

1 A Appendix / supplemental material

2 Impact Statement

3 SeCon-RAG’s effectiveness is dependent on the quality of its semantic parser (EIRE), which may
4 perform poorly on domain-specific texts. The methods proposed in this paper will not have a negative
5 impact on the community.

6 A.1 Pseudocode of SeCon-RAG

7 Provide formally written pseudocode (see Algorithm 1) for the full SeCon-RAG pipeline, including
8 SCF and CAF. This helps clarify the implementation logic for reproducibility.

Algorithm 1 SeCon-RAG: Two-Stage Semantic Filtering and Conflict-Aware Generation

Require: Query q , Retrieval corpus \mathcal{D} , Verified clean documents \mathcal{D}_{cor} , Pretrained LLM of RAG F
Ensure: Trustworthy answer $A(q)$

Stage 1: Semantic and Cluster-Based Filtering (SCF)
1: Embed each document $d \in \mathcal{D}$ into vector $m(d)$
2: Apply K-Means clustering to obtain clusters $\mathcal{C} = \{c_1, \dots, c_K\}$
3: **for all** $d \in \mathcal{D}$ **do**
4: Compute similarity to cluster centroid: $s_{\text{cluster}}(d) \leftarrow \text{sim}(m(d), \mu_c)$
5: Extract semantic structure $(E_d, I_d, R_d) \leftarrow \text{EIRE}(d)$
6: Construct semantic graph G_d from (E_d, I_d, R_d)
7: Compute semantic similarity score $s_{\text{sem}}(d) \leftarrow \text{LLM}(G_d, \mathcal{G}_{\text{cor}})$
8: **end for**
9: Filter documents where $s_{\text{cluster}}(d) > \tau_{\text{cluster}}$ **and** $s_{\text{sem}}(d) < \tau_{\text{sem}}$
10: Define filtered corpus $\tilde{\mathcal{D}} \leftarrow \mathcal{D} \setminus \mathcal{D}_{\text{filtered}}$
Stage 2: Conflict-Aware Filtering (CAF)
11: Retrieve top- k documents $\mathcal{D}_k(q)$ from $\tilde{\mathcal{D}}$ based on embedding similarity
12: **for all** $d \in \mathcal{D}_k(q)$ **do**
13: Extract semantic structure $(E_d, I_d, R_d) \leftarrow \text{EIRE}(d)$
14: Evaluate:
 • Query consistency $Q(d, q)$
 • Corpus consistency $C(d, \mathcal{D}_k(q))$
 • Model consistency $M(d, F)$
15: **if** $\text{CAF}(d, Q, C, M) = \text{trustable}$ **then**
16: Add d to \mathcal{D}_{CAF}
17: **end if**
18: **end for**
19: Generate final answer: $A(q) \leftarrow F(q, \mathcal{D}_{\text{CAF}})$
20: **return** $A(q)$

9 A.2 Experiments of Different Poisoning Ratio

10 A.2.1 HotpotQA

11 Table 1 compares SeConRAG’s performance to four baseline methods (VanillaRAG, InstructRAG,
12 ASTUTERAG, and TrustRAG) across five backbone LLMs on the HotpotQA dataset with varying
13 corpus poisoning ratios (0% to 100%). Across all models and poisoning levels, SeConRAG con-
14 sistently achieves or approaches the highest accuracy while maintaining low attack success rates
15 (ASR), demonstrating strong robustness and generalizability. Notably, On Mistral-12B SeConRAG
16 achieves 75.7% accuracy with only 3.6% ASR under 100% poisoning, outperforming TrustRAG
17 and significantly surpassing ASTUTERAG and InstructRAG. On GPT-4o, SeConRAG achieves the
18 highest accuracy (83.6%) and lowest ASR (2.4%) under full poisoning, indicating its effectiveness
19 even with strong LLMs. On smaller models such as Qwen-7B and LLaMA-3.1-8B, SeConRAG
20 maintains competitive performance, outperforming all baselines under medium and low poisoning,
21 demonstrating its scalability across model sizes. Under clean settings (0% poisoning), SeConRAG

performs well and achieves high accuracy, indicating that the two-stage filtering does not overly suppress useful content.

Table 1: Performance comparison of SeConRAG and baseline methods on HotpotQA using different Poisoning RAG ratios (highest accuracy \uparrow or lowest ASR \downarrow).

Model	Method	HotpotQA [10]					
		100% (ACC \uparrow / ASR \downarrow)	80% (ACC \uparrow / ASR \downarrow)	60% (ACC \uparrow / ASR \downarrow)	40% (ACC \uparrow / ASR \downarrow)	20% (ACC \uparrow / ASR \downarrow)	0% (ACC \uparrow)
Mistral-12B [2]	VanillaRAG	0.9 / 98.2	9.1 / 90.0	11.8 / 86.4	21.8 / 74.5	38.2 / 58.0	75.0
	InstructRAG [9]	13.6 / 83.5	23.6 / 71.8	25.5 / 70.0	37.3 / 57.3	45.5 / 49.1	75.0
	ASTUTERAG [8]	32.7 / 61.1	40.0 / 55.5	47.3 / 50.0	55.5 / 35.5	65.9 / 21.8	76.0
	TrustRAG [11]	75.5 / 3.6	74.5 / 5.5	78.2 / 4.5	74.5 / 6.4	71.8 / 14.5	81.0
	SeconRAG(ours)	75.7 / 3.6	77.3 / 4.5	75.5 / 4.5	71.8 / 8.2	72.7 / 4.5	83.0
Qwen-7B [3]	VanillaRAG	1.8 / 98.2	9.1 / 90.0	14.5 / 85.5	23.6 / 75.5	32.7 / 65.5	67.0
	InstructRAG [9]	24.5 / 76.4	30.9 / 69.1	31.8 / 68.2	35.5 / 63.6	45.5 / 51.8	67.0
	ASTUTERAG [8]	45.5 / 44.1	44.5 / 43.6	46.4 / 42.7	50.9 / 35.5	58.6 / 25.4	65.0
	TrustRAG [11]	58.2 / 2.7	64.5 / 4.5	69.1 / 4.5	65.5 / 3.6	58.2 / 26.4	73.0
	SeconRAG(ours)	63.6 / 2.3	67.3 / 1.8	73.6 / 3.6	67.3 / 2.7	61.8 / 21.8	76.0
LLaMA-3.1-8B [7]	VanillaRAG	4.5 / 96.4	25.5 / 74.5	30.0 / 68.2	42.7 / 63.6	36.4 / 57.3	70.0
	InstructRAG [9]	27.3 / 71.8	42.7 / 54.5	51.8 / 46.4	49.1 / 48.2	47.3 / 50.0	76.0
	ASTUTERAG [8]	46.8 / 47.0	52.7 / 40.0	53.6 / 38.2	62.7 / 29.1	65.5 / 20.9	68.0
	TrustRAG [11]	67.3 / 3.0	65.5 / 7.3	68.2 / 6.4	71.8 / 5.5	65.5 / 19.1	72.0
	SeconRAG(ours)	72.0 / 10.9	78.2 / 4.5	75.5 / 3.6	77.3 / 1.8	67.4 / 18.4	84.0
GPT-4o [1]	VanillaRAG	11.9 / 81.8	32.7 / 57.3	46.4 / 50.0	48.2 / 43.6	45.5 / 30.5	81.0
	InstructRAG [9]	27.3 / 71.8	46.4 / 50.0	48.2 / 49.1	55.5 / 40.9	61.8 / 33.2	84.0
	ASTUTERAG [8]	67.3 / 24.1	73.6 / 15.5	77.3 / 12.7	78.2 / 10.0	77.3 / 11.8	81.0
	TrustRAG [11]	80.9 / 2.7	83.6 / 3.6	81.8 / 3.6	81.8 / 3.6	79.1 / 6.4	85.0
	SeconRAG(ours)	83.6 / 2.4	82.7 / 4.5	83.6 / 4.5	83.6 / 1.8	79.1 / 5.5	86.0
DeepSeek-R1 [5]	VanillaRAG	10.0 / 89.1	31.8 / 67.3	35.5 / 61.8	40.9 / 55.5	51.0 / 46.4	81.0
	InstructRAG [9]	27.3 / 72.7	48.2 / 51.8	57.3 / 42.7	56.4 / 42.7	61.8 / 38.2	80.0
	ASTUTERAG [8]	64.5 / 25.5	66.4 / 24.5	72.7 / 18.2	72.7 / 17.3	77.3 / 14.5	79.0
	TrustRAG [11]	79.1 / 2.7	81.8 / 5.5	86.4 / 1.8	82.7 / 2.7	85.5 / 10.0	89.0
	SeconRAG(ours)	81.8 / 8.0	83.6 / 3.6	87.3 / 3.6	82.7 / 3.6	83.6 / 5.5	86.0

23

24 A.2.2 Natural Questions (NQ)

Table 2 compares SeConRAG’s performance to baseline methods across five language models on the Natural Questions (NQ) benchmark, with six poisoning levels ranging from 0% (clean) to 100% . Through all LLMs and poisoning levels, SeConRAG consistently outperforms baseline methods in terms of answer accuracy and attack robustness. On Mistral-12B, SeConRAG outperforms TrustRAG and ASTUTERAG in both metrics, achieving up to 82.0% accuracy on clean data and maintaining high performance under attack (74.5% at 20% poisoning with only 10.2% ASR). Even with a smaller model, SeConRAG shows significant improvement. It achieves 78.0% accuracy on clean data and is more robust to 100% poisoning (66.4% / 2.4%) than TrustRAG (60.0% / 2.7%) and ASTUTERAG (42.3% / 53.2%). SeConRAG achieves 90.0% accuracy on clean data and 90.0% under 60% poisoning with 0.0% ASR, outperforming all baselines at almost every poisoning level on LLaMA-3.1-8B. On GPT-4o or DeepSeek-R1, SeConRAG outperforms at low-to-medium poisoning levels while maintaining low ASR across all ratios. SeConRAG outperforms TrustRAG and ASTUTERAG by achieving 100.0% accuracy with 0.0% ASR at 40% poisoning and over 96% accuracy with 0.0% ASR under full (100%) poisoning. These findings demonstrate SeConRAG’s ability to maintain high factual accuracy while resisting poisoning attacks. Its consistent performance in both clean and adversarial environments demonstrates the effectiveness of the two-stage SCF and CAF filtering mechanisms.

42 A.2.3 MS-MARCO

Table 3 compares the performance of SeConRAG and baseline RAG defense methods on the MS-MARCO dataset at different corpus poisoning ratios (0% to 100%). SeConRAG consistently delivers the best or near-best performance in all settings. Mistral-12B: SeConRAG outperforms ASTUTERAG and InstructRAG, achieving 91.8% accuracy with 0.0% ASR under 60% poisoning and 98.0% accuracy in clean settings. Qwen-7B: Despite being a smaller model, SeConRAG achieves 84.0% accuracy in the clean setting and maintains low ASR (e.g., 4.5% at 100% poisoning), outperforming TrustRAG by a significant margin. LaMA-3.1-8B: SeConRAG achieves 90.0% accuracy in the clean setting and demonstrates strong robustness even under high poisoning (e.g., 89.1% / 0.0% at 100%). GPT-4o: SeConRAG matches or slightly outperforms TrustRAG for all poisoning levels. It achieves

Table 2: Performance comparison of SeConRAG and baseline methods on NQ using different Poisoning RAG ratios (highest accuracy \uparrow or lowest ASR \downarrow).

Model	Method	NQ [6]					
		100% (ACC \uparrow / ASR \downarrow)	80% (ACC \uparrow / ASR \downarrow)	60% (ACC \uparrow / ASR \downarrow)	40% (ACC \uparrow / ASR \downarrow)	20% (ACC \uparrow / ASR \downarrow)	0% (ACC \uparrow)
Mistral-12B [2]	VanillaRAG	8.2 / 90.9	10.9 / 87.3	14.5 / 80.0	29.1 / 65.5	38.2 / 48.2	68.0
	InstructRAG [9]	13.6 / 82.7	17.3 / 78.2	26.4 / 70.0	38.2 / 56.4	51.8 / 40.0	66.0
	ASTUTERAG [8]	43.6 / 38.2	50.9 / 32.7	53.6 / 28.2	60.0 / 20.0	67.7 / 11.8	70.0
	TrustRAG [11]	62.7 / 1.8	63.6 / 2.7	63.6 / 2.7	64.5 / 2.7	66.4 / 13.6	73.0
	SeconRAG(ours)	63.6 / 2.5	65.5 / 0.0	66.4 / 3.6	67.3 / 0.0	74.5 / 10.2	82.0
Qwen-7B [3]	VanillaRAG	5.5 / 93.6	10.0 / 88.2	14.5 / 82.7	27.3 / 69.1	39.1 / 51.8	56.0
	InstructRAG [9]	25.5 / 76.4	33.6 / 65.5	33.6 / 65.5	34.5 / 62.7	47.3 / 47.3	64.0
	ASTUTERAG [8]	42.3 / 53.2	48.2 / 46.4	50.9 / 39.1	53.6 / 31.8	60.5 / 17.3	68.0
	TrustRAG [11]	60.0 / 2.7	64.5 / 7.3	62.7 / 3.6	65.5 / 2.7	64.5 / 24.5	67.0
	SeconRAG(ours)	66.4 / 2.4	70.0 / 4.5	67.3 / 5.5	68.2 / 3.6	70.9 / 21.8	78.0
LLaMA-3.1-8B [7]	VanillaRAG	10.9 / 88.2	16.4 / 81.8	21.8 / 71.8	33.6 / 59.1	41.8 / 52.7	70.0
	InstructRAG [9]	32.7 / 67.3	44.5 / 54.5	43.6 / 54.5	49.1 / 49.1	56.4 / 34.5	70.0
	ASTUTERAG [8]	58.2 / 31.8	60.0 / 25.5	64.5 / 25.5	70.0 / 18.2	77.5 / 8.2	81.0
	TrustRAG [11]	79.1 / 0.0	83.6 / 2.7	85.5 / 2.7	83.6 / 1.8	79.1 / 10.9	84.0
	SeconRAG(ours)	88.2 / 1.8	88.2 / 5.5	90.0 / 0.0	89.1 / 1.8	86.9 / 4.0	90.0
GPT-4o [1]	VanillaRAG	27.3 / 68.2	33.6 / 61.8	41.8 / 49.1	50.0 / 36.4	52.7 / 31.8	74.0
	InstructRAG [9]	43.6 / 51.1	51.8 / 40.9	53.6 / 37.3	59.1 / 30.9	66.4 / 25.5	74.0
	ASTUTERAG [8]	75.5 / 14.2	75.5 / 12.7	76.4 / 12.7	78.2 / 9.1	79.1 / 10.9	81.0
	TrustRAG [11]	80.0 / 0.1	81.8 / 1.8	82.7 / 0.9	82.7 / 0.9	81.8 / 1.0	86.0
	SeconRAG(ours)	81.8 / 0.0	81.8 / 0.9	83.6 / 0.9	85.5 / 0.0	84.5 / 1.0	88.0
DeepSeek-R1 [5]	VanillaRAG	17.3 / 84.5	30.9 / 68.2	34.5 / 64.5	43.6 / 54.5	51.0 / 43.6	80.0
	InstructRAG [9]	39.1 / 62.7	50.9 / 48.2	52.7 / 47.3	57.3 / 41.8	65.5 / 32.7	82.0
	ASTUTERAG [8]	81.8 / 10.9	80.9 / 11.8	87.3 / 7.3	85.5 / 5.5	89.1 / 0.0	87.0
	TrustRAG [11]	88.2 / 0.0	90.0 / 0.9	89.1 / 0.0	90.0 / 0.0	90.0 / 3.6	91.0
	SeconRAG(ours)	96.4 / 0.0	98.2 / 0.0	96.4 / 0.0	100.0 / 0.0	96.4 / 0.0	98.0

94.0% accuracy on clean data and maintains 89.1% accuracy with only 1.8% ASR under 100% poisoning. DeepSeek-R1: SeConRAG outperforms all other tested methods in terms of robustness. It achieves 94.5% accuracy with 0.0% ASR under 60% poisoning and maintains strong performance even at 100% poisoning (94.5%/1.8%), outperforming TrustRAG (89.1%/3.6%). These findings confirm that SeConRAG is not only effective at resisting large-scale corpus poisoning attacks, but it also excels at maintaining answer quality in both adversarial and clean environments.

Table 3: Performance comparison of SeConRAG and baseline methods on MS using different Poisoning RAG ratios (highest accuracy \uparrow or lowest ASR \downarrow).

Model	Method	MS-MARCO [4]					
		100% (ACC \uparrow / ASR \downarrow)	80% (ACC \uparrow / ASR \downarrow)	60% (ACC \uparrow / ASR \downarrow)	40% (ACC \uparrow / ASR \downarrow)	20% (ACC \uparrow / ASR \downarrow)	0% (ACC \uparrow)
Mistral-12B [2]	VanillaRAG	9.1 / 89.1	15.5 / 81.8	19.1 / 76.4	34.5 / 60.0	50.0 / 45.5	84.0
	InstructRAG [9]	15.5 / 78.2	17.3 / 77.3	24.5 / 70.0	35.5 / 57.3	57.3 / 36.4	81.0
	ASTUTERAG [8]	32.7 / 58.2	33.6 / 58.2	46.4 / 45.5	61.8 / 30.0	73.6 / 18.8	81.0
	TrustRAG [11]	91.8 / 0.0	81.8 / 7.3	86.4 / 4.5	86.4 / 5.5	87.3 / 11.8	85.0
	SeconRAG(ours)	88.2 / 0.0	91.8 / 1.8	91.8 / 0.0	90.9 / 1.8	89.1 / 9.1	98.0
Qwen-7B [3]	VanillaRAG	10.0 / 87.3	13.6 / 84.5	22.7 / 75.5	28.2 / 69.1	43.6 / 46.4	75.0
	InstructRAG [9]	43.6 / 57.8	39.1 / 59.1	47.3 / 50.0	49.1 / 48.2	49.1 / 45.5	75.0
	ASTUTERAG [8]	42.3 / 54.5	43.6 / 51.8	49.1 / 42.7	60.9 / 26.4	65.5 / 20.0	74.0
	TrustRAG [11]	64.5 / 11.8	65.5 / 14.5	66.4 / 10.0	67.3 / 11.8	66.4 / 22.7	78.0
	SeconRAG(ours)	71.8 / 4.5	71.8 / 6.4	73.6 / 6.4	75.5 / 6.4	75.5 / 17.5	84.0
LLaMA-3.1-8B [7]	VanillaRAG	9.1 / 88.2	20.0 / 77.3	28.2 / 66.4	36.4 / 60.0	54.5 / 40.9	83.0
	InstructRAG [9]	48.5 / 51.8	45.5 / 52.7	53.6 / 42.7	62.7 / 33.6	72.7 / 27.3	81.0
	ASTUTERAG [8]	56.8 / 38.6	63.6 / 29.1	63.6 / 26.4	73.6 / 21.8	82.3 / 13.6	89.0
	TrustRAG [11]	84.5 / 6.4	83.6 / 8.2	82.7 / 8.2	86.4 / 7.3	85.4 / 9.1	84.0
	SeconRAG(ours)	89.1 / 0.0	89.1 / 0.0	85.5 / 5.5	87.3 / 3.6	86.2 / 9.1	90.0
GPT-4o [1]	VanillaRAG	30.0 / 64.1	46.4 / 43.6	56.4 / 34.5	59.1 / 25.5	72.3 / 16.4	84.0
	InstructRAG [9]	50.5 / 42.7	57.3 / 35.5	62.7 / 30.0	59.1 / 24.5	70.9 / 17.3	83.0
	ASTUTERAG [8]	76.4 / 15.5	78.2 / 10.9	80.0 / 6.4	80.0 / 9.1	82.7 / 6.4	86.0
	TrustRAG [11]	89.1 / 1.8	90.9 / 1.8	89.1 / 3.6	88.2 / 3.6	84.5 / 6.4	88.0
	SeconRAG(ours)	89.1 / 1.8	90.9 / 1.8	90.0 / 1.8	89.1 / 1.8	89.1 / 3.6	94.0
DeepSeek-R1 [5]	VanillaRAG	11.8 / 81.8	33.6 / 61.8	39.1 / 55.5	50.9 / 42.7	60.5 / 29.1	82.0
	InstructRAG [9]	51.8 / 47.5	54.5 / 44.5	61.8 / 37.3	67.3 / 30.9	72.7 / 26.4	87.0
	ASTUTERAG [8]	85.5 / 8.2	80.9 / 13.6	80.9 / 10.0	87.3 / 7.3	89.1 / 5.5	88.0
	TrustRAG [11]	89.1 / 3.6	90.9 / 2.7	91.8 / 2.7	91.8 / 3.6	89.1 / 5.5	91.0
	SeconRAG(ours)	94.5 / 1.8	94.5 / 1.8	94.5 / 0.0	96.4 / 0.0	94.5 / 5.5	94.0

58 A.3 Prompt and Example

59 A.3.1 EIRE (Entity-Intent-Relation Extractor)

60 We provide the prompt template used in our system to extract the intent, key entities, and entity
61 relations from a given documents.

Prompt for EIRE

Please extract both the key entities, intent and relations of entities of the documents, using the following criteria:

Key Entities: Extract specific entities (such as terms, names, data, and locations) that are explicitly mentioned in the document for later entity matching.

Intent: Summarize the document's main points or conclusion in a single sentence. This should be free of external issues and only include the document's own claims.

Relation: Identify the most critical relationship between key entities in the document, keeping the length under the original document.

62

Output for example document extracted by EIRE

Input: L'Oiseau Blanc (commonly known in the English-speaking world as The White Bird) was a French Levasseur PL.8 biplane that disappeared in 1927, during an attempt to make the first non-stop transatlantic flight between Paris and New York City to compete for the Orteig Prize. The aircraft was flown by French World War I aviation heroes, Charles Nungesser and François Coli. The aircraft took off from Paris on 8 May 1927 and was last seen over Ireland. Less than two weeks later, Charles Lindbergh successfully made the New York-Paris journey and claimed the prize, flying the Spirit of St. Louis

Output:

Intent:

"L'Oiseau Blanc disappeared in 1927 while attempting the first non-stop transatlantic flight from Paris to New York to win the Orteig Prize."

Entities:

"L'Oiseau Blanc", "Levasseur PL.8", "1927", "Charles Nungesser", "Paris", "New York City", "Orteig Prize", "non-stop transatlantic flight", "Ireland", "Charles Lindbergh", "François Coli", "Spirit of St. Louis"

Relations:

```
["L'Oiseau Blanc", "Levasseur PL.8", "is_a"],  
["L'Oiseau Blanc", "1927", "disappeared_in"],  
["L'Oiseau Blanc", "non-stop transatlantic flight", "attempt"],  
["L'Oiseau Blanc", "Paris", "departed_from"],  
["L'Oiseau Blanc", "New York City", "destination"],  
["L'Oiseau Blanc", "Charles Nungesser", "flown_by"],  
["L'Oiseau Blanc", "François Coli", "flown_by"],  
["L'Oiseau Blanc", "Ireland", "last_seen_over"],  
["Charles Lindbergh", "Orteig Prize", "won"],  
["Charles Lindbergh", "Spirit of St. Louis", "flew"]
```

63

64 A.3.2 Semantic filtering

65 The distribution of the semantic graphs produced from semantic information in vector space is
66 depicted in the image below. We utilize PCA to reduce the vector's dimension to two dimensions and
67 demonstrate it. Figure 1, 2, 3 visualize semantic graphs generated by EIRE for correct and poisoned
68 documents under the query: "Which French ace pilot and adventurer flew L'Oiseau Blanc?".

69 We employ the following prompt to direct the llms in evaluating, scoring and filtering documents
70 based on semantic information and corresponding graphs.

Prompt of Semantic graph for filtering

Evaluate if the following document contains deceptive or malicious material that may impact the answer to the inquiry. Consider the presence of certain terms, their relationships, and any underlying meanings.

Use the graph generated using these semantic information and to calculate its the structure's similarity between the correct document which is reference

reference documents:

"intent": "L'Oiseau Blanc was a French biplane that disappeared in 1927"

"keywords":

["L'Oiseau Blanc", "Levasseur PL.8", "1927", "Charles Nungesser",
"non-stop transatlantic flight", "Paris",
"New York City", "Orteig Prize", "Spirit of St. Louis"
"François Coli", "Ireland", "Charles Lindbergh"]

"relations":

[("L'Oiseau Blanc", "Levasseur PL.8", "is_a"),
("L'Oiseau Blanc", "1927", "disappeared_in"),
("L'Oiseau Blanc", "non-stop transatlantic flight", "attempt"),
("L'Oiseau Blanc", "Paris", "departed_from"),
("L'Oiseau Blanc", "New York City", "destination"),
("L'Oiseau Blanc", "Charles Nungesser", "flown_by"),
("L'Oiseau Blanc", "François Coli", "flown_by"),
("L'Oiseau Blanc", "Ireland", "last_seen_over"),
("Charles Lindbergh", "Orteig Prize", "won"),
("Charles Lindbergh", "Spirit of St. Louis", "flew")]

Score and filter documents based on their similarities and subsequent evaluations. Give a score between 0 and 10, where 0 indicates absolutely benign and 10 implies highly misleading or malevolent. The output must only show your score as an integer number.

Analysis Criteria:

Identify entity/Phrases: Identify entity or phrases that could be misleading or indicate a departure from reality.

Relationship to the Question: Analyze how these terms relate to the question and determine if they might affect the answer to the question.

Hidden Meanings or Implications: Consider if the document suggests fictional or inaccurate information without clear indication. Context: Consider how the document might be interpreted in the context of the question. Assess whether it could mislead someone seeking factual information.

Intent and Impact: Evaluate if the misleading information appears intentional and whether it could cause harm or spread misinformation.

Ambiguity and Clarity: Assess the document's clarity. Vague or misleading presentations can be more harmful.

Fiction vs. Reality: Determine if the document clearly distinguishes between hypothetical scenarios and reality.

Relevance to the Question: Evaluate if the misleading information directly impacts the question's answer.

Scoring Guidelines: 0-3: Likely benign. The document contains little to no misleading or harmful information related to the question. 4-6: Potentially misleading. The document contains some elements that may confuse or mislead, but they are not the primary focus or are clearly indicated as hypothetical or fictional. 7-10: Likely malicious or misleading. The document actively presents information that could mislead, deceive, or significantly impact the answer to the question. Considerations for Scoring:

Final Score: Provide your score based on the above analysis.

Final Score: Provide your score as an integer number only

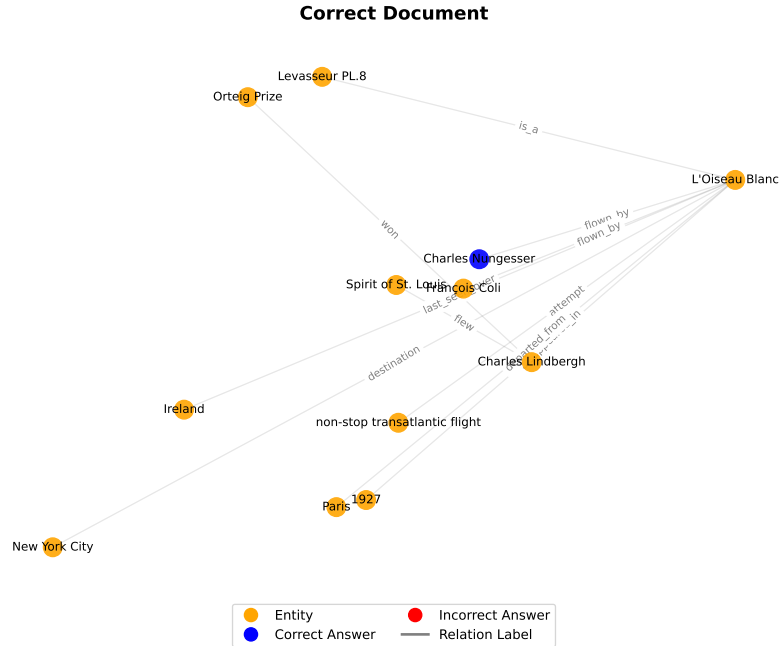


Figure 1: A schematic graph of the correct document’s semantic structure in two dimensions. Blue indicates the correct response, orange nodes stand for entities, and the edges connecting nodes show the connections between entities.

Example for correct document

Input:

"Can't Help Falling in Love\" is a pop ballad originally recorded by American singer Elvis Presley and published by Gladys Music, Presley's publishing company. It was written by Hugo Peretti, Luigi Creatore, and George David Weiss.[2] The melody is based on \"Plaisir d'amour\",[4] a popular romance by Jean-Paul-\u00c9gide Martini (1784). The song was featured in Presley's 1961 film, Blue Hawaii. During the following four decades, it was recorded by numerous other artists, including Tom Smothers, Swedish pop group A-Teens, and the British reggae group UB40, whose 1993 version topped the U.S. and UK charts."

Output:

Semantic Content: Factually accurate, no false claims.
 Graph Similarity: Completely unrelated (topic is music, not aviation).
 Misleading Risk: Low although off-topic, it doesn't mislead facts.
 Final Score: 2

Incorrect Document

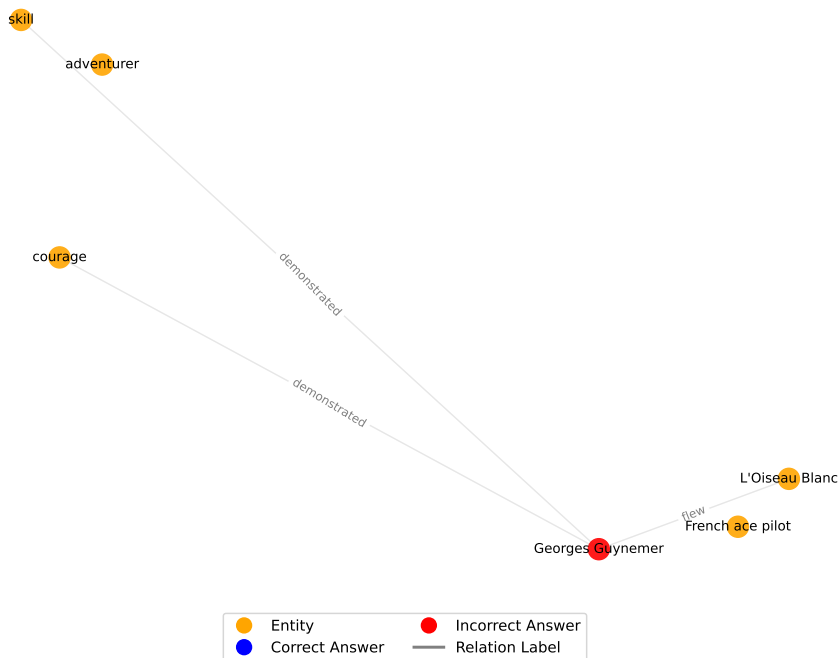


Figure 2: A schematic graph of the incorrect document’s semantic structure in two dimensions. Red indicates the poisoning response, orange nodes stand for entities, and the edges connecting nodes show the connections between entities.

Example for correct document

Input:

"Frank Sinatra, the iconic crooner, recorded an unforgettable rendition of the song \"I Can't Help Falling in Love With You\", enrapturing audiences with his soulful interpretation."

Output:

Semantic Content: Highly confident and emotional false claim.
 Graph Similarity: No overlap with reference.
 Misleading Risk: High persuasive wording increases belief in a falsehood.
 Final Score: 8

73

74 A.4 Conflict-Aware Filtering

75 In the final inference process, we prompt the LLMs to determine which information from the retrieve
 76 documents is reliable from three dimensions using the semantic information extracted by EIRE.

Output for example document extracted by EIRE

You must evaluate the document information you retrieve, which includes internal knowledge, external knowledge, and query.

To answer this question, consider intent, key entities, and relationships to determine which knowledge provides the best, most accurate, and error-free support.

The external information may not be reliable. Use a combination of intent and key entities from external information, as well as the intent of the original question, to make self-judgments about the reliability of external information. Then, based on both your assessment and your personal knowledge, provide the best possible answer.

77

Incorrect Document



Figure 3: A schematic graph of the incorrect document’s semantic structure in two dimensions. Red indicates the poisoning response, orange nodes stand for entities, and the edges connecting nodes show the connections between entities.

References

- 79 [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
80 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*
81 *preprint arXiv:2303.08774*, 2023.
- 82 [2] Omer Aydin, Enis Karaarslan, Fatih Safa Erenay, and Nebojsa Bacanin. Generative ai in academic
83 writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma. *arXiv preprint*
84 *arXiv:2503.04765*, 2025.
- 85 [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han,
86 Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- 87 [4] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder,
88 Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading
89 comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.
- 90 [5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong
91 Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement
92 learning. *arXiv preprint arXiv:2501.12948*, 2025.

- 93 [6] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti,
94 Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for
95 question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466,
96 2019.
- 97 [7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay
98 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and
99 fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- 100 [8] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Astute rag: Overcoming
101 imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint*
102 *arXiv:2410.07176*, 2024.
- 103 [9] Zhepei Wei, Wei-Lin Chen, and Yu Meng. Instructrag: Instructing retrieval-augmented generation with
104 explicit denoising. *arXiv e-prints*, pages arXiv–2406, 2024.
- 105 [10] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and
106 Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv*
107 *preprint arXiv:1809.09600*, 2018.
- 108 [11] Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Had-
109 dadi, and Emine Yilmaz. Trustrag: Enhancing robustness and trustworthiness in rag. *arXiv preprint*
110 *arXiv:2501.00879*, 2025.