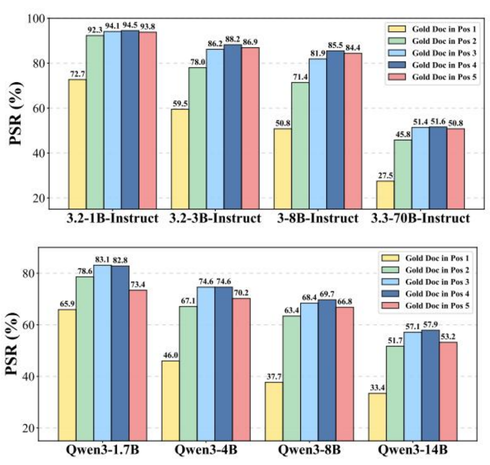
Stable-RAG: Mitigating Retrieval-Permutation-Induced Hallucinations in Retrieval-Augmented Generation

Qianchi Zhang[1], Hainan Zhang[1]*, Liang Pang[3], Hongwei Zheng[2], Zhiming Zheng[1]

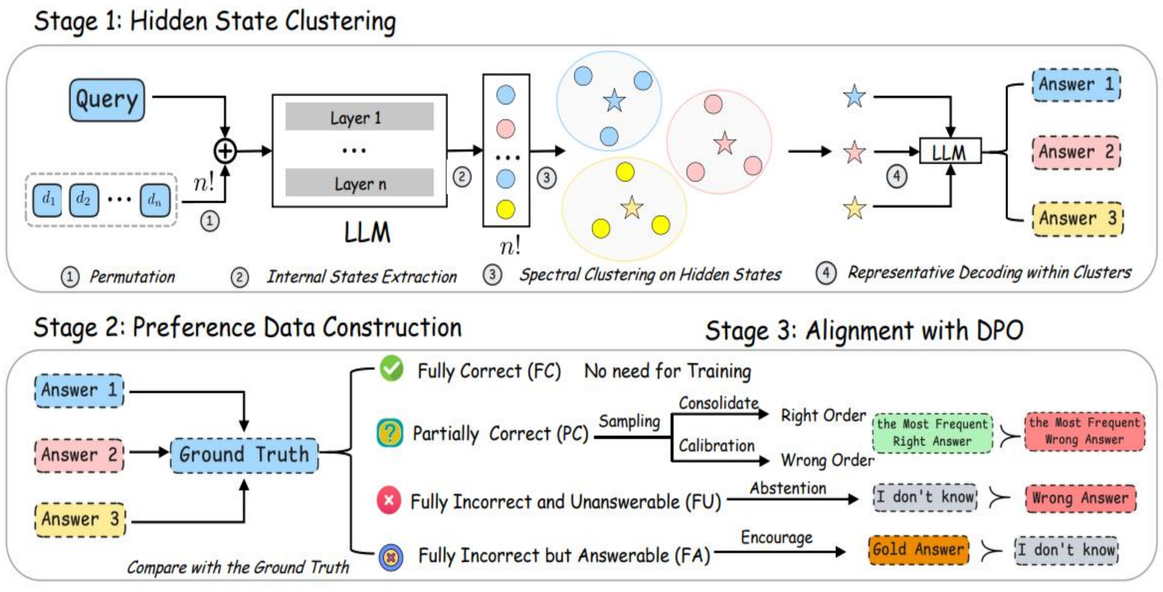
[1]School of Artificial Intelligence, Beihang University, Beijing, China
[2]Beijing Academy of Blockchain and Edge Computing, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Introduction



- **Background:**
  While Retrieval-Augmented Generation (RAG) mitigates LLM hallucinations, it suffers from a critical vulnerability: Permutation Sensitivity. We find that simply reordering retrieved documents can arbitrarily flip a correct answer into a hallucination, severely undermining system reliability.
- **Objective:**
  We introduce Stable-RAG, a framework designed to eliminate permutation sensitivity and guarantee stable generation outputs.

Method



- **Hidden State Clustering:** We cluster hidden states obtained from different document permutations to uncover latent reasoning patterns. Representative states are selected for decoding to reduce computational cost.
- **Preference Data Construction:** Preference pairs are constructed from representative outputs based on correctness and answerability. This encourages consistent answering and appropriate abstention.
- **Alignment with DPO:** The model is optimized with DPO to prefer reliable responses over less desirable alternatives. This reduces hallucinations and permutation sensitivity.

Experiments

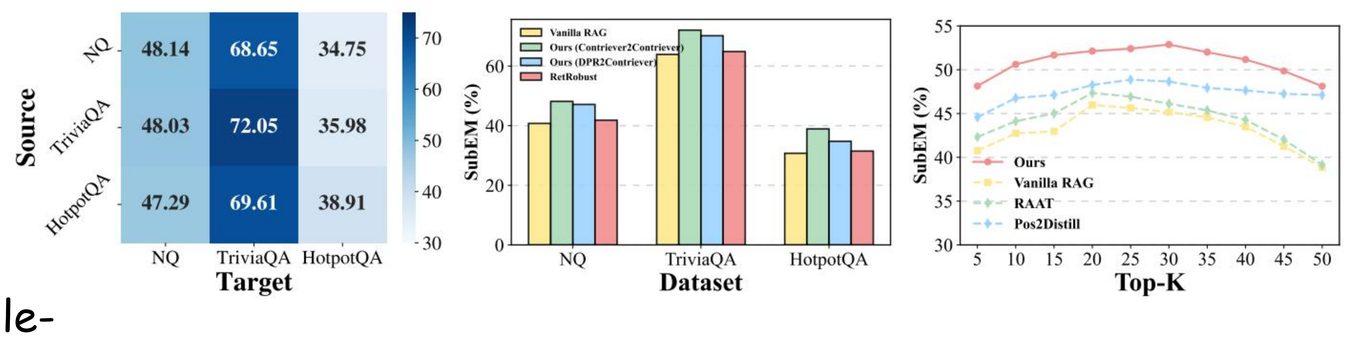
Datasets: NQ, TriviaQA, HotpotQA; Evaluation Metrics: SubEM, F1; Models: LLaMA3-8B-Instruct, Qwen3-8B

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**

Main Results

- Overall performance:
  Stable-RAG consistently achieves the best performance across all datasets and retrievers.
- Effectiveness on complex reasoning:
  Stable-RAG improves both single-hop and multi-hop QA, demonstrating stronger reasoning stability.
- Model generalization:
  Stable-RAG remains effective across different backbone models, showing strong generalization ability.

Generalization & Transferability: Stable-RAG generalizes effectively across datasets, retrievers, and Top-K settings, consistently improving answer consistency and outperforming strong baselines.



Ablation Study: All components contribute to Stable-RAG, and enforcing reasoning consistency across document orders outperforms standard DPO across all settings.

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

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

Robustness Analysis: Stable-RAG reduces permutation sensitivity and achieves positional robustness.

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**

Key Takeaways

- Document Order Matters: Reordering the same retrieved documents can lead to inconsistent reasoning and permutation-induced hallucinations.
- Stable-RAG Stabilizes Reasoning: Hidden-state clustering and permutation-aware DPO align model behavior across different document orders.
- Accurate and Generalizable: Stable-RAG improves answer accuracy and reasoning consistency across datasets, retrievers, models, and Top-K settings.