outperforming MoA majority voting by +0.19 SubEM and +0.17 F1. The improvement is consistent across both metrics, demonstrating that document-ID agreement provides a more

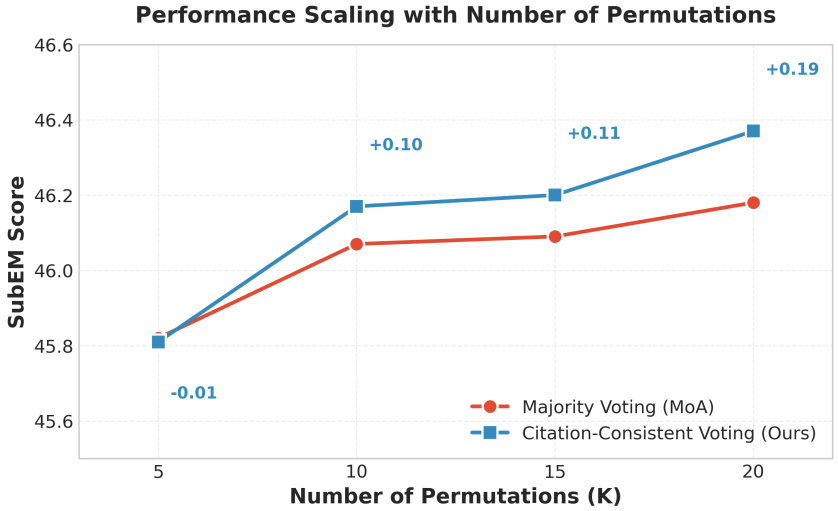

Figure 2: Performance scaling with number of permutations (K) on NaturalQuestions. Citation-consistent voting shows monotonically increasing improvement over majority voting as K increases, reaching +0.19 SubEM at $K = 20$.

Table 2: K-scaling analysis showing SubEM at different numbers of permutations. Δ indicates improvement of citation-consistent voting over majority voting. Improvement scales monotonically with K.

K	MoA SubEM (%)	Citation SubEM (%)	Δ
5	45.82	45.81	-0.01
10	46.07	46.17	+0.10
15	46.09	46.20	+0.11
20	46.18	46.37	**+0.19**

reliable aggregation signal than answer frequency alone. We note that the standard Vanilla RAG prompt achieves higher SubEM (48.98%) but lower F1 (41.10%) compared to the JSON citation prompt, reflecting differences in output format rather than method quality. Among methods using the same JSON citation format, multi-permutation approaches consistently outperform single-pass generation.

4.3 K-Scaling Analysis

Figure 2 and Table 2 show how the improvement of citation-consistent voting over majority voting scales with the number of permutations. At $K = 5$, the two methods perform nearly identically ($\Delta = -0.01$), but as K increases, citation-consistent voting shows monotonically increasing gains: +0.10 at $K = 10$, +0.11 at $K = 15$, and +0.19 at $K = 20$. This scaling behavior suggests that more permutations provide more reliable citation agreement signals, as the document-ID voting mechanism benefits from a larger sample of evidence-valid responses to identify consistent citation patterns.

4.4 Ablation: Quote Constraint

Table 3 compares citation-consistent voting with and without the verbatim quote constraint at $K = 5$. Requiring exact quote matches reduces citation path coverage to 96.37% (the fraction of queries with at least one evidence-valid response) and yields only 4.12 valid runs per query on average. Relaxing this constraint to only require valid document indices increases coverage to 99.70% and valid runs to 4.96, resulting in a substantial improvement from 45.73% to 46.37%

Table 3: Ablation study on quote requirement at $K = 5$. Removing the verbatim quote constraint improves performance by increasing citation path coverage.

Method	SubEM (%)	Citation Path (%)	Avg Valid Runs
MoA Majority Vote	45.71	–	–
Citation Vote (with quote)	45.73	96.37	4.12
Citation Vote (no quote)	**46.37**	99.70	4.96

SubEM (+0.64). This demonstrates that the document-ID agreement signal is more important than strict quote verification for effective aggregation.

4.5 Diagnostic Analysis

To validate the core hypothesis that citation agreement correlates with answer correctness, we analyze the relationship between document-ID agreement scores and prediction accuracy. Correct predictions have significantly higher mean agreement scores than incorrect ones (3.04 vs. 2.85, Mann-Whitney $p = 1.14 \times 10^{-5}$), confirming that citation consistency is a meaningful quality signal.

We further examine where citation-consistent voting differs from majority voting. The method primarily acts on unstable queries, defined as those producing three or more unique answers across permutations. Among the 699 unstable queries (19.4% of the test set), citation-consistent voting selects a different answer than majority voting in 19.89% of cases. This targeted behavior confirms that the method intervenes precisely where frequency-based voting is least reliable, while leaving stable queries unchanged.

5 Conclusion

We presented Citation-Consistent Voting, a training-free method for improving RAG robustness by aggregating answers based on document-ID agreement rather than answer frequency. On NaturalQuestions, CCV achieves +0.19 SubEM improvement over majority voting at $K = 20$ permutations, with gains scaling monotonically with K. Diagnostic analysis confirms that citation agreement correlates significantly with answer correctness ($p = 1.14 \times 10^{-5}$).

Our work has several limitations. The improvement, while consistent, is modest in absolute terms. Evaluation is limited to a single dataset and model. The method requires K forward passes, adding computational cost. Future work could explore combining CCV with training-based approaches like Stable-RAG (Zhang et al., 2026), evaluating on additional datasets, and investigating other citation signals beyond document-ID agreement.

References

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022, 2021.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020.

T. Kwiatkowski, J. Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, D. Epstein, I. Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural

questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Youngwon Lee, Seung won Hwang, Daniel F. Campos, Filip Grali'nski, Zhewei Yao, and Yuxiong He. Inference scaling for bridging retrieval and augmented generation. *ArXiv*, abs/2412.10684, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, F. Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023.

Raphael Tang, Xinyu Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *ArXiv*, abs/2310.07712, 2023.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *ArXiv*, abs/2406.04692, 2024.

Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *ArXiv*, abs/2601.02993, 2026.