Stable-RAG: Mitigating Retrieval-Permutation-Induced Hallucinations in Retrieval-Augmented Generation

Qianchi Zhang[1,2], Hainan Zhang[1,2]✉, Liang Pang[4], Hongwei Zheng[3], Zhiming Zheng[1,2]
[1]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing
[2]School of Artificial Intelligence, Beihang University, China
[3]Beijing Academy of Blockchain and Edge Computing, China
[4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{zhangqianchi, zhanghainan}@buaa.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) has become a key paradigm for reducing factual hallucinations in Large Language Models (LLMs), yet little is known about how the order of retrieved documents affects model behavior. We empirically show that under a Top-5 retrieval setting with the gold document included, LLM answers vary substantially across permutations of the retrieved set, even when the gold document is fixed in the first position. This reveals a previously underexplored sensitivity to retrieval permutations. Although existing robust RAG methods focus primarily on enhancing LLM robustness to low-quality retrieval and mitigating positional bias to distribute attention fairly over long contexts, neither approach directly addresses permutation sensitivity. In this paper, we propose **Stable-RAG**, which exploits permutation sensitivity estimation to mitigate permutation-induced hallucinations. Stable-RAG runs the generator under multiple retrieval orders, clusters hidden states, and decodes from a cluster-center representation that captures the dominant reasoning pattern. It then uses these reasoning results to align hallucinated outputs toward the correct answer, encouraging the model to produce consistent and accurate predictions across document permutations. Experiments on three QA datasets show that Stable-RAG improves answer accuracy, reasoning consistency, and generalization across datasets, retrievers, and input lengths compared with strong baselines [1].

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance on language understanding and generation tasks, but still often generate confident yet incorrect statements, known as hallucinations (Fan et al., 2024; Xu et al., 2026), especially in knowledge-intensive settings (Chen et al., 2022; Huang et al., 2023; Zhang et al., 2024c; Chen et al., 2025; Luo et al., 2025; Li et al., 2025b; Yuan and Zhang, 2025b; Ji et al., 2026; Hu et al., 2026). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Gao et al., 2024; Li et al., 2026a) reduces factual hallucinations by grounding model outputs in externally retrieved documents rather than relying only on parametric knowledge, improving factuality, interpretability, and updatability without additional retraining (Zhou et al., 2024).

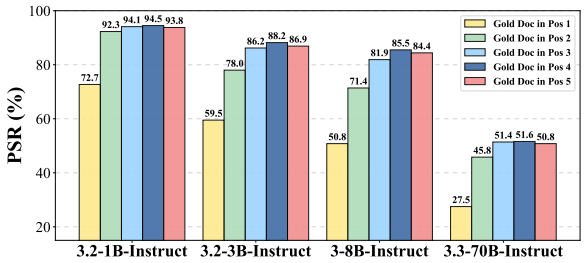


Figure 1: **Perturbation Success Rate (PSR)** on the NQ test set across different LLaMA models. PSR is computed as the proportion of successful document-order perturbations to produce hallucination results among 1,000 randomly sampled instances, with the gold document fixed in different positions. Qwen models' results can be seen in Appendix C.1.

Despite these benefits, RAG systems are far from hallucination-free (Hamman et al., 2025). We identify a critical but overlooked vulnerability in existing RAG systems: a strong sensitivity to the order of retrieved documents. When the retrieved content remains the same, with the gold document included, merely reordering them can lead the model to follow entirely different reasoning paths and produce inconsistent answers, referred to as **Permutation-Induced Hallucinations**. As shown in Figure 1, we retrieve the Top-5 documents (Xu et al., 2024; Zhu et al., 2024b), place the gold document in different positions, and observe that LLM answers vary substantially across retrieval permutations.

✉Corresponding author.

[1] Our code is available at https://github.com/zqc1023/Stable-RAG.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25907–25926
July 2-7, 2026

Even when the gold document is fixed first, models may ignore it and produce answers that conflict with the evidence. This reveals a previously underexplored sensitivity to retrieval permutations, even in short contexts under 1,000 tokens.

Existing robust RAG methods mainly focus on retrieval quality and positional bias. The former enhances LLM robustness to low-quality retrieval via uncertainty estimation and adversarial training, such as noise injection (Fang et al., 2024; Yoran et al., 2024) of weakly relevant documents. The latter alleviates attention bias toward specific positions in long contexts, promoting more balanced use of retrieved documents (Zhang et al., 2024d; Wang et al., 2025b). However, these approaches overlook a critical issue: permutation sensitivity is neither caused by weakly relevant documents, because the input documents are the same, nor confined to long-context reasoning tasks, since only the Top-5 documents fall within one thousand tokens.

Instead, permutation sensitivity stems from structural instability in the internal reasoning dynamics of LLMs. As model depth increases, document permutations induce a growing number of distinct reasoning trajectories, leading to more frequent branching and a higher risk of hallucinations or unreliable outputs. As shown in Figure 2, we measure the average number of clusters obtained via spectral clustering over document-permuted representations across different LLM layers on the NQ and HotpotQA datasets. The results indicate that reasoning trajectories in shallow layers are relatively concentrated, while divergence emerges in the middle layers and becomes more pronounced in higher layers. Furthermore, sensitive samples (i.e., 10+) exhibit substantially greater divergence than non-sensitive ones (i.e., 1-2), with this effect primarily localized to the higher layers. These findings highlight the importance of mitigating permutation sensitivity, enabling LLMs to produce stable and accurate outputs regardless of the ordering of retrieved documents, which is critical for improving the robustness of RAG systems.

In this paper, we introduce **Stable-RAG** that explicitly leverages permutation sensitivity estimation to mitigate the permutation-induced hallucinations. Specifically, we apply spectral clustering to the last token hidden states of the final layer before response generation, across all document permutations to identify dominant reasoning clusters. For each cluster, we select a representative hidden state and decode it to obtain candidate answers, thereby

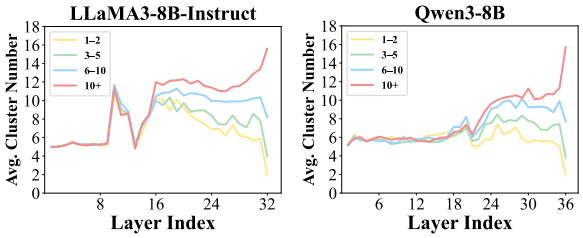


Figure 2: Hidden-state clustering behaviors across layers for LLaMA3-8B-Instruct on the NQ train set with DPR retriever and Qwen3-8B on the HotpotQA train set with Contriever retriever, using 1,000 randomly sampled instances. Different colored lines indicate the number of clusters of final reasoning states produced by the LLM under all $5!(=120)$ permutations of the Top-5 retrieved documents (e.g., the green line indicates 3–5 cluster states). Other scales are reported in Appendix C.2.

capturing the model's core reasoning modes. Then, we perform cross-cluster consistency alignment over these candidates, encouraging the model to prioritize semantically consistent and factually correct answers across different document orders. This cluster-based alignment substantially reduces the uncertainty induced by order perturbations and fundamentally improves the robustness of RAG.

Experiments on three QA datasets demonstrate that Stable-RAG improves answer accuracy, reasoning consistency, and generalization across datasets, retrievers, and input lengths compared with strong baselines.

Our main contributions are as follows:

- We find that RAG systems are highly sensitive to document order, leading to inconsistent reasoning. Layer-wise hidden state clustering reveals divergent reasoning trajectories across layers.
- We propose Stable-RAG, which can mitigate permutation-induced hallucinations, achieving model-agnostic stable reasoning.
- Across three QA datasets, Stable-RAG outperforms strong baselines in accuracy and reasoning consistency.

2 Related Work

RAG mitigates factual hallucinations (Yang et al., 2026; Kong et al., 2026) in LLMs on knowledge-intensive tasks by providing explicit evidence from external documents (Lewis et al., 2020; Fan et al., 2024; Qian et al., 2025; Yuan et al., 2026; Luo et al., 2026b; Ge et al., 2026). Prior work on im-

proving the robustness of RAG systems has primarily focused on enhancing retrieval quality (Liu et al., 2021a; Xu et al., 2024; Ma et al., 2024; Li et al., 2026b) and reranking performance (Liu et al., 2021b, 2025), or strengthening the generator's robustness. For instance, AdaComp (Zhang et al., 2024b) and CompSelect (Zhang et al., 2026c) apply noise filtering to boost generation accuracy. RetRobust (Yoran et al., 2024) and RAAT (Fang et al., 2024) expose the model to retrieval noise or irrelevant documents during training, enhancing robustness. However, these methods generally assume a stable document order and do not systematically assess its impact on reasoning. Although ATM (Zhu et al., 2024a) considers order perturbations, it does not explicitly model reasoning trajectories across permutations and thus cannot ensure consistency.

In addition, another line of research focuses primarily on positional bias in long-context scenarios. Most LLMs use relative positional encodings (Peysakhovich and Lerer, 2023), such as RoPE (Su et al., 2024) or ALiBi (Press et al., 2021), which introduce systematic biases: early tokens receive excessive attention due to attention sinks (Xiao et al.; Gu et al.), while long-range decay favors recent tokens. Prior work mitigates these issues by modifying positional encodings (Zhang et al., 2024d; Chen et al., 2024; Lin et al., 2024; Egressy and Stühmer, 2025), adjusting causal masks, reweighting attention or hidden states (Hsieh et al., 2024), or using Pos2Distill (Wang et al., 2025b) to distill knowledge from advantageous to less favorable positions to promote fair attention across tokens. These methods focus on long contexts and do not explicitly address reasoning inconsistency induced by different permutations of the same document set.

3 Preliminary Study

3.1 Problem Formulation

Given a query q and its retrieved document set $\mathcal{S} = \{d_1, d_2, \ldots, d_n\}$, the goal is to ensure that the model f_θ produces consistent outputs across different document orderings. Let $\mathrm{Perm}(\mathcal{S})$ denote all possible permutations of $\mathcal{S}$. For any two permutations $\pi_1, \pi_2 \in \mathrm{Perm}(\mathcal{S})$, the model's outputs should be as similar as possible:

$$f_\theta(q, \pi_1) \approx f_\theta(q, \pi_2).$$

In this task, the model is expected to produce consistent answers regardless of the document order.

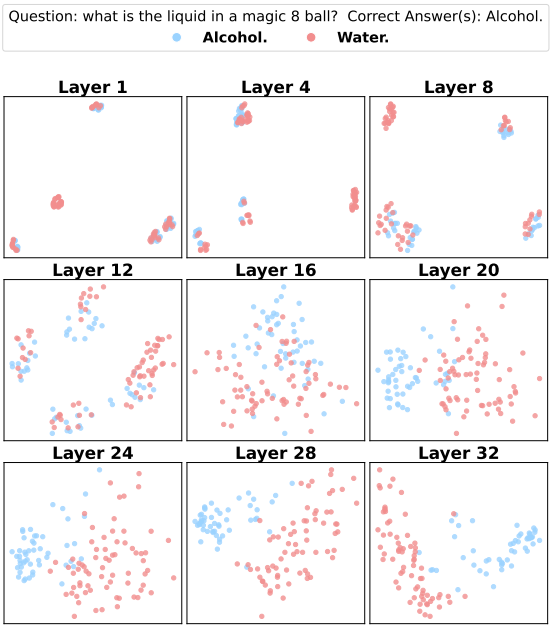


Figure 3: Layer-wise visualization of a case study from the NQ train set using LLaMA-3-8B-Instruct. Each point corresponds to a document order, and its color represents the model's final answer.

3.2 Permutation Sensitivity Estimation

Recent work (Liang et al., 2025; Lee et al., 2025) exploits hidden states to uncover latent reasoning trajectories, often as indicators of generative uncertainty. Accordingly, we propose to quantify model generation uncertainty via the spectral clustering of hidden states. In this section, we validate the feasibility of the spectral clustering (Ng et al., 2001; Von Luxburg, 2007) through both layer-wise visualization and quantitative analysis.

Layer-wise Visualization. For each question, we permute the Top-5 documents to generate $5! = 120$ orders and extract the hidden states of the last token from each layer before response generation. Representative layers are then projected to two dimensions via PCA for visualization, as shown in Figure 3. We observe that hidden states in shallow layers form mixed clusters with points corresponding to different answers interleaved, while in deeper layers the clusters become increasingly well-separated and points with the same answer clearly group together. This indicates that variations in document order induce distinct reasoning trajectories, which manifest as progressively separable clusters in hidden state space, reflecting the model's internal reasoning patterns. More results are presented in Appendix C.3.

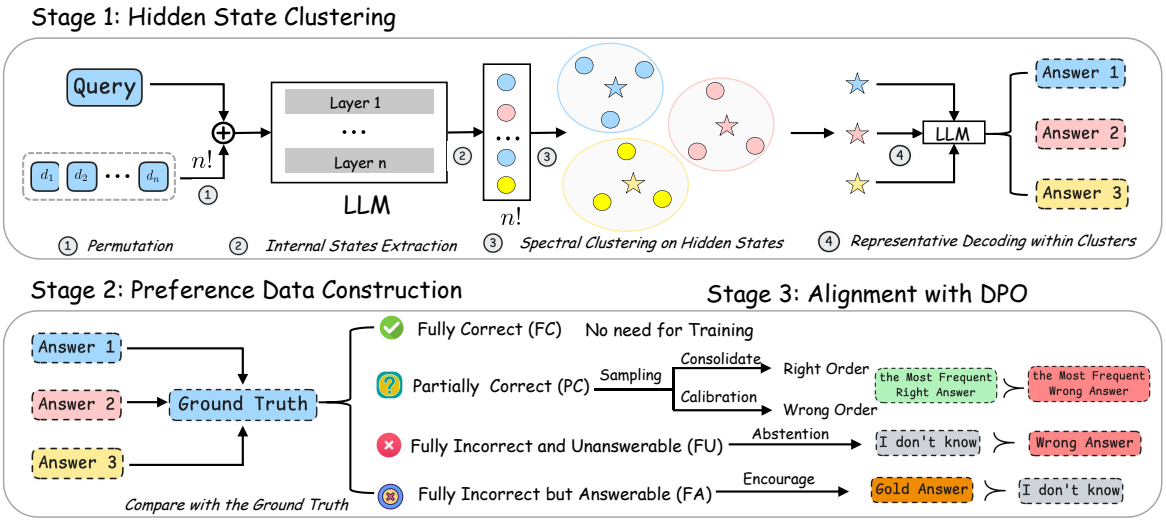


Figure 4: Overall framework of our Stable-RAG.

Model	Layer	Precision	Recall	F1
QWEN3-8B	8	78.1	79.3	77.9
	16	79.9	81.3	79.6
	24	86.8	87.5	86.6
	36	87.8	88.4	87.6
LLAMA3-8B-INSTRUCT	8	69.2	71.8	69.3
	16	81.4	82.5	81.3
	24	82.3	83.7	82.2
	32	84.1	85.2	83.9

Table 1: Clustering performance (%) of hidden states across different layers for Qwen3-8B and LLaMA3-8B-Instruct on the NQ train set using DPR retriever, averaged over 10,000 randomly sampled instances.

Quantitative Analysis of Clustering. To assess each cluster's reasoning performance, we select the hidden state closest to the cluster center, decode it as a representative answer of the cluster, and match this answer with the real reasoning answers of all hidden states in the same cluster to compute overall Precision, Recall, and F1 scores. As shown in Table 1, clustering metrics improve with network depth, indicating that hidden states for different answers become more separable in deeper layers. Notably, the clustering performance is already satisfactory for practical use, with F1 scores of 83.9 using LLaMA3 and 87.6 using Qwen3, respectively. Thus, we use the final layer hidden states for spectral clustering in our method.

4 Methodology

Overview. Our method comprises three stages: hidden state clustering, preference data construction and alignment with DPO, as shown in Figure 4. For each permutation, we extract the last token hidden state of the final layer before response generation, capturing the model's reasoning states. Spectral clustering is then applied to uncover latent reasoning modes, and representative states from each cluster are decoded. By aligning hidden states across permutations, our approach improves generation consistency across different retrieval orders.

4.1 Hidden State Clustering

Internal States Extraction. For each query q and its retrieved document set $\mathcal{S} = \{d_1, d_2, \ldots, d_n\}$, we enumerate all permutations of the documents and run the model for each permutation. Let $i \in \{1, \ldots, N\}$ denote the permutation index, where $N = n!$. To reduce computational cost, we extract only the last token hidden state of the final layer before response generation, $h^{(i)} \in \mathbb{R}^d$. Prior work (Azaria and Mitchell, 2023; Ni et al., 2025) has shown that this hidden state sufficiently captures the model's perception of its knowledge boundaries. We organize all hidden states into a matrix H:

$$H = [h^{(1)}, h^{(2)}, ..., h^{(N)}]^\top \in \mathbb{R}^{N \times d},$$

which represents the distribution of the model's final reasoning states across document permutations.

Spectral Clustering on Hidden States. To determine the number of clusters adaptively and capture the global structure of the hidden state space, we apply spectral clustering (Ng et al., 2001) to H, where each cluster corresponds to a latent reasoning mode (Lee et al., 2025). We compute the similarity between each pair of hidden states $h^{(i)}$ and

$h^{(j)}$ using the exponential of the cosine distance:

$$A_{ij} = \exp\Big(-\frac{1-\frac{h^{(i)}\cdot h^{(j)}}{\|h^{(i)}\|\,\|h^{(j)}\|}}{\sigma}\Big),$$

where σ is a hyperparameter controlling sensitivity. Here, $A \in \mathbb{R}^{N\times N}$ denotes the weighted adjacency matrix of all N hidden states.

The normalized graph Laplacian L is then constructed as:

$$D = \mathrm{diag}\Big(\sum_{j=1}^{N} A_{ij}\Big), \quad L = I - D^{-1/2}AD^{-1/2},$$

where D is the degree matrix, with each diagonal entry D_{ii} representing the sum of edge weights connected to the i-th hidden state (treated as a graph node), and I is the identity matrix.

The number of clusters K is determined adaptively via the eigengap of L. Let $\lambda_1 \le \cdots \le \lambda_N$ be the eigenvalues of L, and define the consecutive gaps $\mathrm{gap}_i = \lambda_{i+1} - \lambda_i$ between each pair of adjacent eigenvalues. The number of clusters is then set as $K = \max\big(2, (\arg\max_i \mathrm{gap}_i) + 1\big)$ to ensure clear separation between latent reasoning modes. Once K is determined, we obtain normalized spectral embeddings for all hidden states and assign each $h^{(i)}$ to one of the clusters $C_1, C_2, \ldots, C_K$. See more details in Appendix B.

Representative Decoding within Clusters. Within each cluster C_k, we identify a representative hidden state through centroid-based sampling. The cluster centroid is computed as:

$$\mu_k = \frac{1}{|C_k|}\sum_{h^{(i)}\in C_k} h^{(i)}.$$

We select the representative hidden state:

$$h^{(r_k)} = \arg\min_{h^{(i)}\in C_k} \|h^{(i)} - \mu_k\|_2.$$

Only the representative hidden states selected within each cluster $\{h^{(r_1)}, h^{(r_2)}, \ldots, h^{(r_K)}\}$ are decoded into textual answers, reducing the number of runs from $N = n!$ to K and substantially lowering computational and annotation overhead.

Exhaustive Full-Permutation Decoding. We study an exhaustive permutation decoding setting in which the model is evaluated under all ($N = n!$) permutations of retrieved documents. While this fully characterizes permutation-induced output variability, it is computationally and annotationally prohibitive at scale. We therefore use it only as a reference to assess the efficiency gains of our representative decoding strategy.

4.2 Preference Data Construction

Targets. Our goal is to build a robust RAG system. When the model cannot produce a reliable answer, it is encouraged to abstain to effectively suppress hallucinations and improve system reliability. When an answer is available, the output should remain consistent regardless of document order, thereby reducing permutation sensitivity and further enhancing overall reasoning robustness.

Data Construction Procedure. We construct preference data $\mathcal{P} = (x, y_w, y_l)$ for training. For each query q with its retrieved documents set $\mathcal{S} = \{d_1, d_2, \ldots, d_n\}$, the input x is formed by concatenating q with a specific document permutation π. Model outputs are obtained via representative decoding of hidden-state clusters induced by document permutations. Each instance is then compared with the ground truth and categorized into the following four types: FC (**Fully Correct**): the base model produces correct answers under all document permutations. Such instances are stable and excluded from training. PC (**Partially Correct**): the base model produces both correct and incorrect answers across permutations. Two representative outputs are sampled: y_w is the most frequent right answer to consolidate correct predictions, and y_l is the most frequent wrong answer for calibration. FU (**Fully Incorrect and Unanswerable**): the base model answers incorrectly under all permutations and no gold answers exist in the documents. y_w is set to *"I don't know"* to encourage abstention, and y_l is the most frequent wrong answer. FA (**Fully Incorrect but Answerable**): the base model answers incorrectly under all permutations but a gold answer exists in the documents. y_w is set to the gold answer to encourage correct prediction, and y_l is *"I don't know"*.

4.3 Alignment with DPO

We employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) to train the base model on the constructed preference tuples. For each tuple (x, y_w, y_l), DPO maximizes the likelihood of the preferred answer y_w over the less

Method	NQ				TriviaQA				HotpotQA				Average	
	Contriever		DPR		Contriever		DPR		Contriever		DPR			
	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
LLaMA3-8B-Instruct														
Direct Generation	25.18	29.11	25.18	29.11	55.92	58.95	55.92	58.95	21.39	22.87	21.39	22.87	34.16	36.98
Vanilla RAG	40.75	42.82	45.81	47.80	63.89	65.43	67.12	68.61	30.73	34.08	25.66	28.22	45.66	47.83
Vanilla SFT	42.10	44.78	46.20	49.44	55.52	51.40	57.10	52.51	27.25	31.58	24.63	29.85	42.13	43.26
RetRobust	41.82	44.26	48.70	49.29	64.85	66.72	68.67	70.42	31.46	35.34	26.96	30.36	47.08	49.40
ATM	43.75	44.88	49.78	50.19	66.37	67.12	70.12	70.35	34.36	36.97	28.55	29.31	48.82	49.80
RAAT	42.33	43.85	49.12	49.85	65.58	66.94	68.03	69.12	33.58	36.12	26.35	28.79	47.50	49.11
Pos2Distill	44.58	43.12	49.25	48.37	64.13	65.78	66.57	68.12	32.73	35.79	26.45	28.91	47.29	48.35
Ms-PoE	40.32	42.49	45.58	47.53	64.21	66.14	66.48	67.73	30.17	33.65	26.12	28.57	45.48	47.69
Stable-RAG (Ours)	**48.14**	**45.80**	**52.02**	**50.72**	**72.05**	**71.56**	**73.43**	**73.76**	**38.91**	**39.87**	**29.48**	**31.68**	**52.34**	**52.23**
Stable-RAG♣ (Ours)	**48.75**	**46.58**	**52.88**	**51.78**	**72.13**	**71.89**	**74.01**	**74.12**	**39.12**	**40.16**	**30.41**	**32.12**	**52.88**	**52.78**
Qwen3-8B														
Naive Generation	21.94	24.07	21.94	24.07	45.77	48.16	45.77	48.16	19.54	24.86	19.54	24.86	29.08	32.36
Vanilla RAG	44.65	45.34	50.55	50.67	64.35	66.29	69.62	71.03	33.14	38.66	26.17	31.33	48.08	50.55
Vanilla SFT	41.41	45.05	45.60	49.19	51.87	47.62	54.46	50.17	28.36	34.15	25.35	29.77	41.18	42.66
RetRobust	43.10	44.99	49.50	50.81	63.49	65.39	69.12	70.33	32.77	39.39	26.83	33.06	47.47	50.66
ATM	45.47	45.86	50.94	51.03	64.78	66.57	70.06	71.67	35.12	40.69	29.07	33.43	49.24	51.54
RAAT	45.13	45.87	50.12	50.03	63.12	65.17	68.54	69.88	33.54	39.06	27.21	**33.75**	47.94	50.63
Pos2Distill	44.89	45.52	50.71	50.93	64.95	66.81	69.87	71.35	33.72	39.11	26.53	31.88	48.45	50.93
Ms-PoE	44.39	45.12	50.04	50.08	64.88	66.72	69.03	70.84	32.98	38.21	25.93	31.02	47.88	50.33
Stable-RAG (Ours)	**46.12**	**46.79**	**51.69**	**51.78**	**66.58**	**68.13**	**71.32**	**72.89**	**35.73**	**41.78**	**30.15**	33.26	**50.27**	**52.44**
Stable-RAG♣ (Ours)	**46.94**	**47.13**	**52.12**	**52.38**	**67.11**	**68.79**	**71.74**	**73.40**	**36.89**	**42.94**	**31.77**	**35.78**	**51.10**	**53.40**

Table 2: Main results (%) on three QA benchmarks using two retrievers. ♣ denotes our method trained on exhaustive full-permutation decoding.

preferred y_l:

$$\mathcal{L}_{\text{DPO}} = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Big],$$

where θ denotes the model parameters, σ is the sigmoid function, and β is a scaling hyperparameter controlling the sharpness of preference. The model policy π_θ is initialized using the base reference policy π_{ref}.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our method on three QA benchmark datasets, including (1) Open-Domain QA, represented by NaturalQuestions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017); (2) Multi-Hop QA, represented by HotpotQA (Yang et al., 2018). Dataset statistics are provided in Appendix A.1.

Evaluation Metrics. Since answer style mismatch may cause additional variance, we follow prior and concurrent work (Zhu et al., 2024a; Wang et al., 2025c; Zhang et al., 2026c; Li et al., 2026d; Luo et al., 2026a) and adopt Substring Exact Match (**SubEM**) and **F1** for evaluation. SubEM checks whether the gold answer appears as a substring in the prediction, while F1 measures token-level overlap with the reference.

Baselines. We compare our method with the following baseline strategies on the same test set. Vanilla methods include *Direct Generation*, *Vanilla RAG* (Lewis et al., 2020), and *Vanilla SFT* (Zhang et al., 2024a). Robust RAG methods include *RetRobust* (Yoran et al., 2024), *ATM* (Zhu et al., 2024a), and *RAAT* (Fang et al., 2024). Positional Bias methods include *Pos2Distill* (Wang et al., 2025b) and *Ms-PoE* (Zhang et al., 2024d). The details of these baselines are presented in Appendix A.2.

Implementation Details. We use LLaMA3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025a) as backbone models for experiments. To ensure high and consistent evaluation quality (Cuconasu et al., 2024) and further assess the stability of our method under different retrieval settings, we follow prior work (Xu et al., 2024; Zhu et al., 2024b; Zhang et al., 2026c) and use the same Top-5 Wikipedia passages retrieved by DPR (Karpukhin et al., 2020) and Contriever-MS MARCO (Izacard et al., 2021) for all baselines and our method. Additional implementation details are provided in Appendix A.3.

5.2 Main Results

We conduct a comprehensive comparison of Stable-RAG against all the baseline methods, as shown in Table 2. The results indicate the following: (i) **Overall performance.** Stable-RAG consistently achieves the best overall performance across all the datasets with both Contriever and DPR retrievers,

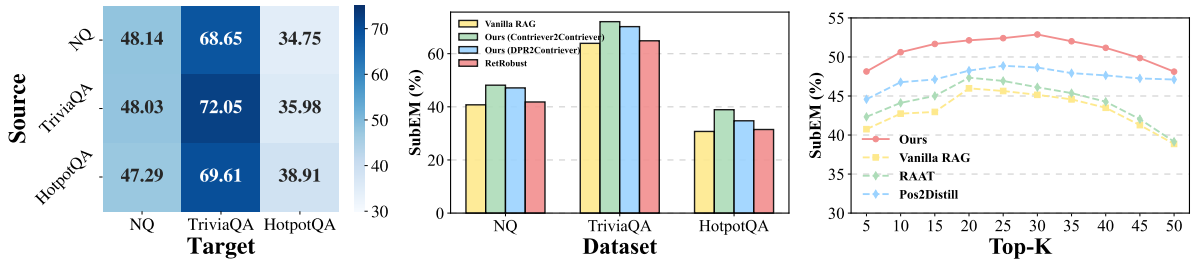


Figure 5: **(Left)** *Cross-Dataset Generalization*. We evaluate on three test sets with the Contriever retriever using SubEM. **(Middle)** *Cross-Retriever Transferability*. **(Right)** *Cross-Top-K Robustness*. We evaluate on the NQ test set with the Contriever retriever. All the three experiments are conducted on LLaMA3-8B-Instruct.

outperforming all strong baselines; (ii) **Effectiveness on complex reasoning.** Stable-RAG consistently improves performance on both single-hop and multi-hop QA tasks, demonstrating its ability to stabilize intermediate reasoning for complex questions; (iii) **Model generalization.** Stable-RAG performs robustly across backbone models, indicating model-agnostic generalization.

5.3 Further Analysis

Ablation Study. We conduct an ablation study to assess the contribution of each component in Stable-RAG, as shown in Table 3. Removing any component consistently degrades performance, demonstrating that all components are essential. In particular, excluding the *PC* component (Index a) causes significant drops across datasets, indicating the importance of partially correct signals for stabilizing reasoning. Removing *FA* (Index c) mainly impacts overall performance, while removing *FU* (Index b,d) sharply reduces the abstention rate (Wen et al., 2025; Sun et al., 2025), underscoring its role in handling unanswerable or hallucinated cases. Overall, Stable-RAG achieves the best trade-off between performance and abstention.

Comparison with Standard DPO. To isolate the effect of the order-stability mechanism, we compare Stable-RAG with standard DPO using the same base model and optimization strategy, differing only in whether reasoning consistency across document orders is enforced. In standard DPO, the model is trained to prefer the gold answer when evidence is available over other wrong answers obtained via sampling, or *"I don't know"* when the query is unanswerable. Results in Table 4 demonstrate that adding the order-stability constraint consistently improves RAG performance across datasets and retrievers without modifying the preference optimization framework.

Index	component			Dataset			Average	AR
	PC	*FA*	*FU*	NQ	TriviaQA	HotpotQA		
(a)	✗	✓	✓	37.62	61.37	28.54	42.51	**35.1**
(b)	✓	✗	✗	<u>47.17</u>	<u>71.28</u>	37.44	51.96	0.0
(c)	✓	✗	✓	46.73	70.14	35.75	50.87	17.3
(d)	✓	✓	✗	46.70	70.69	**38.93**	<u>52.11</u>	0.5
Ours	✓	✓	✓	**48.14**	**72.05**	<u>38.91</u>	**53.03**	<u>21.8</u>

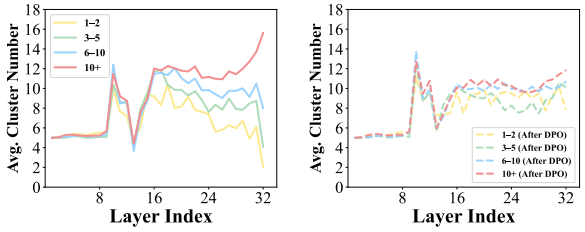
Table 3: Ablation results (%) on LLaMA3-8B-Instruct with the Contriever retriever measured by SubEM. **AR**(Abstention Rate) denotes the proportion of abstentions on 1,000 randomly sampled questions from three datasets when no retrieval evidence is available and the base model cannot answer. Higher AR indicates better awareness of model limitations and evidence availability. Best and second-best results are bolded and <u>underlined</u>, respectively.

Method	NQ		TriviaQA		HotpotQA	
	Contriever	DPR	Contriever	DPR	Contriever	DPR
Standard DPO	44.76	50.88	68.03	71.67	35.96	**30.43**
Ours	**48.14**	**52.02**	**72.04**	**73.43**	**38.91**	29.48

Table 4: SubEM results (%) between our method and Standard DPO using LLaMA3-8B-Instruct.

Cross-Dataset Generalization. We further evaluate the transferability of Stable-RAG across different data distributions. As shown in Figure 5 (Left), permutation-sensitivity patterns are learned on an in-domain dataset and directly applied to multiple out-of-distribution datasets to assess cross-dataset generalization. Experimental results demonstrate that Stable-RAG exhibits robust transfer across tasks and knowledge domains, consistently outperforming the best baseline regardless of the source–target dataset combination, and achieving stable improvements in answer consistency.

Cross-Retriever Transferability. We further evaluate the model's transferability by training on the DPR retriever and evaluating on the Contriever retriever. Figure 5 (Middle) shows that the model



(a) LLaMA3-8B-Instruct. (b) Ours.

(c) Ours (*w/o* **FU**, *w/o* **FC**). (d) Standard DPO.

Figure 6: Comparison of internal model behaviors across Base Model (**a**), Ours (**b**), one variant of Ours (**c**), and Standard DPO (**d**) on a random subset of 500 samples from the NQ test set with Contriever retriever.

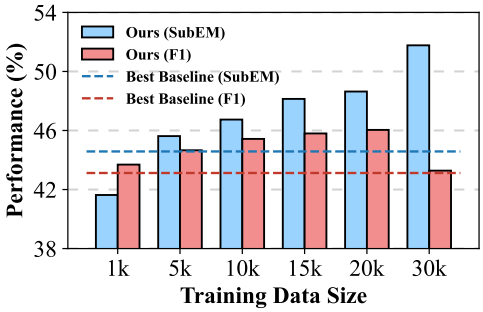


Figure 7: Effect of training data size on LLaMA3-8B-Instruct with Contriever retriever on the NQ dataset.

Method	Position of Gold Document				
	in Pos 1	in Pos 2	in Pos 3	in Pos 4	in Pos 5
Vanilla RAG	50.8	71.4	81.9	85.5	84.4
Vanilla SFT	47.2	66.2	74.8	80.0	82.6
RetRobust	35.5	75.5	85.3	88.6	88.9
ATM	33.7	64.2	71.8	77.4	77.8
Pos2Distill	29.5	55.8	69.4	72.8	73.2
Ms-PoE	31.4	63.8	72.1	73.9	74.3
Ours	**28.3**	**54.7**	**67.3**	**72.6**	**73.0**

Table 5: **PSR** (%) on the NQ test set with DPR retriever across different document positions, same as Figure 1.

maintains stable performance under cross-retriever settings, demonstrating strong transferability to different retrieval methods. Additionally, the results of training on the Contriever retriever and evaluating on the DPR retriever are shown in Appendix C.4.

Cross-Top-K Robustness. We train the model under a Top-5 setting and evaluate its performance on contexts retrieved with different Top-K values. Experimental results in Figure 5 (Right) show that the model maintains stable performance across various Top-K configurations and achieves significant improvements over corresponding baselines, demonstrating strong generalization when handling different numbers of candidate documents.

Effect of Training Data Size. As shown in Figure 7, we analyze the effect of training sample size on learning permutation sensitivity. Performance improves steadily with more data and saturates beyond 15k samples, indicating relatively small datasets suffice to capture core permutation-sensitivity patterns. However, with very limited data (e.g., 1k), performance drops markedly, reflecting difficulty in modeling fine-grained order differences. Given this trade-off, we adopt 15k samples as default, since gains over 20k do not justify the added computational cost.

Internal Model Behaviors after DPO. We label samples by their sensitivity according to the Base Model and exam hidden-state clustering after training. Figure 6b shows our method reduces clusters for high-sensitivity samples, keeps medium-sensitivity samples stable, and slightly increases low-sensitivity clusters. Figure 6c shows training on sensitive samples only, and Figure 6d shows standard DPO results. We can see that the increased clusters mainly stems from DPO-induced answer diversity rather than direct training on sensitive samples. For instance, for the same query *"when was the cat and mouse act introduced?"* and order, the response changes from *"1913."* to "*"introduced in April 1913."* after DPO. Overall, our method stabilizes high-sensitivity representations while preserving diversity for less sensitive samples.

External Positional Robustness after DPO. Following prior settings in Figure 1, we evaluate PSR on 1,000 randomly sampled instances by inserting the gold document in varying positions in the retrieved context to assess external positional robustness using LLaMA3-8B-Instruct. As shown in Table 5, our method consistently achieves lower PSR across all positions than the baselines, indicating reduced sensitivity to document ordering and improved external robustness under positional perturbations, even when the gold document appears in less favorable or later positions within the context.

Method	NQ			TriviaQA			HotpotQA		
	Original	Shuffled	Drop	Original	Shuffled	Drop	Original	Shuffled	Drop
Vanilla SFT	42.10	36.43	5.67	55.52	53.19	2.33	27.25	22.48	4.77
RetRobust	41.82	38.06	3.76	64.85	62.86	1.99	31.46	29.18	2.28
ATM	43.75	42.47	1.28	66.37	63.60	2.77	34.36	32.46	1.90
RAAT	42.33	40.54	1.79	65.58	62.19	3.39	33.58	29.75	3.83
Pos2Distill	44.58	43.63	0.95	64.13	63.57	0.56	32.73	32.09	**0.64**
Ms-PoE	40.32	39.17	1.15	64.21	62.96	1.25	30.17	29.14	1.03
Ours	**48.14**	**47.23**	**0.91**	**72.05**	**71.76**	**0.29**	**38.91**	**37.50**	1.41

Table 6: Performance comparison of LLaMA3-8B-Instruct with Contriever retriever under original and shuffled document order across three QA datasets. We report SubEM for evaluation.

Original vs. Shuffled Order As shown in Table 6, we present a comparison of answer performance under the original document order and a randomly shuffled order across three QA datasets. Our method achieves the highest SubEM scores in both original and shuffled conditions across all datasets, demonstrating its robustness to retrieval order permutations and its ability to maintain stable answer consistency.

6 Conclusion

We identify an underexplored vulnerability in RAG: LLMs are highly sensitive to document order, producing divergent reasoning and inconsistent or hallucinatory outputs from identical evidence. Layer-wise analysis traces this instability to the model's middle and higher layers. We propose Stable-RAG, which reduces permutation-induced uncertainty by clustering permuted hidden states and aligning reasoning modes via DPO optimization. Experiments across multiple QA benchmarks show consistent gains in accuracy, reasoning stability, and strong transferability. Enforcing layer-wise reasoning constraints while reducing training costs offers a promising approach to mitigate permutation-induced hallucinations.

Limitations

While this work demonstrates the effectiveness of Stable-RAG in mitigating permutation-induced hallucinations, it has several limitations that warrant further investigation.

First, our approach focuses on stabilizing reasoning at the final-layer representation level, without explicitly enforcing layer-wise reasoning path constraints throughout the model. Although our analysis reveals that permutation-induced divergence primarily emerges in the middle and higher layers, Stable-RAG does not directly regularize intermediate-layer reasoning trajectories. Incorporating explicit layer-wise constraints or trajectory-level alignment may further improve reasoning stability, but would require more fine-grained supervision or architectural modifications, which we leave for future work.

Second, Stable-RAG relies on spectral clustering over document-permuted hidden representations to estimate dominant reasoning modes and construct preference signals for DPO alignment. While this strategy reduces annotation cost by approximately threefold compared to exhaustive full-permutation decoding, it still incurs non-trivial computational and labeling overhead. More efficient clustering strategies, weak supervision signals, or fully unsupervised alignment objectives could further reduce annotation requirements and improve scalability. Exploring such cost-effective supervision mechanisms is important for building more robust and practical RAG systems.

Ethical Considerations

While Stable-RAG aims to enhance RAG robustness and accuracy, ethical considerations remain. First, although our method reduces hallucinations from document order, it cannot guarantee fully correct outputs. Users should avoid over-reliance in high-stakes domains (e.g., healthcare, law, and finance). Second, Stable-RAG relies on external documents for grounding, which may contain biases or errors, potentially propagating or amplifying them.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. U25B2070 and 62406013, the Beijing Advanced Innovation Center Funds for Future Blockchain and Privacy Computing (GJJ-24-034), and the Fundamental Research Funds for the Central Universities.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Gere: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2184–2189.

Jinwen Chen, Hainan Zhang, Liang Pang, Yongxin Tong, Haibo Zhou, Yuan Zhan, Wei Lin, and Zhiming Zheng. 2025. Privacy-preserving reasoning with knowledge-distilled parametric retrieval augmented generation. *arXiv preprint arXiv:2509.01088*.

Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. 2024. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11160–11174.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Beni Egressy and Jan Stühmer. 2025. Set-llm: A permutation-invariant llm. *arXiv preprint arXiv:2505.15433*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039.

Rong Fu, Yang Li, Zeyu Zhang, Jiekai Wu, Yaohua Liu, Shuaishuai Cao, Yangchen Zeng, Yuhang Zhang, Xiaojing Du, Chuang Zhao, Kangning Cui, and Simon Fong. 2026a. Neurosymactive: Differentiable neural-symbolic reasoning with active exploration for knowledge graph question answering. *arXiv preprint arXiv:2602.15353*.

Rong Fu, Yemin Wang, Tianxiang Xu, Yongtai Liu, Weizhi Tang, Wangyu Wu, Xiaowen Ma, and Simon Fong. 2026b. S-path-rag: Semantic-aware shortest-path retrieval augmented generation for multi-hop knowledge graph question answering. *arXiv preprint arXiv:2603.23512*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Xueren Ge, Sahil Murtaza, Anthony Cortez, and Homa Alemzadeh. 2026. Expert-guided prompting and retrieval-augmented generation for emergency medical service question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30798–30806.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*.

Faisal Hamman, Chenyang Zhu, Anoop Kumar, Xujun Peng, Sanghamitra Dutta, Daben Liu, and Alfy Samuel. 2025. Improving consistency in retrieval-augmented systems with group similarity reward. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junan Hu, Shudan Guo, Wenqi Liu, Jianhua Yin, and Yinwei Wei. 2026. Context-agent: Dynamic discourse trees for non-linear dialogue. *arXiv preprint arXiv:2604.05552*.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 2523–2540.

Shiting Huang, Zhen Fang, Zehui Chen, Siyu Yuan, Junjie Ye, Yu Zeng, Lin Chen, Qi Mao, and Feng Zhao. 2025. CRITICTOOL: Evaluating self-critique capabilities of large language models in tool-calling error scenarios. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26672–26704.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Hongru Ji, Yuyin Fan, Meng Zhao, Xianghua Li, Lianwei Wu, and Chao Gao. 2026. Stride-ed: A strategy-grounded stepwise reasoning framework for empathetic dialogue systems. *Preprint*, arXiv:2604.07100.

Angqing Jiang, Jianlyu Chen, Zhe Fang, Yongcan Wang, Xinpeng Li, Keyu Ding, and Defu Lian. 2026a. Cmedteb & care: Benchmarking and enabling efficient chinese medical retrieval via asymmetric encoders. *arXiv preprint arXiv:2604.10937*.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. 2026b. Magma: A multi-graph based agentic memory architecture for ai agents. *arXiv preprint arXiv:2601.03236*.

Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, Alysa Zhao, Dingyi Kang, Xu Hu, Feng Chen, Qiannan Li, and Bingzhe Li. 2026c. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations. *arXiv preprint arXiv:2602.19320*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Linggang Kong, Lei Wu, Yunlong Zhang, Xiaofeng Zhong, Zhen Wang, Yongjie Wang, and Yao Pan. 2026. Causalgaze: Unveiling hallucinations via counterfactual graph intervention in large language models. *Preprint*, arXiv:2604.11087.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Sungjae Lee, Hoyoung Kim, Jeongyeon Hwang, Eunhyeok Park, and Jungseul Ok. 2025. Efficient latent semantic clustering for scaling test-time computation of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24126–24144.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2026a. Modeling uncertainty trends for timely retrieval in dynamic rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31527–31535.

Bo Li, Mingda Wang, Gexiang Fang, Shikun Zhang, and Wei Ye. 2026b. Retrieval as generation: A unified framework with self-triggered information planning. *Preprint*, arXiv:2604.11407.

Fanxiao Li, Jiaying Wu, Tingchao Fu, Dayang Li, Herun Wan, Wei Zhou, and Min-Yen Kan. 2026c. What's left unsaid? detecting and correcting misleading omissions in multimodal news previews. *arXiv preprint arXiv:2601.05563*.

Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. 2025a. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9342–9354.

Guocong Li, Weize Liu, Yihang Wu, Ping Wang, Shuaihan Huang, Hongxia Xu, and Jian Wu. 2025b. From misleading queries to accurate answers: A three-stage fine-tuning method for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1192–1209.

Guocong Li, Jinjian Zhang, Ping Wang, Dongnan Liu, Tian Liang, Qiuyi Qi, Hao Huang, Siyan Guo, Mutian Bao, Wei Zhou, Linjian Mo, Hongxia Xu, and Jian Wu. 2026d. Mol: Adaptive mixture-of-length reasoning for efficient question answering with context. In *The Fourteenth International Conference on Learning Representations*.

Jiaqian Li, Yanshu Li, Ligong Han, Ruixiang Tang, and Wenya Wang. 2025c. Towards generalizable implicit in-context learning with attention routing. *Preprint*, arXiv:2509.22854.

Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Ligong Han, Hongyang He, Zhengtao Yao, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and Ruixiang Tang. 2025d. Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning. *Preprint*, arXiv:2505.17097.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. 2025. Clue: Non-parametric verification from experience via hidden-state clustering. *Preprint*, arXiv:2510.01591.

Ailiang Lin, Zhuoyun Li, Kotaro Funakoshi, and Manabu Okumura. 2025. Causal2vec: Improving decoder-only llms as versatile embedding models. *Preprint*, arXiv:2507.23386.

Hongzhan Lin, Ang Lv, Yuhan Chen, Chen Zhu, Yang Song, Hengshu Zhu, and Rui Yan. 2024. Mixture of in-context experts enhance llms' long context awareness. *Advances in Neural Information Processing Systems*, 37:79573–79596.

Peiyang Liu, Xi Wang, Ziqiang Cui, and Wei Ye. 2025. Queries are not alone: Clustering text embeddings for video search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 874–883.

Peiyang Liu, Xi Wang, Lin Wang, Wei Ye, Xiangyu Xi, and Shikun Zhang. 2021a. Distilling knowledge from bert into simple fully connected neural networks for efficient vertical retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3965–3975.

Peiyang Liu, Xi Wang, Sen Wang, Wei Ye, Xiangyu Xi, and Shikun Zhang. 2021b. Improving embedding-based large-scale retrieval via label enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 133–142.

Guanran Luo, Zhongquan Jian, Wentao Qiu, Meihong Wang, and Qingqiang Wu. 2025. DTCRS: Dynamic tree construction for recursive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10948–10963.

Guanran Luo, Wentao Qiu, Zhongquan Jian, Meihong Wang, and Qingqiang Wu. 2026a. Gcot-decoding: Unlocking deep reasoning paths for universal question answering. *Preprint*, arXiv:2604.06794.

Guanran Luo, Wentao Qiu, Wanru Zhao, Wenhan Lv, Zhongquan Jian, Meihong Wang, and Qingqiang Wu. 2026b. Agsc: Adaptive granularity and semantic clustering for uncertainty quantification in long-text generation. *Preprint*, arXiv:2604.06812.

Kexin Ma, Ruochun Jin, Wang Haotian, Wang Xi, Huan Chen, Yuhua Tang, and Qian Wang. 2024. Context-driven index trimming: A data quality perspective to enhancing precision of RALMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4886–4901.

Kexin Ma, Bojun Li, Yuhua Tang, Liting Sun, and Ruochun Jin. 2026. Cast: Character-and-scene episodic memory for agents. *Preprint*, arXiv:2602.06051.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24315–24329.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Cheng Qian, Hainan Zhang, Yongxin Tong, Hong-Wei Zheng, and Zhiming Zheng. 2025. Hyfedrag: A federated retrieval-augmented generation framework for heterogeneous and privacy-sensitive data. *arXiv preprint arXiv:2509.06444*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. 2025. CausalAbstain: Enhancing multilingual LLMs with causal reasoning for trustworthy abstention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14060–14076.

Yuxi Sun, Aoqi Zuo, Haotian Xie, Wei Gao, Mingming Gong, and Jing Ma. 2026. Fact-e: Causality-inspired evaluation for trustworthy chain-of-thought reasoning. *Preprint*, arXiv:2604.10693.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2025a. Think-while-generating: On-the-fly reasoning for personalized long-form generation. *arXiv preprint arXiv:2512.06690*.

Chengbing Wang, Wuqiang Zheng, Yang Zhang, Fengbin Zhu, Junyi Cheng, Yi Xie, Wenjie Wang, and Fuli Feng. 2026a. Perm: Psychology-grounded empathetic reward modeling for large language models. *arXiv preprint arXiv:2601.10532*.

Tongxi Wang. 2026. Fbs: Modeling native parallel reading inside a transformer. *arXiv preprint arXiv:2601.21708*.

Yifei Wang, Feng Xiong, Yong Wang, Linjing Li, Xiangxiang Chu, and Daniel Dajun Zeng. 2025b. POSITION BIAS MITIGATES POSITION BIAS: Mitigate position bias through inter-position knowledge distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1495–1512.

Yu Wang, Emmanuele Chersoni, and Chu-Ren Huang. 2026b. This one or that one? a study on accessibility via demonstratives with multimodal large language models. In *Language Resources and Evaluation Conference 2026*.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025c. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25434–25442.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Xi Xiao, Chenrui Ma, Yunbei Zhang, Chen Liu, Zhuxuanzi Wang, Yanshu Li, Lin Zhao, Guosheng Hu, Tianyang Wang, and Hao Xu. 2026. Not all directions matter: Toward structured and task-aware low-rank adaptation. *arXiv preprint arXiv:2603.14228*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Shijia Xu, Zhou Wu, Xiaolong Jia, Yu Wang, Kai Liu, and April Xiaowen Dong. 2026. Self-correcting rag: Enhancing faithfulness via mmkp context selection and nli-guided mcts. *Preprint*, arXiv:2604.10734.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Junqi Yang, Yuecong Min, Jie Zhang, Shiguang Shan, and Xilin Chen. 2026. Infact: A diagnostic benchmark for induced faithfulness and factuality hallucinations in video-llms. *arXiv preprint arXiv:2603.11481*.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025b. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Haohan Yuan, Sukhwa Hong, and Haopeng Zhang. 2026. StrucSum: Graph-structured reasoning for long document extractive summarization with LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3708–3721.

Haohan Yuan and Haopeng Zhang. 2025a. DomainSum: A hierarchical benchmark for fine-grained domain shift in abstractive text summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2219–2231.

Haohan Yuan and Haopeng Zhang. 2025b. Understanding llm reasoning for abstractive summarization. *arXiv preprint arXiv:2512.03503*.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. 2026. Vision-deepresearch benchmark: Rethinking visual and textual search for

multimodal large language models. *arXiv preprint arXiv:2602.02185*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139.

Jiajun Zhang, Zeyu Cui, Jiaxi Yang, Lei Zhang, Yuheng Jing, Zeyao Ma, Tianyi Bai, Zilei Wang, Qiang Liu, Liang Wang, Binyuan Hui, and Junyang Lin. 2026a. From completion to editing: Unlocking context-aware code infilling via search-and-replace instruction tuning. *arXiv preprint arXiv:2601.13384*.

Jiajun Zhang, Yuying Li, Zhixun Li, Xingyu Guo, Jingzhuo Wu, Leqi Zheng, Yiran Yang, Jianke Zhang, Qingbin Li, Shannan Yan, Zhetong Li, Changguo Jia, Junfei Wu, Zilei Wang, Qiang Liu, and Liang Wang. 2026b. Realchart2code: Advancing chart-to-code generation with real data and multi-task evaluation. *arXiv preprint arXiv:2603.25804*.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. 2025a. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. 2026c. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pages 1971–1982.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2024b. Adacomp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv preprint arXiv:2409.01579*.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025b. SOTOPIA-Ω: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24669–24697.

Wenyuan Zhang, Shuaiyi Nie, Xinghua Zhang, Zefeng Zhang, and Tingwen Liu. 2025c. S1-bench: A simple benchmark for evaluating system 1 thinking capability of large reasoning models. *arXiv preprint arXiv:2504.10368*.

Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. 2026d. Expseek: Self-triggered experience seeking for web agents. *Preprint*, arXiv:2601.08605.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Jianing Wang, Zhuangzhuang Ye, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2024c. Trendfact: A benchmark for explainable hotspot perception in fact-checking with natural language explanation. *arXiv preprint arXiv:2410.15135*.

Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid, and Mourad Oussalah. 2026e. Towards reliable multimodal disaster severity assessment through preference optimization and explainable vision-language reasoning. *Reliability Engineering & System Safety*, page 112674.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024d. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37:60755–60775.

Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu, and Di Wang. 2025a. Flattery in motion: Benchmarking and analyzing sycophancy in video-llms. *arXiv preprint arXiv:2506.07180*.

Wenrui Zhou, Qiyu Liu, Jingshu Peng, Aoqian Zhang, and Lei Chen. 2025b. Carpo: Leveraging listwise learning-to-rank for context-aware query plan optimization. *arXiv preprint arXiv:2509.03102*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. 2024a. ATM: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10902–10919.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024b. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1069.

Appendix

Contents

A Implementation Details

A.1 Datasets

We conduct experiments on three widely used QA datasets that cover both single-hop and multi-hop question-answering scenarios. Table 7 summarizes the key statistics of these datasets. Specifically, **NQ** (Kwiatkowski et al., 2019) and **TriviaQA** (Joshi et al., 2017) are representative single-hop datasets, where each question can typically be answered using information from a single passage retrieved from the corpus. These datasets primarily evaluate a model's ability to locate and extract factual evidence efficiently. In contrast, **HotpotQA** (Yang et al., 2018) is a challenging multi-hop dataset that requires integrating and reasoning over multiple pieces of evidence distributed across different documents to derive the final answer. This dataset is particularly useful for testing a model's reasoning and compositional understanding capabilities. Together, these datasets provide a comprehensive benchmark for evaluating both the retrieval quality and reasoning robustness of our proposed method under diverse task settings.

Dataset	Type	# Train	# Dev	# Test
NQ	single-hop	79.1k	8.7k	3.6k
TriviaQA	single-hop	78.7k	11.3k	8.8k
HotpotQA	multi-hop	88.9k	5.6k	5.6k

Table 7: Statistics for the datasets.

A.2 Baseline Details

We compare Stable-RAG with the following baseline strategies. To ensure a fair comparison, all methods are evaluated on the same test set and retrieved set.

Vanilla Methods. (i) *Direct Generation*. This baseline relies solely on the generator's parametric knowledge to produce answers without consulting any retrieved documents. (ii) *Vanilla RAG* (Lewis et al., 2020). This baseline concatenates all retrieved documents as model input without any additional processing. (iii) *Vanilla SFT*. We implement vanilla SFT following Zhang et al. (2024a). For each training example, this baseline uses the gold answer as the training label if it appears in the retrieved documents; otherwise, it assigns *"I don't know"* as the training label to guide the model to abstain when the necessary information is missing.

Robust RAG. (i) *RetRobust* (Yoran et al., 2024). This baseline improves retrieval-augmented QA models by filtering out irrelevant retrieved passages and fine-tuning the model on a mix of relevant and irrelevant contexts, enabling it to leverage relevant information while remaining robust to irrelevant content. (ii) *ATM* (Zhu et al., 2024a). This baseline optimizes a retrieval-augmented Generator using an Adversarial Tuning Multi-agent system, where an auxiliary Attacker agent iteratively steers the Generator to better discriminate useful documents from noisy or fabricated ones, improving robustness and performance on knowledge-intensive question answering tasks. (iii) *RAAT* (Fang et al., 2024). This baseline dynamically adjusts the model's learning process in response to various types of retrieval noise through adaptive adversarial training, while employing multi-task learning to enable the model to internally recognize and handle noisy contexts, thereby improving robustness and answer quality in retrieval-augmented generation.

Positional Bias. (i) *Pos2Distill* (Wang et al., 2025b). This baseline mitigates positional bias in long-context tasks by transferring knowledge from advantageous positions to less favorable ones through position-to-position knowledge distillation. (ii) *Ms-PoE* (Zhang et al., 2024d). This baseline uses Multi-scale Positional Encoding to mitigate the "lost-in-the-middle" issue in LLMs by rescaling positional indices and assigning different scaling ratios to attention heads, enabling multi-scale context fusion without fine-tuning or extra overhead.

A.3 Training Details

We use LLaMA3-8B-Instruct [2] (Grattafiori et al., 2024) and Qwen3-8B [3] (Yang et al., 2025a) as backbone models following prior and concurrent work (Zhang et al., 2025c,b; Lin et al., 2025; Huang et al., 2025; Yuan and Zhang, 2025a; Zhang et al., 2026d; Li et al., 2025d,c). These models and their variants have been widely used across a variety of tasks and applications (Zhang et al., 2025a; Yang et al., 2025b; Li et al., 2025a, 2026c; Zhang et al., 2026a; Wang et al., 2026b; Zhang et al., 2026b; Wang et al., 2025a, 2026a; Jiang et al., 2026c,b; Fu et al., 2026b,a; Sun et al., 2026; Jiang et al., 2026a; Zeng et al., 2026; Ma et al., 2026). We implement DPO training pipeline using the HuggingFace Transformers (Wolf et al., 2020), incorporating PEFT LoRA (Hu et al., 2022; Xiao et al., 2026) for parameter-efficient fine-tuning. Both the base model and reference model are initialized from pre-trained checkpoints, with the reference model kept in evaluation mode to provide stable policy targets during training. Each dataset is randomly shuffled and split into 85% training and 15% validation samples, with a maximum of 18,000 samples per dataset to control computational overhead. We fix the random seed to 42 to ensure reproducibility. LoRA is applied to all projection layers with rank $r = 128$, alpha $= 128$, dropout $= 0$ and no additional bias terms. The DPO configurations (Zhang et al., 2026e) use a per-device batch size of 2 with gradient accumulation of 8, a learning rate of 5×10^{-6}, a linear warmup ratio of 0.1, and a preference scaling hyperparameter β of 0.4. We train LLaMA-3-8B-Instruct for 1 epoch and Qwen3-8B for 2 epochs on two NVIDIA RTX PRO 6000 GPUs, with each epoch taking roughly two hours. Notably, we use greedy decoding during data construction, and set the temperature to 0.01 during inference, which is nearly equivalent to greedy decoding. This ensures that output variations primarily reflect document-order sensitivity rather than sampling randomness.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[3] https://huggingface.co/Qwen/Qwen3-8B

```
<system>
You are a helpful, respectful, and honest assistant. Answer the question with couple of words using the provided documents. For example: Question: What is the capital of France? Output: Paris.
</system>
<user>
Question: {query}
Documents:
Doc1: {Document 1}
Doc2: {Document 2}
.....
</user>
```

Table 8: Prompt for the backbone LLMs.

A.4 Prompts

We adopt a system-user style prompting scheme (White et al., 2023; Zhou et al., 2025a,b) to guide the backbone LLMs to generate concise, document-grounded answers, as presented in Table 8.

B Mathematical Derivations

We employ spectral clustering on hidden states to identify dominant reasoning modes across permutations of retrieved documents. Compared with conventional clustering methods, spectral clustering captures the global structure of the hidden state space. This enables Stable-RAG to robustly group similar reasoning behaviors, reduce noise from spurious variations, and improve the consistency of preference signals used for DPO alignment.

B.1 Spectral Clustering on Hidden States

Spectral clustering is applied to the hidden states matrix:

$$H = [h^{(1)}, h^{(2)}, \ldots, h^{(N)}]^\top \in \mathbb{R}^{N \times d}$$

to adaptively determine the number of clusters and capture the global structure of the hidden state

space, where each cluster corresponds to a latent reasoning mode (Lee et al., 2025).

B.2 Similarity Graph and Adjacency Matrix

We construct a weighted similarity graph $G = (V, E)$ where each node corresponds to a hidden state $h^{(i)}$ and edges encode pairwise similarities. The adjacency matrix $A \in \mathbb{R}^{N \times N}$ is computed as the exponential of the cosine distance:

$$A_{ij} = \exp\Big(-\frac{1 - \frac{h^{(i)} \cdot h^{(j)}}{\|h^{(i)}\| \, \|h^{(j)}\|}}{\sigma}\Big),$$

where σ is a hyperparameter controlling sensitivity.

B.3 Degree Matrix and Normalized Laplacian

The degree matrix D is a diagonal matrix with entries:

$$D_{ii} = \sum_{j=1}^{N} A_{ij}.$$

The normalized graph Laplacian is:

$$L = I - D^{-1/2} A D^{-1/2},$$

where I is the identity matrix.

B.4 Eigen-decomposition and Determining Cluster Number

Let $\lambda_1 \leq \cdots \leq \lambda_N$ be the eigenvalues of L. Define the consecutive eigengaps as:

$$\text{gap}_i = \lambda_{i+1} - \lambda_i.$$

The number of clusters K is set adaptively as:

$$K = \max\big(2, (\arg\max_i \text{gap}_i) + 1\big),$$

ensuring clear separation between latent reasoning modes following standard practice (Ng et al., 2001; Von Luxburg, 2007).

B.5 Spectral Embedding and Clustering

We then compute the first K eigenvectors of L, normalize each row to unit length, and apply standard clustering to assign each hidden state $h^{(i)}$ to one of the clusters:

$$C_1, C_2, \ldots, C_K,$$

exactly following the procedure described in the main text.

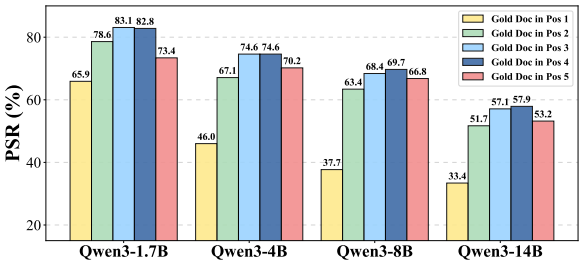


Figure 8: **Perturbation Success Rate (PSR)** on the NQ test set across different Qwen3 models. PSR is computed as the proportion of successful document-order perturbations to produce hallucination results among 1,000 randomly sampled instances, with the gold document fixed in different positions.

C More Experimental Results

C.1 Permutation Sensitivity in Qwen3 Models

We further investigate whether document-order sensitivity generalizes to different model families by reporting PSR results on the Qwen3 series. Following the same evaluation protocol as in Figure 1, we fix the gold document in different positions and measure the proportion of document-order perturbations that lead to hallucinated outputs over 1,000 randomly sampled instances on the NQ test set.

Figure 8 compares the PSR trends of the Qwen3 models with those observed in the LLaMA3 Instruct series. Overall, Qwen3 models exhibit clear document-order sensitivity across all model sizes. When the gold document is placed at early positions, the PSR is relatively low, indicating stronger robustness to document-order perturbations. However, as the gold document is shifted to later positions, PSR increases substantially, suggesting a higher likelihood of hallucinations induced purely by document reordering.

We observe a consistent monotonic pattern across Qwen3 variants: PSR generally rises from Top-1 to Top-3 or Top-4 and slightly saturates or declines afterward. This behavior closely mirrors the trends observed in LLaMA3 models, despite differences in model architecture and pretraining data. Moreover, smaller Qwen3 models tend to exhibit higher sensitivity to document order changes, while larger models demonstrate comparatively improved robustness, though the issue remains non-negligible even at larger scales.

These results show that document-order sensitivity is a general property of modern LLMs, highlighting the need for order-robust RAG methods.

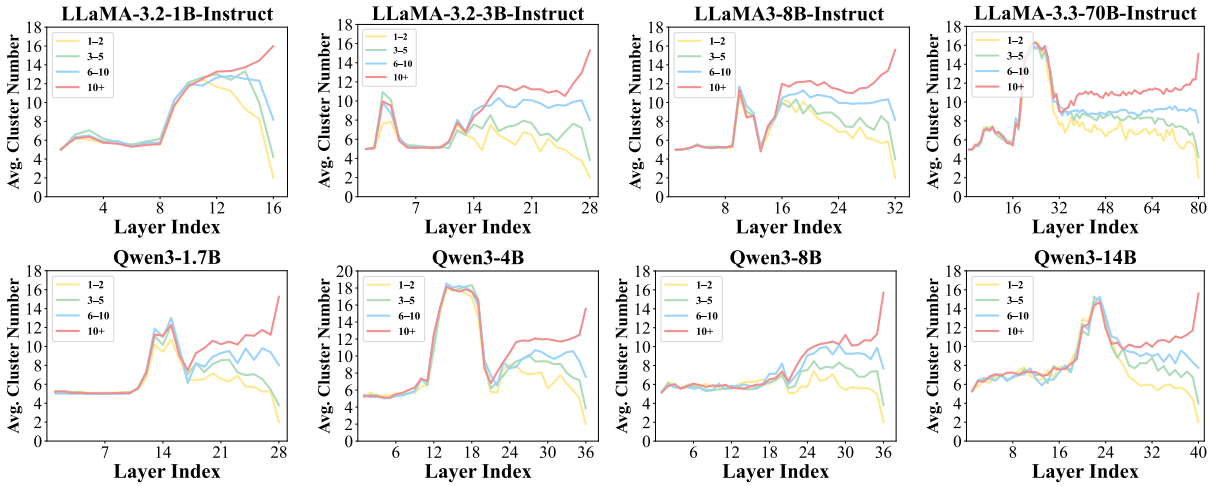


Figure 9: Hidden-state clustering behaviors across layers for LLaMA3 series on the NQ train set with DPR retriever and Qwen3 series on the HotpotQA train set with Contriever retriever, using 1,000 randomly sampled instances. Different colored lines indicate the number of clusters of final reasoning states produced by the LLM under all $5!(=120)$ permutations of the Top-5 retrieved documents (e.g., the green line indicates 3–5 cluster states).

C.2 Structural Instability Across Model Families

We provide additional visualizations of the structural instability in internal reasoning dynamics for both the LLaMA3 and Qwen3 model families as shown in Figure 9. We analyze how document permutations induce representation divergence across layers. Despite differences in architecture, scale, and pretraining data, both model families show consistent structural instability. Specifically, shallow-layer representations remain relatively concentrated under document permutations, while strong divergence emerges in middle layers and becomes more pronounced in higher layers. Moreover, high-sensitivity samples consistently exhibit greater representational divergence than stable ones.

These observations suggest that permutation sensitivity originates from a shared structural instability in the reasoning dynamics of large language models rather than from model-specific design choices (Wang, 2026). Consistent trends across LLaMA3 and Qwen3 further highlight the need to address structural instability to improve RAG robustness.

C.3 Visualization of Layer-wise Hidden States

Figure 11 shows LLaMA3-8B-Instruct on NQ using the Contriever retriever, illustrating the hidden state evolution across all layers for a selected example. Figure 12 displays Qwen3-8B on HotpotQA dataset using Contriever dataset, showing the layer-wise progression of hidden states for a representative sample. In both cases, shallow layers exhibit mixed clusters with points corresponding to different answers interleaved, while deeper layers form increasingly well-separated clusters according to the final answers. These visualizations reinforce that the structural evolution of reasoning trajectories is consistent across multiple models and datasets.

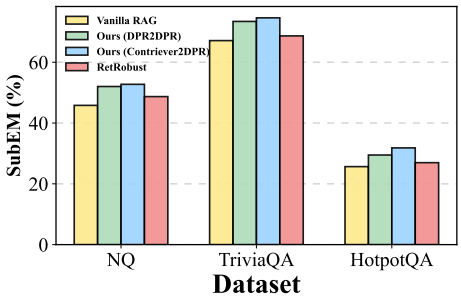


Figure 10: Cross-Retriever Transferability.

C.4 Cross-Retriever Transferability

As shown in Section 5.3, we evaluate transferability from DPR to Contriever. We additionally test transferability from Contriever to DPR in Figure 10. Both experiments confirm that Stable-RAG consistently improves answer consistency and reduces permutation-induced variance across retrievers, demonstrating cross-retriever transferability.

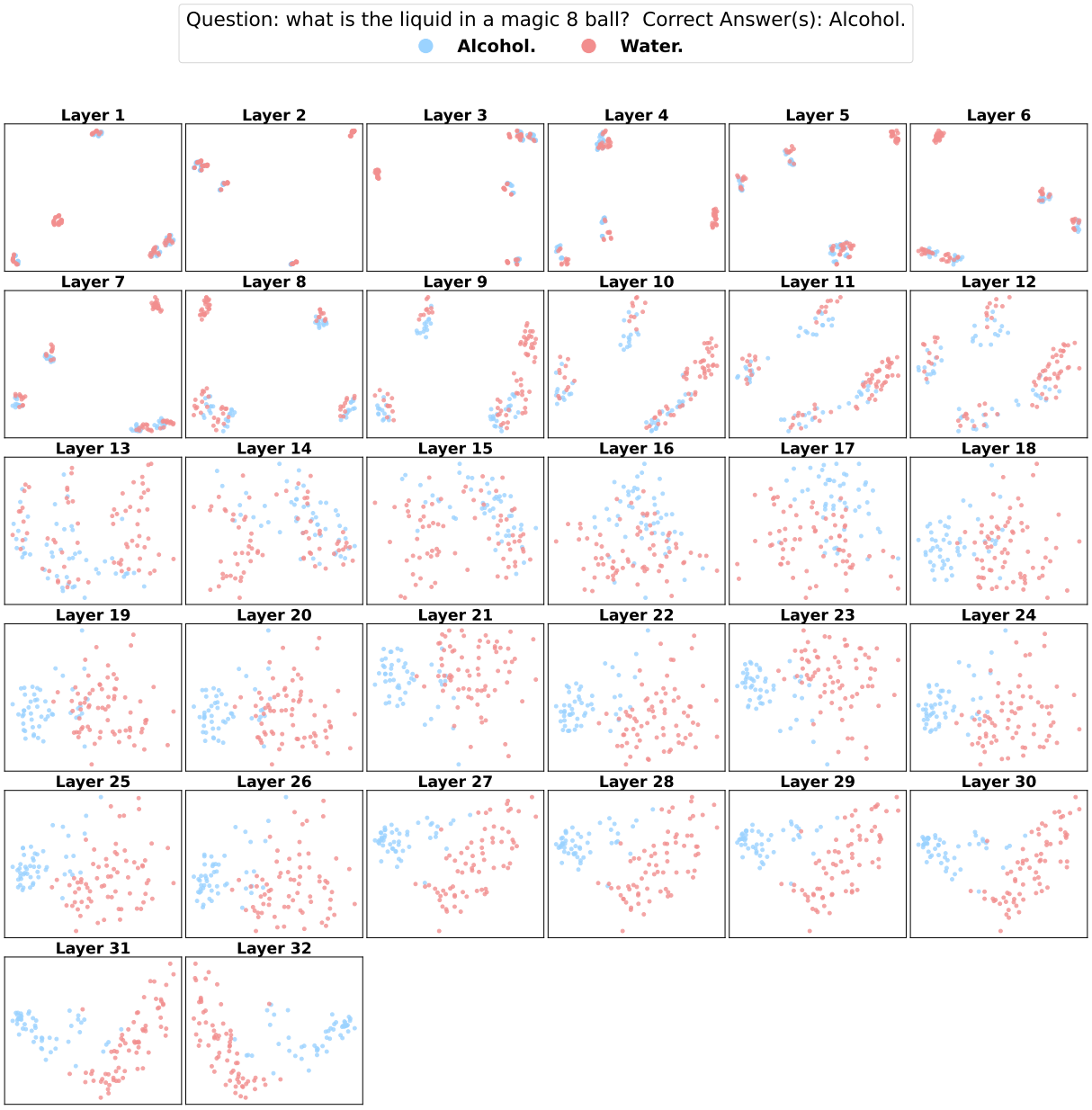


Figure 11: 2D PCA visualization of hidden state representations across all layers in LLaMA3-8B-Instruct for a single example. Each point corresponds to a document order, and its color represents the model's final answer.

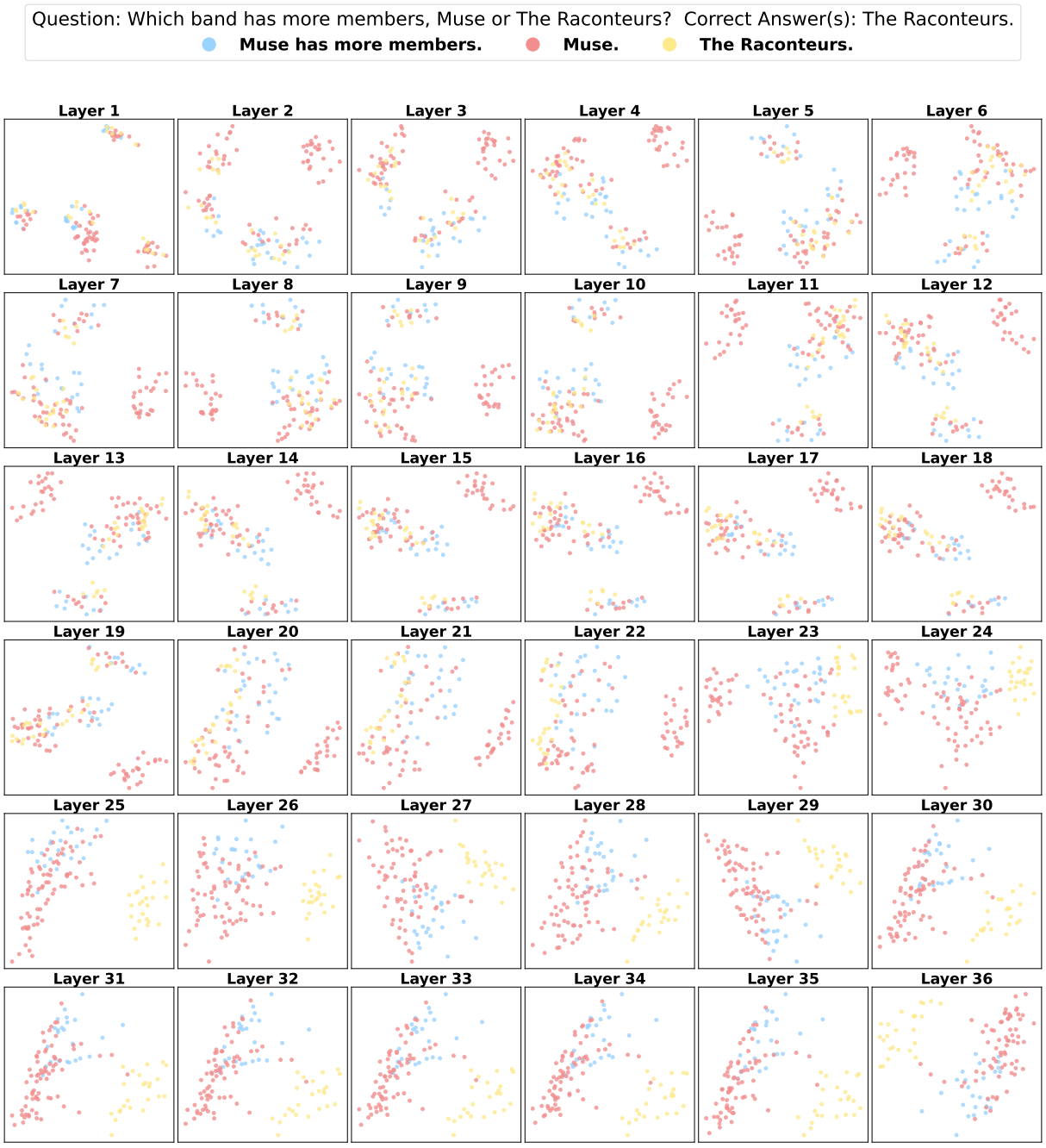


Figure 12: 2D PCA visualization of hidden state representations across all layers in Qwen3-8B for a single example. Each point corresponds to a document order, and its color represents the model's final answer.