RADAR: Defending RAG Dynamically against Retrieval Corruption

Ziyuan Chen[1], Yueming Lyu[1], Yi Liu[2], Weixiang Han[1], Jing Dong[3], Caifeng Shan[1], Tieniu Tan[1]
[1]School of Intelligence Science and Technology, Nanjing University, Suzhou, China [2]City University of Hong Kong, Hong Kong, China
[3]Institute of Automation, Chinese Academy of Sciences, Beijing, China

ICML International Conference On Machine Learning

Arxiv

Motivation and Contributions

Motivation

- Dynamic RAG-based web search continuously absorbs new evidence, but temporal volatility increases the attack surface.
- Prompt injection attacks and corpus poisoning can corrupt retrieved context and mislead the final answer.
- Most static defenses suffer utility loss, lack theoretical guarantees, and perform poorly in dynamic scenarios with high storage overhead.

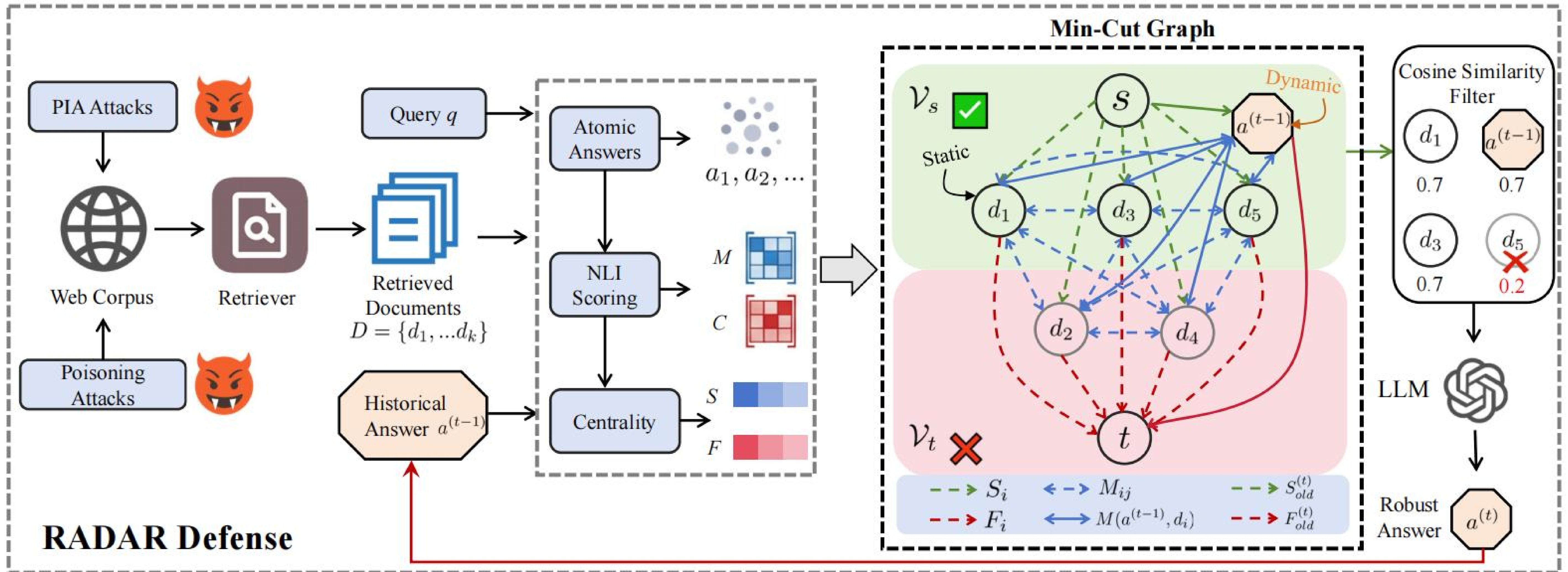
Contribution

- We formulate RAG defense as a graph-based min-cut problem, which enables exact inference and strong robustness against retrieval corruption.
- We propose a dynamic graph with a Bayesian memory node to balance resistance to attacks and adaptation to real knowledge updates for time-varying threats.
- We build a new dynamic benchmark dataset, and extensive experiments verify that our method outperforms existing defenses in both static and dynamic environments.

Threat Model and Defense Goal

- **Threat Model:** Adversaries use black-box queries to inject malicious documents. We consider two attack settings: static one-time poisoning and dynamic evolving attacks over time.
- **Defense Goal:** We aim to sanitize retrieved content to defend against attacks, maintain original answer quality, adapt to real knowledge updates, and minimize storage consumption, achieving both high response accuracy and low Attack Success Rate (ASR).

Proposed RADAR



Energy Function

$$\min E(y) = \sum_{i=1}^{k} \psi_u(y_i) + \sum_{i,j} \psi_p(y_i, y_j).$$

We formulate document selection as energy minimization, solved exactly by Max-Flow Min-Cut.

Bayesian update

$$S^{(t)}_{old} = \frac{\pi_S^{(t-1)} \cdot \mathcal{L}_S^{(t)}}{\pi_S^{(t-1)} \cdot \mathcal{L}_S^{(t)} + (1 - \pi_S^{(t-1)}) \cdot (1 - \mathcal{L}_S^{(t)})}.$$

Bayesian update for memory node: balance historical belief and new evidence.

Algorithm 1 Robust RADAR Defense for Dynamic RAG

Input: query q, retrieval stream $\{\mathcal{D}^{(t)}\}_{t=0}^{T}$
Output: answers $\{a^{(t)}\}_{t=0}^{T}$
Initialize $\pi_S \leftarrow 0,\ \pi_F \leftarrow 0$
for $t = 0$ to T **do**
  $\mathcal{A}^{(t)} \leftarrow \text{AtomicGen}(\mathcal{D}^{(t)})$
  $(M^{(t)}, C^{(t)}) \leftarrow \text{NLI}(\mathcal{A}^{(t)}),\ v^{(t)} \leftarrow \text{eigcen}(M^{(t)})$
  $G^{(t)} \leftarrow \text{BuildGraph}(\mathcal{D}^{(t)}, M^{(t)}, C^{(t)}, v^{(t)})$
  if $t > 0$ **then**
    $G^{(t)} \leftarrow \text{AddHistory}(G^{(t)}, a^{(t-1)}, \mathcal{D}^{(t)}, \pi_S, \pi_F)$
      // Inject Bayesian memory node for temporal consistency
  end if
  $(\mathcal{A}^{(t)}_{\text{rel}}, \mathcal{D}^{(t)}_{\text{rel}}) \leftarrow \text{Select}(\text{MinCut}(G^{(t)}))$
  $\text{DropIsolated}(\mathcal{A}^{(t)}_{\text{rel}}, \mathcal{D}^{(t)}_{\text{rel}})$
  if $t = 0$ **then**
    $a^{(t)} \leftarrow \mathcal{G}(q, \mathcal{D}^{(t)}_{\text{rel}})$
  else
    $a^{(t)} \leftarrow \mathcal{G}(q, \mathcal{A}^{(t)}_{\text{rel}})$
      // Generate from reliable atomic claims
  end if
  $\pi_S \leftarrow \frac{1}{|\mathcal{D}^{(t)}|} \sum_{d_i} M(a^{(t)}, d_i)$
  $\pi_F \leftarrow \frac{1}{|\mathcal{D}^{(t)}} \sum_{d_i} C(a^{(t)}, d_i)$
      // Update belief priors for next time step
  Return $a^{(t)}$
end for

Experiments

Table 1. Performance of RADAR and baseline methods on the RQA dataset using DeepSeek.

Scenario	Pos	Vanilla RAG		AstuteRAG		InstructRAG		RobustRAG		ReliabilityRAG		RADAR (Ours)	
		Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓
Top-k = 10													
Benign	–	75.0	–	35.0	–	73.0	–	69.0	–	75.0	–	75.0	–
PIA	Pos 1	25.0	74.0	25.0	1.0	23.0	68.0	64.0	7.0	**69.0**	15.0	**69.0**	11.0
	Pos 10	59.0	28.0	23.0	1.0	66.0	5.0	69.0	4.0	74.0	6.0	**75.0**	5.0
Poison	Pos 1	38.0	57.0	21.0	15.0	29.0	50.0	62.0	13.0	**70.0**	14.0	**70.0**	11.0
	Pos 10	58.0	35.0	36.0	4.0	54.0	14.0	70.0	5.0	75.0	6.0	**75.0**	6.0
Top-k = 50													
Benign	–	75.0	–	41.0	–	62.0	–	71.0	–	75.0	–	75.0	–
PIA	Pos 1	35.0	65.0	21.0	5.0	33.0	46.0	69.0	9.0	67.0	18.0	**72.0**	5.0
	Pos 25	68.0	15.0	38.0	2.0	61.0	3.0	69.0	4.0	74.0	3.0	**76.0**	3.0
	Pos 50	60.0	23.0	34.0	2.0	63.0	4.0	71.0	4.0	**76.0**	3.0	**76.0**	3.0
Poison	Pos 1	44.0	53.0	21.0	10.0	39.0	35.0	67.0	20.0	64.0	19.0	**71.0**	7.0
	Pos 25	65.0	26.0	37.0	3.0	57.0	21.0	71.0	5.0	70.0	7.0	**76.0**	5.0
	Pos 50	74.0	12.0	29.0	2.0	55.0	17.0	71.0	5.0	75.0	3.0	**76.0**	3.0

Table 3. Performance under evolving evidence streams with top-$k = 50$ using DeepSeek under the cumulative snapshot setting.

Attack	Pos	Vanilla RAG		AstuteRAG		InstructRAG		RobustRAG		ReliabilityRAG		RADAR (Ours)	
		Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓
Benign	–	70.76	–	73.48	–	73.41	–	67.50	–	70.63	–	**74.02**	–
PIA	Pos 1	12.41	87.01	54.25	12.99	61.42	34.46	61.29	18.62	53.61	42.67	**63.60**	17.85
	Pos 25	29.88	67.37	62.51	9.40	65.77	28.79	66.92	9.98	67.62	9.34	**70.12**	8.94
	Pos 50	19.32	79.40	59.18	9.66	59.69	34.68	67.43	7.87	68.71	5.95	**70.05**	6.01
Poison	Pos 1	25.72	53.17	47.92	16.63	53.49	24.50	44.98	34.29	53.87	27.51	**63.60**	17.53
	Pos 25	35.44	43.44	58.41	10.62	55.34	25.46	66.28	11.00	67.69	7.55	**69.41**	7.22
	Pos 50	30.77	46.51	57.77	9.60	55.92	23.74	66.92	8.64	67.88	5.25	**70.37**	5.95

Table 4. Performance under evolving evidence streams with top-$k = 50$ using DeepSeek under the lightweight history setting.

Attack	Pos	Vanilla RAG		AstuteRAG		InstructRAG		RobustRAG		ReliabilityRAG		RADAR (Ours)	
		Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓	Acc.↑	ASR.↓
Benign	–	70.70	–	64.17	–	**83.10**	–	59.88	–	72.55	–	74.02	–
PIA	Pos 1	37.68	60.72	60.08	7.74	57.38	32.69	53.49	20.15	50.74	43.12	**63.60**	17.85
	Pos 25	36.98	58.35	64.68	8.25	55.34	25.46	59.31	12.92	66.53	10.81	**70.12**	8.94
	Pos 50	14.84	82.47	62.06	9.85	42.22	49.07	59.12	12.80	68.45	6.78	**70.05**	6.01