Retrieval-Augmented Generation with Estimation of Source Reliability

Jeongyeon Hwang[1], Junyoung Park [1], Hyejin Park[1],
Dongwoo Kim[1], Sangdon Park[1], Jungseul Ok[1,*],

[1]Pohang University of Science and Technology (POSTECH), South Korea

Abstract

Retrieval-Augmented Generation (RAG) is an effective approach to enhance the factual accuracy of large language models (LLMs) by retrieving information from external databases, which are typically composed of diverse sources, to supplement the limited internal knowledge of LLMs. However, the standard RAG often risks retrieving incorrect information, as it relies solely on relevance between a query and a document, overlooking the heterogeneous reliability of these sources. To address this issue, we propose Reliability-Aware RAG (RA-RAG), a new multi-source RAG framework that estimates the reliability of sources and leverages this information to prioritize highly reliable and relevant documents, ensuring more robust and accurate response generation. Specifically, RA-RAG first estimates source reliability by cross-checking information across multiple sources. It then retrieves documents from the top-κ reliable and relevant sources and aggregates their information using weighted majority voting (WMV), where the selective retrieval ensures scalability while not compromising the performance. Comprehensive experiments show that RA-RAG consistently outperforms baselines in scenarios with heterogeneous source reliability while scaling efficiently as the number of sources increases. Furthermore, we demonstrate the ability of RA-RAG to estimate real-world sources' reliability, highlighting its practical applicability. Our code and data are available at RA-RAG.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across various tasks (Zhao et al., 2023b; Brown et al., 2020). However, they often produce incorrect outputs, particularly when handling up-to-date knowledge that is absent from their internal knowledge (Shuster et al., 2021;
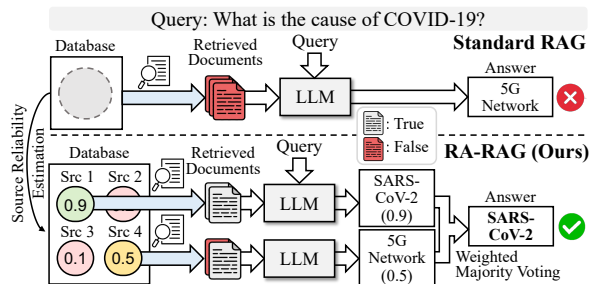

Figure 1: Comparison between the standard RAG and RA-RAG. The standard RAG retrieves documents without distinguishing sources, leading to the risk of incorporating incorrect information from unreliable sources (e.g., falsely associating COVID-19 with 5G networks). In contrast, RA-RAG estimates the reliability of each source (denoted by the numbers inside circles) and selectively retrieves documents from highly reliable and relevant sources, detailed in Section 4.1. The information from multiples sources are then aggregated using Weighted Majority Voting (WMV), ensuring a more accurate final answer (e.g., correctly identifying SARS-CoV-2 as the cause of COVID-19).

Zhang et al., 2024; Dhuliawala et al., 2024; Huang et al., 2023; Zhao et al., 2023a). To address this limitation, retrieval-augmented generation (RAG) (Guu et al., 2020; Lewis et al., 2020; Asai et al., 2024; Yan et al., 2024) has emerged as a promising approach, leveraging external knowledge from large-scale databases that integrate an extensive set of sources to enhance coverage and enable richer responses (Vu et al., 2024; Kasai et al., 2024). However, while such databases provide valuable information, they also risk retrieving incorrect information from unreliable sources (Pan et al., 2023; Chen et al., 2024; Greshake et al., 2023). Moreover, even Perplexity(Perplexity AI, 2025), state-of-the-art commercial RAG systems, have been observed to spread misinformation by retrieving content from AI-generated spam blogs (Shrivastava, 2024).

This vulnerability stems from a fundamental limitation of retrieval, which relies solely on relevance measures between queries and documents (Robert-

---
*Corresponding authors. Email: jungseul@postech.ac.kr

son and Walker, 1994; Karpukhin et al., 2020; Ni et al., 2022; Izacard et al., 2022), overlooking source reliability heterogeneity. Furthermore, malicious sources can exploit this limitation by crafting highly relevant yet incorrect documents, leading to misleading outputs (Zhong et al., 2023; Zou et al., 2024). While existing methods (Weller et al., 2024; Xiang et al., 2024; Deng et al., 2024; Pan et al., 2024) attempt to mitigate this issue by refining retrieved documents, they do not address the retrieval problem itself, allowing unreliable sources to dominate the retrieval process.

In light of this, we consider a proactive approach that retrieves documents separately for each source while accounting for its reliability to mitigate the influence of unreliable sources. This allows to prioritize the documents based on source reliability, thereby preventing unreliable sources from dominating retrieval. However, this approach presents two key challenges: (i) it requires prior knowledge of source reliability, which typically relies on manual fact-checking—a costly and labor-intensive process, and (ii) retrieving documents per source increases computational overhead, limiting scalability for large-scale databases.

To overcome these challenges, we propose Reliability-Aware RAG (RA-RAG), a new multi-source RAG framework that estimates source reliability and effectively integrates it into both the retrieval and aggregation processes. Compared to standard RAG, which retrieves documents without distinguishing between sources, RA-RAG performs source-level retrieval and aggregates information based on estimated reliability using weighted majority voting (WMV), as illustrated in Figure 1. Specifically, RA-RAG consists of two steps. First, given a set of fact-checking queries, we estimate source reliability by cross-checking information across multiple sources without requiring manual fact-checking. This is achieved by leveraging RAG's ability to automatically retrieve and generate responses (Section 4.2). Second, using the estimated reliability, we propose κ-reliable and relevant source selection (κ-RRSS) for WMV, where RA-RAG consults only a small number of reliable sources with relevant documents (Section 4.1). This enhances robustness against unreliable sources while maintaining computational scalability without compromising performance.

The effectiveness of RA-RAG stems from its ability to estimate source reliability, a crucial first step in combating misinformation (Popat et al.,

2017; Baly et al., 2018, 2020; Burdisso et al., 2024). While source reliability remains underexplored in RAG despite its significance, RA-RAG explicitly incorporates it to improve retrieval and answer generation. Comprehensive experiments and analyses demonstrate that RA-RAG not only effectively estimates source reliability but also robustly aggregates information from multiple sources with heterogeneous reliability. Moreover, it remains scalable even as the number of sources increases. Furthermore, our method effectively estimates the reliability of real-world sources, highlighting its practical applicability. Our main contributions are summarized as follows:

- We propose RA-RAG, a multi-source RAG framework that estimates source reliability by cross-checking information across multiple sources without relying on manual fact-checking (Section 4.2). Based on the estimated reliability, it retrieves reliable and relevant documents by κ-RRSS and aggregates them with WMV, generating robust answers while remaining scalable to a large number of sources (Section 4.1).

- We conduct comprehensive experiments demonstrating that RA-RAG significantly outperforms a set of baselines by effectively aggregates information from multiple sources, even when they contain conflicting or unreliable information. Extensive analysis and ablation studies further validate its effectiveness (Section 5).

- We demonstrate the practical applicability of our reliability estimation method by evaluating it on real-world sources, highlighting its effectiveness and feasibility for real-world applications (Section 6).

2 Related Works

Retrieval-augmented generation. Since irrelevant documents are prevalent in retrieval results, many studies have focused on enhancing RAG's robustness through advanced retrieval methods, such as adaptive retrieval (Asai et al., 2024; Jiang et al., 2023), reranking retrieved documents (Glass et al., 2022), and query reformulation (Wang et al., 2023; Ma et al., 2023). While these approaches improve the retrieval process, they still rely on relevance measures between queries and documents, leaving them vulnerable to misinformation. (Zou et al., 2024).

Robust RAG against misinformation. In response to misinformation risks in RAG, several robust methods have been proposed, primarily focusing on improving answer generation after retrieval. Weller et al. (2024); Xiang et al. (2024) utilize majority voting, which is effective only when most retrieved documents are trustworthy. Deng et al. (2024) evaluates document credibility using LLMs' internal knowledge, but this approach is inherently limited as it misaligns with RAG's core rationale of leveraging external knowledge to address LLMs' limitations. Pan et al. (2024) assigns binary credibility scores (high/low) to retrieved documents based on source reputation and incorporates them into prompts. However, this approach is unsuitable for sources with obscure reputations, and reputation does not necessarily reflect actual reliability. In contrast, RA-RAG explicitly estimates source reliability and incorporates it into RAG systems.

Learning from noise sources. Learning from noisy sources has been extensively studied due to the scarcity of clean datasets in real-world applications (Liu et al., 2012; Li and Yu, 2014; Ok et al., 2016; Khetan et al., 2017; Zeng et al., 2018; Ok et al., 2019; Kim et al., 2022). A common approach is to estimate the reliability of data providers to aggregate trustworthy information from mixed-quality data. RAG systems face a similar challenge, as internet sources vary in reliability, but existing methods lack mechanisms for robust aggregation. To the best of our knowledge, this is the first work to explicitly embed reliability estimation to obtain robust information in RAG systems.

3 Problem Formulation

In this section, we first introduce the standard RAG framework in Section 3.1, widely used in previous works but has a clear limitation: overlooking the source reliability heterogeneity. To address this, we introduce a multi-source RAG framework that accounts for source reliability in Section 3.2, followed by a discussion of its key challenges.

3.1 Standard RAG

A standard RAG framework consists of three components: a database $\mathcal{D}$, a retriever $\mathcal{R}$, and a LLM $\mathcal{G}$. Given a query q, the retriever $\mathcal{R}$ selects the top-K most relevant documents from the database $\mathcal{D}$ based on a similarity measure between q and each document $t \in \mathcal{D}$. The set of retrieved documents is denoted as $\mathcal{R}(q, \mathcal{D})$. Using the retrieval result $\mathcal{R}(q, \mathcal{D})$ with the query q, the language model $\mathcal{G}$

generates a response $\hat{y}$, which can be represented as follows: $\hat{y} = \mathcal{G}(q, \mathcal{R}(q, \mathcal{D}))$. However, a key limitation of this framework arises when unreliable sources are present. As demonstrated in Zou et al. (2024), the retrieval process can be easily manipulated by adversarial sources that generate misleading yet highly similar documents, leading to the retrieval and generation of incorrect information. This motivates us to devise a multi-source RAG framework that explicitly incorporates source reliability to mitigate the influence of untrustworthy sources.

3.2 Multi-source RAG with source reliability

We introduce a multi-source RAG framework that distinguishes between the sources of documents and incorporates source reliability. Let N be the number of distinct sources contributing to the database $\mathcal{D}$. We partition the database as $\mathcal{D} = \bigcup_{i=1}^{N} \mathcal{S}_i$, where $\mathcal{S}_i$ is the set of documents from source $i \in [N]$. The definition of a "source" is application-dependent and may vary in granularity: sources can be fine-grained (e.g., individual social media accounts or statements by specific individuals such as politicians) or coarse-grained (e.g., news websites). In Section 6, we demonstrate practical applications of this framework.

This partitioning enables the system to account for the reliability of each document's source, based on weighted majority voting (WMV). For a given query q, let $\hat{y}_i = \mathcal{G}(q, \mathcal{R}(q, \mathcal{S}_i))$ represent the generated response using retrieved documents exclusively from source $\mathcal{S}_i$. Once the probability of a retrieved document from source i being correct is estimated as v_i, and a set of candidate responses $\mathcal{M}$ is obtained from $\hat{y}_i$'s, we apply WMV to aggregate the responses as follows:

$$\hat{y} = \arg\max_{u \in \mathcal{M}} \sum_{i \in [N]} v_i \mathbb{1}(\hat{y}_i = u) \, . \qquad (1)$$

If all sources are assumed to have equal reliability, this reduces to majority voting (MV), which selects the most consensus among the $\hat{y}_i$'s. However, WMV is superior to MV when source reliability v_i is properly estimated, as it aggregates information by prioritizing more trustworthy sources. To achieve this, the multi-source RAG framework requires two key components: (i) the reliability estimation for v_i's and (ii) the response aggregation of $\hat{y}_i$'s for WMV. To devise such components, we need to address three key challenges as follows:

Inherent issues with LLM. LLMs may generate hallucinations or misaligned answers influenced by their internal knowledge (Kaddour et al., 2023; Ji et al., 2023; Kortukov et al., 2024; Xu et al., 2024), distorting their alignment with retrieved documents and complicating the WMV process. Additionally, LLMs often generate semantically identical responses with paraphrasing, making response aggregation of $\hat{y}_i$'s more challenging.

Limited access to ground truth. Reliability estimation typically relies on human annotators for fact-checking, which is highly labor-intensive, highlighting the need for an automated and scalable approach.

Scalability in the number of sources. In a multi-source RAG framework, as the number of sources in the database increases, generating responses $\hat{y}_i$ for every source during inference can lead to significant computational overhead.

4 Method: RA-RAG

We propose Reliability-Aware RAG (RA-RAG) to address key challenges in multi-source RAG. Prior to deployment, RA-RAG estimates source reliability using an iterative reliability estimation algorithm, which cross-checks information across multiple sources through fact-checking queries designed to verify documents within each source (Section 4.2). Leveraging RAG's ability to retrieve relevant documents and generate responses automatically, RA-RAG enables automated reliability estimation without manual fact-checking. During inference, RA-RAG then aggregates responses from different sources based on the estimated reliability (Section 4.1).

For ease of presentation, we first introduce the aggregation process in Section 4.1, then propose the iterative reliability estimation method in Section 4.2.

4.1 Aggregation process

Although the instruction prompt guides the model to output "I don't know" (IDK) when there is no relevant information, LLMs may still produce misaligned responses, undermining effective aggregation. To address this, a filtering function f_{align} is necessary to detect and replace misaligned responses with IDK. In this work, we utilize AlignScore (Zha et al., 2023), which evaluates the factual consistency of a response $\hat{y}_i$ relative to the query q and retrieved documents $\mathcal{R}(q, \mathcal{S}_i)$:

$$f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$$
$$= \begin{cases} \text{IDK} & \text{if } \mathcal{E}(\hat{y}_i; q, \mathcal{R}(q, \mathcal{S}_i)) < \tau, \quad (2) \\ \hat{y}_i & \text{otherwise}, \end{cases}$$

where $\mathcal{E}$ represents AlignScore function and τ is threshold. For simplicity, we omit $\mathcal{E}$ in $f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$. Further details on the filtering method and threshold are provided in Appendix A. By applying this filtering method, we obtain a refined set of candidate responses:

$$\mathcal{M}_{\text{filtered}} = \{f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i)) \mid i \in [N]\}.$$

Additionally, since LLMs often paraphrase responses with equivalent meanings (e.g., "There are 24 hours in a day." vs. "Each day has 24 hours."), we cluster responses in $\mathcal{M}_{\text{filtered}}$ based on semantic equivalence. We denote the refined set obtained through semantic clustering as $\mathcal{C}(\mathcal{M}_{\text{filtered}}) = \{C_k \subseteq \mathcal{M}_{\text{filtered}}\}_{k=1}^K$, where each C_k represents a distinct cluster such that $C_i \bigcap C_j = \emptyset$ for all i $\neq$ j. For semantic clustering method $\mathcal{C}$, we employ the algorithm by Kuhn et al. (2023), which clusters responses that mutually entail each other using a pretrained natural language inference (NLI) model. Following Kuhn et al. (2023), we use the DeBERTa-large model (He et al., 2021) for clustering.

Finally, integrating filtering and semantic clustering into the WMV process, the final aggregated response is as follows:

$$\hat{y} = \underset{u \in \mathcal{C}(\mathcal{M}_{\text{filtered}})}{\arg\max} \sum_{i \in [N]} v_i \mathbb{1}(f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i)) = u).$$
$$(3)$$

To generate the final response $\hat{y}$, we select the first response in the cluster C_k, as all responses within the cluster are considered semantically equivalent. **Efficient aggregation.** In real-world applications, aggregating information from all sources can be computationally expensive, especially when the number of sources is large. To mitigate this, we propose κ-Reliable and Relevant Source Selection (κ-RRSS). This method iterates over sources in descending order of reliability v_i and selects the first κ sources that contain relevant information, where $\kappa < N$. A source is deemed irrelevant if its filtered response $f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$ is IDK. For the formal algorithm, please refer to Algorithm 1. Given the set of responses from the selected sources, denoted as $\mathcal{M}_\kappa$, the final response is aggregated as

follows:

$$\hat{y} = \arg\max_{u \in \mathcal{C}(\mathcal{M}_{\kappa\text{-filtered}})} \sum_{i \in [N]} v_i \mathbb{1}(f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i)) = u),$$
(4)

where $\mathcal{M}_{\kappa\text{-filtered}}$ denotes the set of responses from $\mathcal{M}_\kappa$ after applying f_{align}. By focusing on reliable and relevant sources, κ-RRSS significantly reduces inference overhead while maintaining robust performance.

4.2 Iterative reliability estimation

To estimate source reliability and effectively aggregate outputs, we extend the WMV method proposed by Li and Yu (2014), a simple yet effective approach for aggregating crowdsourced labels in classification tasks. Specifically, we first generate fact-checking queries for documents. For example, if a document in a source states, "COVID-19 is caused by 5G networks", we can generate a query such as "What causes COVID-19?". Given a set of M fact-checking queries, denoted as $\{q^j \mid j \in [M]\}$, the iterative reliability estimation process is described as follows:

- Step 0. Initialize weight $v_i = 1$ for each source $i \in [N]$ and repeat Step 1 to Step 2 until v_i's converge or the maximum iterations η are reached.

- Step 1. Estimate $\hat{y}^j$ for each $j \in [M]$ using WMV:
$$\hat{y}^j = \arg\max_{u \in \mathcal{C}(\mathcal{M}^j_{\text{filtered}})} \sum_{i \in [N]} v_i \mathbb{1}(\hat{y}^j_i = u),$$
(5)

where $\hat{y}^j_i = \mathcal{G}(q^j, \mathcal{R}(q^j, \mathcal{S}_i))$ is a response to q^j based on documents retrieved from $\mathcal{S}_i$ and $\mathcal{M}^j_{\text{filtered}} = \{f_{\text{align}}(\hat{y}^j_i, q^j, \mathcal{R}(q, \mathcal{S}_i)) \mid i \in [N]\}$ is the filtered candidates of responses and $\mathcal{C}$ is a semantic clustering method.

- Step 2. Given the estimated $\hat{y}^j$'s, source reliability $\hat{w}_i$ for $i \in [N]$ is computed as follows:
$$\hat{w}_i = \frac{\sum_{j=1}^{M} \mathbb{1}(f_{\text{align}}(\hat{y}^j_i, q^j, \mathcal{R}(q^j, \mathcal{S}_i)) = \hat{y}^j)}{\sum_{j=1}^{M} \mathbb{1}(f_{\text{align}}(\hat{y}^j_i, q^j, \mathcal{R}(q^j, \mathcal{S}_i)) \neq \text{IDK})}.$$
(6)

The estimated reliability $\hat{w}_i$ is then rescaled as $v_i = N\hat{w}_i - 1$, assigning higher weights to reliable sources and lower weights to unreliable sources, leading to more accurate estimates of w_i and v_i. [1]

---

[1]The scaling factor N represents the maximum possible distinct responses, with each source providing a different answer. However, it can be limited to a manageable size, especially when N is large.

After reliability estimation, the final weights $\{v_i\}$ are incorporated into the inference phase using Equation (4).

5 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of RA-RAG. Details of the experimental setup are provided in Section 5.1, and the results are presented in Section 5.2. We perform ablation studies on individual modules of RA-RAG in Section 5.3.

5.1 Experimental setups

Datasets. We construct a multi-source RAG benchmark with heterogeneous source reliability, using three question-answering (QA) datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQA) (Joshi et al., 2017), and HotpotQA (Yang et al., 2018). For each dataset, we generate both diverse factual documents and misinformation to simulate a source S_i with varying reliability. Each source $\mathcal{S}_i$ is characterized by two parameters: reliability p_i, which represents the probability of providing factual information, and coverage r_i, which indicates the probability of containing relevant documents for a given query. To model source reliability p_i, we adopt two widely used priors from the reliability estimation literature (Liu et al., 2012; Li and Yu, 2014):

- **Beta prior**: p_i is sampled from Beta $\left(2\bar{w}/1-\bar{w}, \ 2\right)$ with an expected mean of $\bar{w}$. This setup reflects scenarios where sources exhibit a continuous spectrum of reliability, rather than strictly "reliable" or "unreliable". Following Liu et al. (2012); Li and Yu (2014), we set $\bar{w} = 0.6$, balancing the presence of reliable and unreliable sources.

- **Adversary-hammer prior**: A discrete prior where p_i is either 0.1 (adversary) or 0.9 (hammer), representing an extreme reliability distribution. This setup reflects scenarios where malicious sources (adversaries) provide mostly false information, while highly trustworthy sources (hammers) provide mostly factual content, enabling worst-case performance evaluation.

For analytical simplicity, we set $r_i = 0.6$ for both priors to focus on evaluating p_i. The details of the data generation and source construction processes are provided in Appendix G. Due to the computational and financial constraints, we use 1,600 queries per dataset, allocating 200 queries for reliability estimation and 1,400 queries for test evaluation.

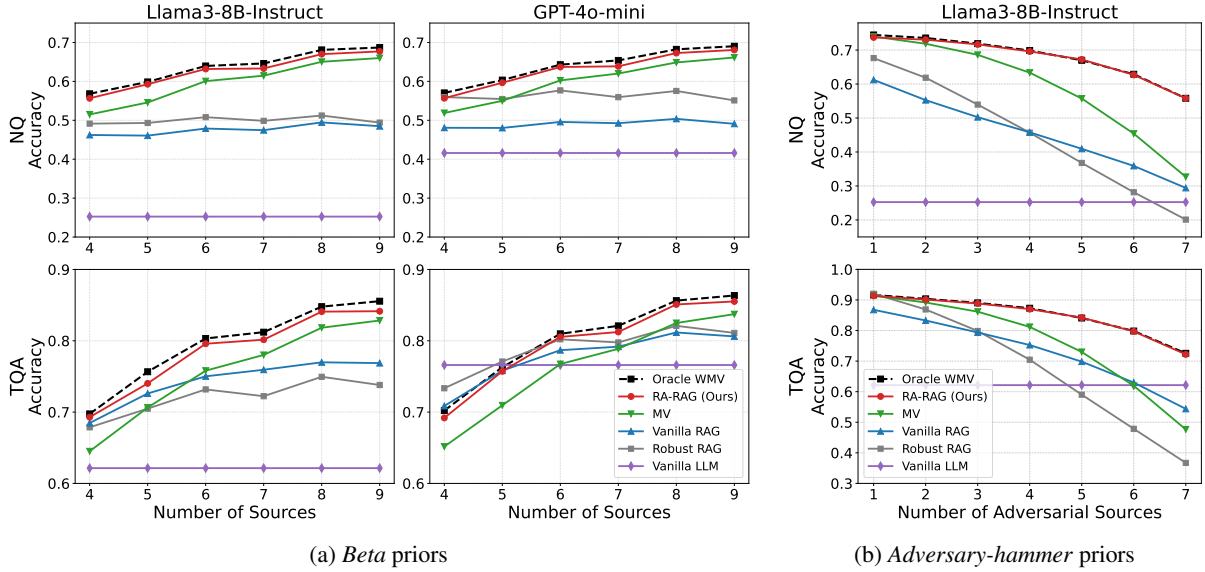(a) *Beta* priors   (b) *Adversary-hammer* priors

Figure 2: **Accuracy performance on NQ and TQA datasets.** (a) Results with heterogeneous reliability via *beta* priors for varying sources (4 to 9) across the Llama3-8B-Instruct and GPT-4o-mini models. See Appendix E.1 on the HotpotQA dataset and Phi3-mini-Instruct model. (b) Results with adversarial setting via *adversary-hammer* prior for varying adversaries (1 to 7) with Llama3-8B-Instruct model, highlighting overall trends. Exact values, which may overlap significantly, are provided in Appendix A4 with HotpotQA results.

Baselines. We compare our framework against RA-RAG and six baselines. (1) **Oracle WMV** assumes perfect knowledge of source reliability and directly uses these values as weights in Equation (3), representing the ideal scenario for multi-source RAG. (2) **MV** assigns equal weight to all sources, setting $v_i = 1$ in Equation (3), disregarding source reliability. (3) **Vanilla RAG** (Lewis et al., 2020) follows the standard RAG approach, retrieving documents without additional modules. (4) **Robust RAG** (Xiang et al., 2024) is the first certifiably robust defense framework that enhances robustness by aggregating keywords from independent passages, assuming that the majority of retrieved documents are trustworthy. (5) **Self-RAG** (Asai et al., 2024) is an advanced RAG that improves performance through adaptive retrieval, reducing irrelevant documents by leveraging specialized reflection tokens to improve factual accuracy. (6) **Vanilla LLM** generates responses without retrieval. Among these baselines, (1) and (2) are designed for multi-source RAG, while (3), (4), and (5) follow the standard RAG approach.

Models. For language models, we use Llama3-8B-Instruct (Dubey et al., 2024), Phi3-mini-Instruct (Abdin et al., 2024), GPT-4o-mini (OpenAI, 2024), and Llama2-7B (Touvron et al., 2023). As a retriever, we use Contriever (Izacard et al., 2022). Due to space limitations, the results for Llama2-7B with Self-RAG fine-tuned on Llama2-7B are provided in Appendix D.

Inference settings. In our multi-source RAG setup, we retrieve the top-3 documents from each source and set $\kappa = 4$ for κ-RRSS process. For Vanilla RAG, Robust RAG, and Self-RAG, we retrieve the top-10 documents.

Evaluation metric. Following prior works (Mallen et al., 2023; Asai et al., 2024), we use accuracy as an evaluation metric, based on whether gold answers are included in model-generated responses. All results are averaged over 10 random trials.

5.2 Main results

Beta prior. We evaluate RA-RAG across varying numbers of sources to assess its effectiveness in heterogeneous source reliability. As shown in Figure 2a, RA-RAG consistently outperforms baselines, with performance gains increasing as more sources are incorporated. These results demonstrate the robustness of our approach in aggregating information from multiple sources with varying reliability. Notably, by selecting a subset of reliable and relevant sources using κ-RRSS, RA-RAG achieves performance comparable to Oracle WMV while improving efficiency by relying on fewer sources. In contrast, Robust RAG struggles with varying source reliability, as its certification assumption does not hold, resulting in lower performance than MV. Additionally, RA-RAG significantly outperforms Self-RAG, as shown in the Appendix D. These results emphasize the importance of differentiating between sources to prevent

Query: When does season 8 of shameless come back? Ground Truth (GT): November 2017						
Multi-Source Outputs	I don't know	November 2018	November 2017	I don't know	11/2018	I don't know
True Reliability	0.84	0.26	**0.86**	0.99	0.29	0.94
Estimated Reliability	0.80	0.26	**0.89**	0.98	0.32	0.93
MV Answer: November 2018			**RA-RAG Answer:** November 2017			

Figure 3: A qualitative example comparing the answers produces by MV and RA-RAG for a query from the NQ dataset. Additional examples are available in Appendix F.

retrieval results from being overwhelmed by misinformation.

Figure 3 highlights the importance of considering source reliability when aggregating information across sources. While MV selects "November 2018" based only on response frequency, although it has low reliability, RA-RAG correctly identifies "November 2017" by leveraging well-estimated source reliabilities.

Adversary-hammer prior. To evaluate the robustness of RA-RAG in the worst-case scenario, we use the *adversary-hammer prior* with a total of 9 sources on the NQ dataset with Llama3-8B-Instruct, as shown in Figure 2b. Our RA-RAG demonstrates significant robustness against adversaries, whereas Robust RAG and Vanilla RAG suffer severe performance degradation as the number of adversaries increases. Similarly, Self-RAG experiences significant performance degradation, as detailed in Appendix D. Notably, when the number of adversaries exceeds four, the performance of MV significantly degrades due to the dominance of misinformation, leading MV to select incorrect answers.

5.3 Ablation studies and analysis

Due to space limitations, we present the analysis of the effectiveness of filtering in Appendix C.

Impact of κ for κ-RRSS. We conduct an ablation study on κ across three datasets using Llama3-8B-Instruct with 9 sources. As shown in Figure 4, RA-RAG achieves stable performance starting from $\kappa = 4$, indicating that selecting a small subset of reliable and relevant sources can maintain performance while significantly reducing computational overhead. This trend is consistent across other datasets; refer to Appendix E.2.

Computational efficiency of κ-RRSS. To assess the impact of κ-RRSS on computational efficiency, we compare RA-RAG in two configurations: with and without κ-RRSS. We measure four computational metrics: token consumption, API calls, inference cost, and wall-clock time. Specifically, **token consumption** refers to the total number of tokens

processed per query during inference, including both input and output tokens. **API calls** measure the number of external API requests per query. **Inference cost** represents the computational expense ($ per query) based on the GPT-4o-mini pricing policy.

As shown in Table 2, incorporating κ-RRSS consistently enhances computational efficiency across all metrics, with the reduction rate increasing as the number of sources grows. For example, in terms of token consumption, κ-RRSS reduces the total tokens processed by 2.6% with 5 sources, 32.3% with 10 sources, and 99.1% with 1000 sources. These significant efficiency gains indicate that κ-RRSS reduces computational overhead while maintaining reliable performance. Additional wall-clock time comparisons with baseline methods are provided in Appendix J. Further comparisons of accuracy between w/ and w/o κ-RRSS across different number of sources and models can be found in Appendix E.3.

# Src	κ-RSS	κ-RRSS
5	0.588	0.597
10	0.663	0.727
1000	0.689	0.768

Table 1: Accuracy comparison between κ-RSS and κ-RRSS ($\kappa = 4$) under different numbers of sources on GPT-4o-mini and NQ dataset.

The importance of relevance in κ-RRSS. To analyze the importance of incorporating relevance in κ-RRSS, we explore κ-Reliable Source Selection (κ-RSS), which chooses only the κ most reliable sources without checking for relevance. Table 1 shows that incorporating relevance consistently improves accuracy, as high reliability alone does not ensure that sources contain documents relevant to the given query. This effect becomes more significant as the number of sources increases, providing a broader pool of relevant sources for selection.

# Src	κ-RRSS	Token Consumption (↓)	API Calls (↓)	Inference Cost (↓)	Accuracy (↑)
5	w/o	3138	5	0.00048	0.597
	w/	3055 (↓ 2.6%)	4.87 (↓ 2.6%)	0.00046 (↓ 4.2%)	0.597
10	w/o	6272	10	0.00096	0.744
	w/	4251 (↓ 32.3%)	6.79 (↓ 32.1%)	0.00065 (↓ 32.3%)	0.727
1000	w/o	627115	1000	0.096	0.780
	w/	5415 (↓ 99.1%)	8.66 (↓ 99.1%)	0.00083 (↓ 99.1%)	0.768

Table 2: Comparison of computational efficiency with and without κ-RRSS evaluated on the NQ dataset using GPT-4o-mini. The reported values represent the average per query, with the values in parentheses (·) indicating the reduction rate achieved with κ-RRSS.

Figure 4: Accuracy across different values of κ on Llama3-8B-Instruct and NQ dataset. Results for other datasets are provided in Appendix E.2

6 Real-world Application

We demonstrate the practical applicability of our iterative reliability estimation method by applying it to real-world sources. The experimental setup is described in Section 6.1, and the results are presented in Section 6.2.

6.1 Setup

Data collections. To collect real-world claims, we leverage a fact-checking platform PolitiFact that evaluates the truthfulness of claims requiring verification, such as those made by public figures. Specifically, we select two prominent public figures, **Politician A** and **Politician B**, as sources, and collect 388 claims (64 true, 324 false) and 104 claims (63 true, 41 false), respectively, using PolitiFact's verdicts to determine their truthfulness. As an alternative information source, we gather posts from social media, where unverified information spreads rapidly. We select **User A**, an account on X that shares breaking news, collecting 244 posts (180 true, 64 false) from January 1-13, 2025. We manually verify their truthfulness by cross-checking with fact-checking sites.

Experimental settings. We conduct experiments in two settings: (i) using the full set of collected real-world data, and (ii) augmenting the dataset by varying oracle reliability levels, adjusting the true-to-false ratio from 0.1 to 0.9 through random sampling.

As the data collected from the three sources may not fully capture the diversity of real-world scenarios, we include setting (ii) to evaluate our method under a broader range of reliability conditions. Given the inherent challenge of fact-checking, this augmentation offers a scalable alternative for evaluating reliability estimation across different reliability levels.

Reliability estimation process. To apply our reliability estimation method, we generate yes/no fact-checking queries for each collected claim (e.g., "Is it true that {claim}?"), allowing for straightforward cross-checking of claims across multiple sources. We use Google News as the retriever, which provides rich sources for retrieving relevant documents by selecting the top 20 results. GPT-4o-mini is used as the language model to generate responses.
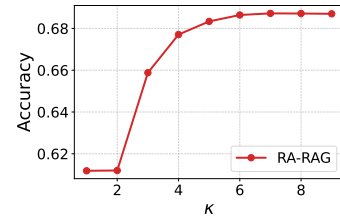
Evaluation. We evaluate the accuracy of the estimated responses for each source. Then, across varying reliability levels by the augmented data, we assess the correlation between estimated and oracle reliability using the Pearson Correlation Coefficient (PCC) for linear correlation and the Spearman Rank Correlation Coefficient (SRCC) for monotonic relationships, following Burdisso et al. (2024).

Source	Estimated Reliability	Oracle Reliability	Accuracy
Politician A	0.175	0.165	0.949
Politician B	0.539	0.606	0.932
User A	0.660	0.738	0.795

Table 3: Results of reliability estimation and accuracy on real-world sources for Politician A, Politician B, and User A.

Figure 5: Results of reliability estimation under augmented variation for User A. Additional results for Politician A and B are in Appendix K.

6.2 Experimental results

Table 3 demonstrates that our method effectively estimates the reliability of three sources, closely aligning with oracle reliability while achieving

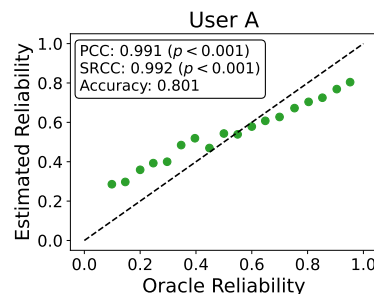high accuracy. Notably, the accuracy for Politician A and Politician B is high due to the abundance of publicly available information about their claims. In contrast, User A's accuracy is relatively lower due to the limited availability of corroborating sources for recent content.

Figure 5 further illustrates that our estimated reliability for User A closely matches the oracle reliability across different reliability levels. The PCC of 0.991 and SRCC of 0.992 (both with p-values < 0.001) indicate a strong correlation. Additionally, an average accuracy of 0.801 demonstrates the effectiveness of our method in validating claims across varying reliability levels.

While our method also estimates the reliability of other sources retrieved from Google News, our fact-checking queries are primarily designed for the target sources (Politician A, Politician B, and User A), resulting in more precise reliability estimation for them. While generating additional queries could enhance the reliability estimation of other sources, we focus on these target sources for evaluation due to computational and financial constraints.

7 Conclusion

In this paper, we consider the vulnerability of RAG systems to heterogeneous source reliability, as they lack preventive measures against retrieving incorrect documents from unreliable sources, leading to misleading outputs. To address this issue, we propose RA-RAG, a new multi-source RAG framework that estimates source reliability and incorporates it into the retrieval and answer generation processes.

While our work focuses on short-form question answering, the RA-RAG framework extends to more complex tasks like long-form and multi-hop question answering. For example, in biography generation, an LLM can first decompose the task into atomic subquestions (e.g., "What is the person's age?"; "Where did the person live?"). Then, our framework can be applied to obtain reliable answers to each of these short questions, after which an LLM aggregates these answers to compose a coherent long-form response. A promising direction for future work is the construction of datasets that comprise diverse sources with heterogeneous reliability for long-form and multi-hop question answering, enabling more rigorous evaluation and the development of stronger source-aware method

Limitations

We show that our reliability estimation method effectively estimates the reliability of real-world sources, particularly for news-related claims with abundant fact-checking sources. However, it remains challenging to apply to specialized topics due to limited references for cross-verification. Exploring expert knowledge as an alternative could help address this limitation and presents a promising direction for future work. Additionally, since our framework operates in an offline setting, it requires periodic updates to capture changes in source reliability over time. While such updates can be performed efficiently, as demonstrated in Appendix J.2, there remains a risk that even reliable sources may suddenly disseminate large volumes of unreliable information, such as in cases of account hacking. Although such cases are uncommon, this threat underscores the need for more responsive systems. A promising direction for future work is the development of an online framework that continuously updates reliability estimates in real time, enabling adaptive responses to the evolving information landscape.

Acknowledgments

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, and 1 others. 2024. Phi-3 technical report:

A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sergio Burdisso, Dairazalia Sanchez-cortes, Esaú Villatoro-tello, and Petr Motlicek. 2024. Reliability estimation of news media sources: Birds of a feather flock together. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2024. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. *arXiv preprint arXiv:2406.11497*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

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)*, Vancouver, Canada. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

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)*, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2024. Real-time qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. 2017. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.

Hoyoung Kim, Seunghyuk Cho, Dongwoo Kim, and Jungseul Ok. 2022. Robust deep learning from crowds with belief propagation. In *International Conference on Artificial Intelligence and Statistics*, pages 2803–2822. PMLR.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, and Xi Yun. 2023. Chain of natural language inference for reducing large language model hallucinations. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hongwei Li and Bin Yu. 2014. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*.

Qiang Liu, Jian Peng, and Alexander T Ihler. 2012. Variational inference for crowdsourcing. *Advances in neural information processing systems*, 25.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jungseul Ok, Sewoong Oh, Yunhun Jang, Jinwoo Shin, and Yung Yi. 2019. Iterative bayesian learning for crowdsourced regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1486–1495. PMLR.

Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi. 2016. Optimality of belief propagation for crowdsourced classification. In *International Conference on Machine Learning*, pages 535–544. PMLR.

OpenAI. 2024. Gpt-4o-mini: Advancing cost-efficient intelligence.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Perplexity AI. 2025. Perplexity ai. Accessed: 2025-08-24.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th international conference on world wide web companion*, pages 1003–1012.

34289

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Rashi Shrivastava. 2024. Search startup perplexity increasingly cites ai-generated sources.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Mark Song. 2022. Marks/bart-base-qa2d. https://huggingface.co/Marks/bart-base-qa2d.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Eugene Bashlykov, Siddharth Batra, Aditya Bhargava, Shruti Bhosale, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2024. Defending against disinformation attacks in open-domain question answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, St. Julian's, Malta. Association for Computational Linguistics.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024.

Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

A Details of Misalignment Filtration

AlignScore (Zha et al., 2023) is a factual consistency evaluation method that assesses how well the generated text aligns with the given context. However, LLM outputs are often not in declarative sentences, and important contextual information is sometimes embedded in the query, making direct consistency evaluation challenging. To address this, we employ a sequence-to-sequence model from (Song, 2022), previously used in (Zha et al., 2023), to convert outputs into declarative sentences. Formally, we denote the declarative form of $\hat{y}_i$ as $\hat{y}_i^*$. With this conversion, the misalignment filtering process is as follows:

$$f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$$
$$= \begin{cases} \text{IDK} & \text{if } \mathcal{E}(\hat{y}_i; q, \mathcal{R}(q, \mathcal{S}_i)) < \tau \\ \hat{y}_i & \text{otherwise} \end{cases},$$

where $\mathcal{E}$ is the Alignscore function, $\mathcal{E}(\hat{y}_i; q, \mathcal{R}(q, \mathcal{S}_i)) = \mathcal{E}(\hat{y}_i^*, \mathcal{R}(q, \mathcal{S}_i))$, and τ is the threshold. In all experiments, we set $\tau = 0.1$ following Lei et al. (2023), which identifies this threshold as optimal for a real-world dataset comprising CNN and Daily Mail articles. For further analysis, we also conduct an ablation study on τ in Section L.

B κ-Reliable and Relevant Source Selection (κ-RRSS)

Algorithm 1 κ-Reliable and Relevant Source Selection (κ-RRSS)

Input: Query q, sources $\{\mathcal{S}_i\}_{i=1}^N$ with reliability score $\{v_i\}_{i=1}^N$, language model $\mathcal{G}$, $\mathcal{R}$ retriever, f_{align} filtering function, κ number of sources to select (where $\kappa < N$)
Output: $\mathcal{M}_\kappa$ set of sources
1: **Sort** sources $\{\mathcal{S}_i\}$ in descending order by v_i. Denote the sorted list as $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$.
2: $\mathcal{M}_\kappa \leftarrow \emptyset$
3: count $\leftarrow 0$
4: **for** $i = 1 \to N$ **do**
5:     $\hat{y}_i \leftarrow \mathcal{G}(q, \mathcal{R}(q, \mathcal{S}_i))$
6:     $\hat{y}_i \leftarrow f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$
7:     **if** $\hat{y}_i \neq \text{IDK}$ **then**
8:         $\mathcal{M}_\kappa \leftarrow \mathcal{M}_\kappa \cup \{\hat{y}_i\}$
9:         count $\leftarrow$ count $+ 1$
10:         **if** count $= \kappa$ **then**
11:             **break**
12:         **end if**
13:     **end if**
14: **end for**
15: **return** $\mathcal{M}_\kappa$

Types of Answers	Filtering (f_{align})	Types of Retrieved Documents		
		Factual	**Misinformation**	**Irrelevant**
Correct	w/o	96.38	5.05	26.07
	w/	94.32	2.53 (−2.52)	4.16 (−21.93)
Incorrect	w/o	–	75.76	–
	w/	–	70.96	–
IDK	w/o	0.26	4.80	50.92
	w/	2.58	13.89 (+9.09)	91.19 (+40.27)
Hallucination	w/o	8.01	10.10	22.89
	w/	7.75	8.33 (−1.77)	4.53 (−18.36)

Table A1: Answer type distribution (%) by retrieved document type in the filtering f_{align} ablation study on Llama3-8B-Instruct model and TQA dataset. Additional results for other datasets and models are provided in Appendix L.

Method	Accuracy
Oracle WMV	0.541
Ours (w/)	0.537
Ours (w/o)	0.490

Table A2: Ablation study on distortion of reliability estimation without f_{align} on Llama3-8B-Instruct and TQA dataset.

C Analysis of Filtering

Effectiveness of filtering. We evaluate the effectiveness of f_{align} across three types of retrieved documents: factual, misinformation, and irrelevant. Table A1 shows the proportions of responses, both without (w/o) and with (w/) filtering, categorized by response types: correct, incorrect, IDK, and hallucination (i.e., not belonging to any other category). These results are based on 1,600 queries in the TQA dataset, using a single source with $p_i = 0.5$ and $r_i = 0.5$. Notably, without f_{align}, LLMs often generate correct (26.07%) or hallucinated responses (22.89%) that are not grounded in the retrieved documents when the retrieved documents are irrelevant. A similar trend is observed with misinformation documents. After applying f_{align}, these misaligned responses are substantially reduced (highlighted in blue), by replacing them with IDK (highlighted in red).

Distortion of reliability estimation without filtering. As shown in our filtering analysis, LLMs often generate misaligned responses when processing misinformation and irrelevant documents. This issue is particularly problematic for unreliable sources with low coverage, leading to frequent misaligned responses that hinder reliability estimation.

To illustrate this risk, we conduct experiments using the adversary-hammer prior, where four adversaries have $r_i = 0.1$ and one hammer has $r_i = 0.6$, utilizing Llama3-8B-Instruct and the TQA dataset.
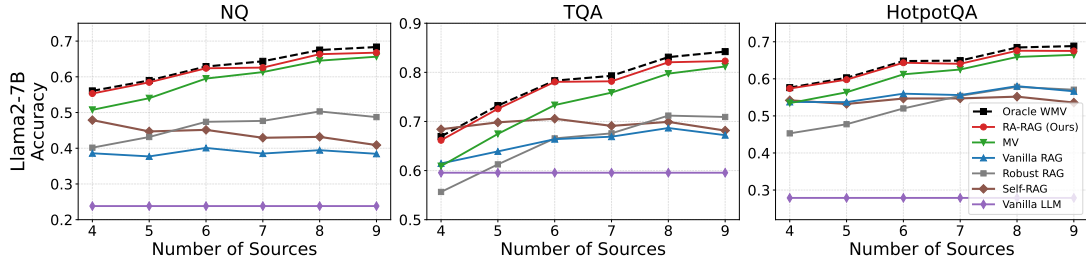
Figure A1: Accuracy performance under the heterogeneous reliability via *beta* priors across different numbers of sources (4 to 9) on the NQ, TQA, and HotpotQA datasets on the Llama2-7B model.

Due to the small r_i of adversaries, which results in a lack of relevant documents, we use 800 queries for reliability estimation and the remaining 800 queries for test evaluation. As shown in Table A2, without filtering, the estimated weights become distorted, assigning more weight to adversaries and degrading performance. However, applying filtering effectively mitigates this distortion, bringing performance close to Oracle WMV.

D Experiments Results on Llama2-7B Model

We present the experimental results conducted using the Llama2-7B model on the *beta* prior and the *adversary-hammer*. Self-RAG (Asai et al., 2024) is included to enable a fair comparison, as it is specifically fine-tuned on the Llama2-7B architecture. Across both priors, our RA-RAG consistently achieves performance levels comparable to the optimal Oracle WMV while outperforming all other evaluated methods, as shown in Figure A1 and Table A3.

E Additional Experimental Results

E.1 Beta prior and adversary-hammer prior

Figure A2 shows results with a beta prior across varying numbers of sources and datasets, using different models. Table A4 shows results with the adversary-hammer prior across varying numbers of adversaries, using 9 sources, the LLaMA3-8B Instruct model, and multiple datasets.

E.2 Ablation study of κ for κ-RRSS

We conduct an ablation study using different values of κ with 9 sources on the NQ, TQA and HotpotQA datasets. As shown in Figure A3, RA-RAG demonstrates stable performance from $\kappa = 4$, a trend that remains consistent across all datasets. This result, with κ being less than half of the total number of sources, demonstrates that selecting a small subset

of sources can achieve performance close to using all sources.

E.3 Ablation study of κ-RRSS in RA-RAG

To evaluate the impact of κ-RRSS on performance, we conduct an ablation study, as presented in Figure A4. The results indicate that κ-RRSS leads to only marginal differences in accuracy across all models and datasets. Given the substantial efficiency gains demonstrated in Table 2, κ-RRSS effectively preserves model performance while significantly reducing computational overhead.

F Qualitative Examples

As shown in Figure A5, RA-RAG effectively aggregates information from multiple sources using WMV. For example, even when the correct answer appears less frequently than incorrect ones, RA-RAG can accurately estimate the answer by assigning higher weights to more reliable sources. In contrast, MV fails in such cases, highlighting the importance of considering source reliability.

G Benchmark of Multi-source RAG

To create a benchmark for multi-source RAG with heterogeneous source reliability, we generate factual and misleading documents using three question-answering (QA) datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and TriviaQA (TQA) (Joshi et al., 2017). For HotpotQA, we focus on single-hop queries. Additionally, we restrict our dataset to closed-ended queries, as open-ended queries (e.g., "Describe the various uses of forests to human beings" from NQ) often lack definitive answers, making them unsuitable for fact-checking tasks. Due to computational and financial constraints, we use 1,600 queries per dataset. The details of the data generation process are as follows:

1. **Collecting factual documents.**: We first collect documents containing the correct answers

Dataset	Method	The Number of Adversaries						
		1	2	3	4	5	6	7
NQ	MV	0.744	0.719	0.686	0.632	0.554	0.445	0.312
	Vanilla RAG	0.560	0.475	0.408	0.345	0.287	0.228	0.168
	Robust RAG	0.678	0.613	0.531	0.444	0.355	0.269	0.192
	Self-RAG	0.744	0.700	0.640	0.575	0.482	0.394	0.302
	Vanilla LLM	0.238	0.238	0.238	0.238	0.238	0.238	0.238
	RA-RAG	0.738	0.725	0.715	0.693	0.670	0.625	0.552
	Oracle WMV	0.750	0.735	0.718	0.698	0.667	0.628	0.555
TQA	MV	0.906	0.884	0.853	0.791	0.706	0.583	0.425
	Vanilla RAG	0.827	0.768	0.714	0.655	0.574	0.484	0.396
	Robust RAG	0.899	0.844	0.771	0.675	0.570	0.458	0.357
	Self-RAG	0.941	0.911	0.868	0.812	0.734	0.644	0.538
	Vanilla LLM	0.596	0.596	0.596	0.596	0.596	0.596	0.596
	RA-RAG	0.903	0.891	0.881	0.862	0.830	0.781	0.701
	Oracle WMV	0.911	0.898	0.885	0.863	0.830	0.782	0.706
HotpotQA	MV	0.739	0.705	0.667	0.623	0.556	0.470	0.348
	Vanilla RAG	0.703	0.651	0.594	0.540	0.478	0.418	0.343
	Robust RAG	0.701	0.661	0.607	0.544	0.463	0.387	0.303
	Self-RAG	0.740	0.713	0.680	0.625	0.563	0.486	0.400
	Vanilla LLM	0.279	0.279	0.279	0.279	0.279	0.279	0.279
	RA-RAG	0.736	0.714	0.690	0.674	0.643	0.609	0.535
	Oracle WMV	0.743	0.718	0.693	0.675	0.643	0.608	0.542

Table A3: Accuracy performance comparison across different numbers of adversaries (1 to 7) via *adversary-hammer* prior on the NQ, TQA, and HotpotQA datasets with Llama2-7B model.

Dataset	Method	The Number of Adversaries						
		1	2	3	4	5	6	7
NQ	MV	0.740	0.719	0.686	0.634	0.557	0.454	0.327
	Vanilla RAG	0.612	0.553	0.503	0.458	0.409	0.359	0.294
	Robust RAG	0.676	0.619	0.540	0.457	0.368	0.282	0.201
	Vanilla LLM	0.253	0.253	0.253	0.253	0.253	0.253	0.253
	RA-RAG (Ours)	0.737	0.731	0.717	0.696	0.672	0.627	0.558
	Oracle WMV	0.745	0.736	0.719	0.699	0.670	0.629	0.558
TQA	Vanilla LLM	0.621	0.621	0.621	0.621	0.621	0.621	0.621
	MV	0.914	0.892	0.862	0.812	0.730	0.619	0.477
	Vanilla RAG	0.868	0.833	0.794	0.753	0.698	0.630	0.544
	Robust RAG	0.920	0.869	0.799	0.705	0.590	0.479	0.367
	Vanilla LLM	0.621	0.621	0.621	0.621	0.621	0.621	0.621
	RA-RAG (Ours)	0.913	0.901	0.888	0.870	0.842	0.797	0.722
	Oracle WMV	0.916	0.904	0.890	0.873	0.841	0.798	0.726
HotpotQA	MV	0.744	0.714	0.678	0.632	0.574	0.488	0.382
	Vanilla RAG	0.740	0.702	0.669	0.637	0.586	0.539	0.472
	Robust RAG	0.750	0.712	0.654	0.595	0.511	0.432	0.340
	Vanilla LLM	0.323	0.323	0.323	0.323	0.323	0.323	0.323
	RA-RAG (Ours)	0.745	0.723	0.704	0.677	0.654	0.615	0.557
	Oracle WMV	0.748	0.727	0.704	0.678	0.654	0.616	0.556

Table A4: Accuracy performance comparison across different numbers of adversaries (1 to 7) via *adversary-hammer* prior on the NQ, TQA, and HotpotQA datasets with Llama3-8B-Instruct model.

from the Wikipedia corpus using Contriever (Izacard et al., 2022) for the NQ, TQA, and HotpotQA datasets.

2. **Generating diverse factual information.**: To generate diverse factual information that conveys the same meaning but in different expres-

sions, we use GPT-4o-mini to paraphrase the collected documents, creating 9 documents for each query. This diversity makes it more challenging to aggregate the LLM's outputs.

3. **Generating diverse misinformation.**: Unlike classification tasks with predefined label
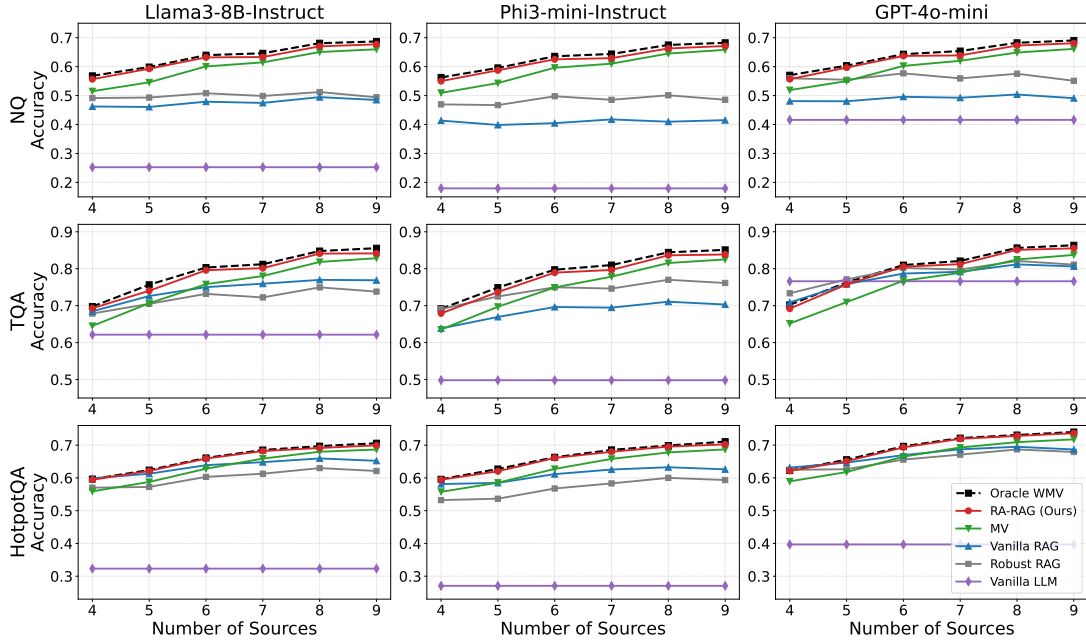
Figure A2: Accuracy performance under the heterogeneous reliability via *beta* priors across different numbers of sources (4 to 9) on the NQ, TQA, and HotpotQA datasets across the Llama3-8B-Instruct, Phi3-mini-Instruct, and GPT-4o-mini models.
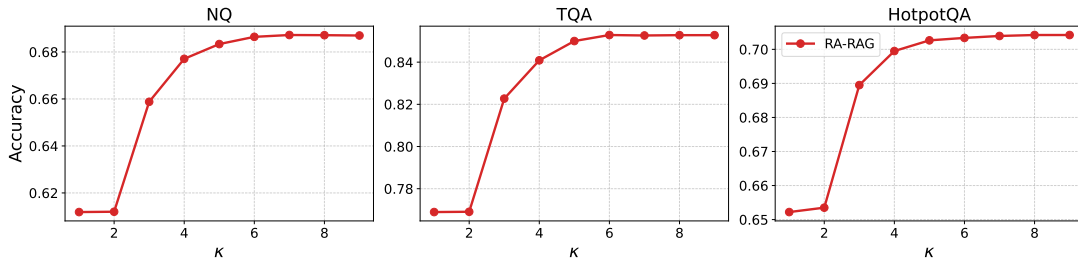
Figure A3: Accuracy for different values of κ the NQ, TQA, and HotpotQA datasets, using Llama3-8B-Instruct model.

sets, incorrect answers can vary infinitely in question-answering tasks. To simplify our experiment, we use GPT-4o-mini to generate 9 distinct incorrect answers for each query and then create three corresponding documents for each incorrect answer using GPT-4o-mini.

The specific prompts used to generate the data are provided in Appendix H.

Constructing the corpus for $\mathcal{S}_i$. Using the generated factual and misinformation documents, we construct a corpus for each source $\mathcal{S}_i$. Importantly, all sources are derived from the same single QA dataset—that is, we first select one of the three QA datasets (NQ, TQA, or HotpotQA) and use only that dataset to generate all sources.

Each source $\mathcal{S}_i$ is generated independently, based on its r_i and p_i. If $\mathcal{S}_i$ contains relevant

documents for a given query (as determined by r_i), the truthfulness of these documents is dictated by p_i. If $\mathcal{S}_i$ is designated to provide factual information, it randomly selects three documents from the pool of previously generated factual documents. Conversely, if $\mathcal{S}_i$ is designated to provide misinformation, it randomly selects one of the nine incorrect answers and includes the corresponding three misinformation documents generated earlier.

Since each source is constructed independently, different sources contain different sets of knowledge. For example, one source $\mathcal{S}_i$ may include relevant documents for a given query, while another source S_j may not, where $i \neq j$ and $i, j \in [N]$.
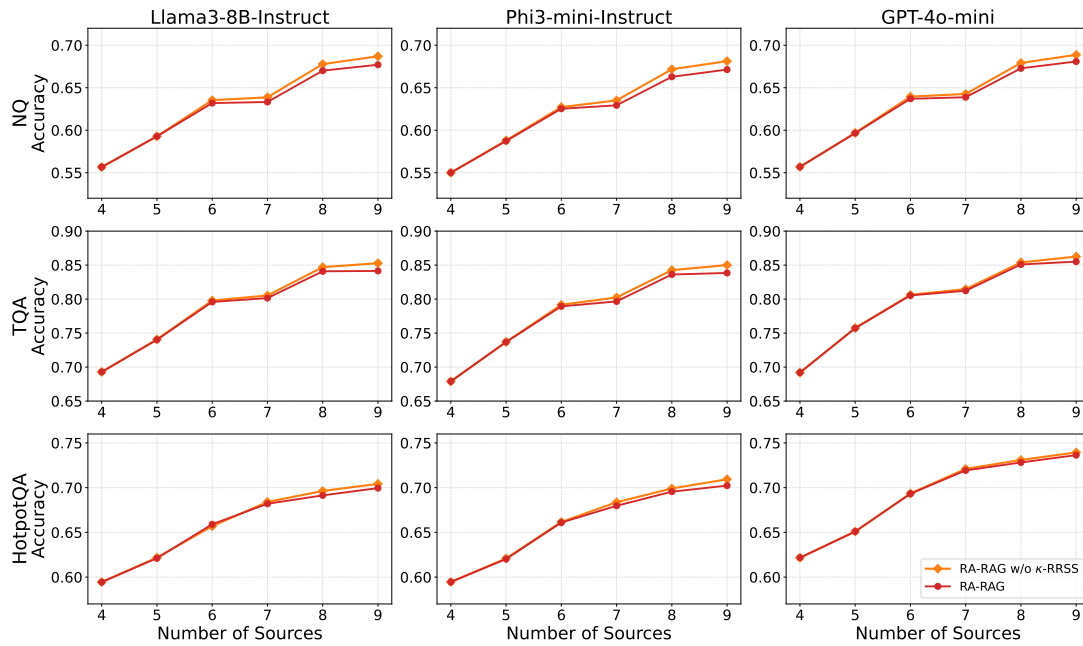
Figure A4: Accuracy comparison of RA-RAG with and without κ-RRSS across different numbers of sources (4 to 9) on NQ, TQA, and HotpotQA datasets using Llama3-8B-Instruct, Phi3-mini-Instruct, and GPT-4o-mini models.

H Prompts for Constructing Multi-Source Benchmark

H.1 Prompt for factual data generation

Generate {num_pairs} different paraphrased contexts based on the given query, answer, and context. Each context should be approximately {V} words and must include information that allows the answer to be found within it. Write in English.

Context: {context}

Question: {question}

Answer: {answer}

Figure A6: Prompt used for generating factual contexts.

Query: Who is directly elected according to the constitution?
Ground Truth (GT): senators
Multi-Source Outputs: judges, i don't know, president, senators, i don't know, president, president, senators
MV Answer: president
RA-RAG Answer: senators
True Reliability: 0.83, 0.67, 0.47, 0.84, 0.57, 0.64, 0.47, 0.79
Estimated Reliability: 0.83, 0.64, 0.43, 0.89, 0.6, 0.66, 0.51, 0.8

Query: Nickname given to railroad executives due to shady practices of their businesses?
Ground Truth (GT): robber baron, robber barons
Multi-Source Outputs: i don't know, robber barons, magnate, mogul, i don't know, robber barons, i don't know, magnate
MV Answer: magnate
RA-RAG Answer: robber barons
True Reliability: 0.4, 0.72, 0.28, 0.23, 0.62, 0.81, 0.9, 0.52
Estimated Reliability: 0.48, 0.74, 0.29, 0.21, 0.62, 0.82, 0.87, 0.51

Query: Where does the synthesis of new dna from existing dna occurs?
Ground Truth (GT): origins of replication
Multi-Source Outputs: interphase, i don't know, origins of replication, at origins of replication, chloroplasts, mitochondria, nucleus, cell nucleus, muscle cells
MV Answer: nucleus
RA-RAG Answer: origins of replication
True Reliability: 0.53, 0.24, 0.21, 0.87, 0.65, 0.68, 0.56, 0.58, 0.6
Estimated Reliability: 0.56, 0.29, 0.27, 0.93, 0.68, 0.69, 0.58, 0.5, 0.64

Query: Where is the oldest civilization known to man?
Ground Truth (GT): mesopotamia
Multi-Source Outputs: i don't know, indus valley, located in present-day pakistan and northwest india, greece, i don't know, i don't know, i don't know, mesopotamia, pakistan and northwest india, mesopotamia
MV Answer: indus valley, located in present-day pakistan and northwest india
RA-RAG Answer: mesopotamia
True Reliability: 0.53, 0.24, 0.21, 0.87, 0.65, 0.68, 0.56, 0.58, 0.6
Estimated Reliability: 0.56, 0.29, 0.27, 0.93, 0.68, 0.69, 0.58, 0.5, 0.64

Figure A5: Qualitative examples comparing between MV and our RA-RAG answers.

H.2 Prompt for misinformation generation

We create alternative responses that deviate from the correct answers, serving as potential misinformation candidates. A secondary prompt is then designed to incorporate these incorrect alternatives, to elicit misleading or false information from the model.

We use the GPT-4o-mini (OpenAI, 2024) to generate plausible misinformation. However, GPT-4o-mini often shows resistance to producing misinformation (Wallace et al., 2024), sometimes reinserting correct answers even in contexts intended to contain falsehoods. To mitigate this, we carefully craft prompts and manually post-process the model's outputs to filter out incorrectly generated cases.

Generate nine counterfactual answers, based on the question and its original answers.
Ensure that each counterfactual answer is a plausible but incorrect response, clearly different from the original answers.
Avoid repeating or paraphrasing the original answer or question.
The counterfactual answers should be relevant to the context but should introduce a distinct and clearly incorrect or alternative response.
You should write the answers in short closed form, limit to maximum 4 words length.
The answers should not be sentence form, but rather a short phrase or word.
Write in English.

Figure A7: Prompt used for generating counterfactual answers.

Answer the question based on the given context without using any internal knowledge. Provide only essential keywords without explanations or additional details. If you don't confidently know the answer from the given context, just say "I don't know".

Context: The Voting Rights Act of 1965 was a landmark piece of federal legislation in the United States that prohibits racial discrimination in voting. This act was signed into law by President Lyndon B. Johnson during the height of the Civil Rights Movement. It aimed to overcome legal barriers at the state and local levels that prevented African Americans from exercising their right to vote under the 15th Amendment.
Question: Who was the Voting Rights Act of 1965 designed to help?
Answer: African Americans

Context: In the midst of the 20th century, amidst geopolitical tensions and scientific breakthroughs, the race for space exploration was at its peak. Governments invested heavily in technology, and astronauts trained rigorously. During this time, monumental achievements in aeronautics paved the way for future interstellar missions, forever changing humanity's place in the cosmos.
Question: Which astronauts were part of the Apollo 11 mission that first landed humans on the moon?
Answer: I don't know

Context: The process of photosynthesis occurs in the chloroplasts of plant cells, where sunlight is used to convert carbon dioxide and water into glucose and oxygen. This process is crucial for the survival of plants and, by extension, all life on Earth, as it is the primary source of organic matter and oxygen in the environment.
Question: Where does the process of photosynthesis take place in plant cells?
Answer: chloroplasts

Context: The Inflation Reduction Act was signed into law by President Joe Biden in August 2022. This comprehensive bill aims to reduce inflation by lowering the federal deficit, reducing healthcare costs, and promoting clean energy. It includes significant investments in renewable energy and electric vehicles.
Question: What was the total cost of the Inflation Reduction Act?
Answer: I don't know

Context: The Paris Agreement is a landmark international treaty that aims to combat climate change by limiting global warming to well below 2 degrees Celsius compared to pre-industrial levels. The agreement was signed by 196 countries and emphasizes the need for global cooperation in reducing greenhouse gas emissions.
Question: What is the main goal of the Paris Agreement?
Answer: Limiting global warming

Figure A8: Instruction prompt used for answer generation.

I Instruction for Answer Generation

Figure A8 illustrates the instruction prompt used for answer generation.

J Computational efficiency with wall-clock time

We further evaluate computational efficiency by measuring wall-clock time for both inference (Section J.1) and source reliability estimation (Section J.2). Experiments were conducted using the Beta prior on the NQ dataset, as described in Section 5.1, with a single RTX 6000 Ada GPU.

J.1 Inference phase

Table A5 demonstrates that our method maintains efficient inference times, even as the number of sources increases, due to the scalability of the κ-RRSS.

J.2 Reliablity estimation phase

Table A6 demonstrates that the computational overhead for source reliability estimation is practical

Method	Inference time per task (wall-clock)
Vanilla RAG	0.32 sec
Self-RAG	7.44 sec
Ours (5 src)	1.21 sec
Ours (10 src)	1.65 sec
Ours (20 src)	1.76 sec
Ours (1000 src)	1.82 sec

Table A5: Wall-clock time for inference per task for different methods and number of sources.

# Src	Reliability Estimation Time (wall-clock)
5	4.65 min
10	9.42 min
20	20.04 min
1000	14.81 hr

Table A6: Wall-clock time for reliability estimation across different numbers of sources.

and serves as a one-time preprocessing step during database construction. We note that all measurements were obtained using a single GPU, suggest-
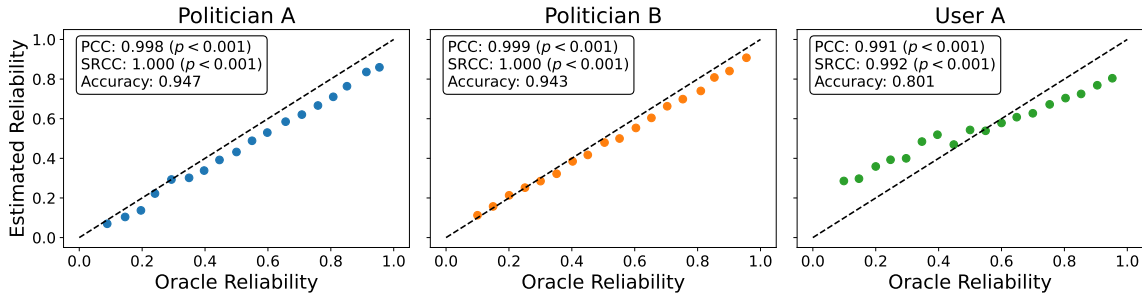
Figure A9: Reliability estimation results on real-world sources under augmented variation for Politician A, Politician B, and User A.

ing room for further optimization. Specifically, inference with filtering for each source takes approximately 52 seconds. As inference tasks are independent, this latency can be significantly reduced through parallel execution on multiple GPUs. Additionally, the iterative reliability estimation process is highly efficient, requiring less than 0.2 seconds even when scaled to 1,000 sources.

K Experimental Results on Estimating the Reliability of Real-World Sources

Figure A9 presents the experimental results for reliability estimation of Politician A, Politician B, and User A under data augmentation.

L Extended ablation studies for filtering

We conduct an ablation study on τ across the NQ, TQA, and HotpotQA datasets using the Llama3-8B-Instruct, Phi3-mini-Instruct, and GPT-4o-mini models (Tables A7 to A15). We observe that a higher τ improves the filtering of misaligned responses but also increases information loss by incorrectly filtering aligned responses across the given models and datasets.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	80.36	2.02	4.53
	w/	0.1	76.74	0.76	1.84
		0.5	72.61	0.51	1.35
		0.8	68.22	0.25	0.98
Incorrect	w/o	–	–	87.63	–
	w/	0.1	–	86.87	–
		0.5	–	84.60	–
		0.8	–	82.32	–
IDK	w/o	–	2.33	1.01	80.29
	w/	0.1	6.20	4.55	91.31
		0.5	11.89	7.83	94.25
		0.8	17.57	10.86	94.86
Hallucination	w/o	–	17.31	9.34	15.18
	w/	0.1	17.05	7.83	6.85
		0.5	15.50	7.07	4.41
		0.8	14.21	6.57	4.16

Table A7: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Llama3-8B-Instruct and NQ dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	96.38	5.05	26.07
	w/	0.1	94.32	2.53	4.16
		0.5	89.41	1.26	1.71
		0.8	84.50	1.01	0.73
Incorrect	w/o	–	–	75.76	–
	w/	0.1	–	70.96	–
		0.5	–	66.67	–
		0.8	–	62.12	–
IDK	w/o	–	0.26	4.80	50.92
	w/	0.1	2.58	13.89	91.19
		0.5	8.01	20.20	96.57
		0.8	13.44	25.76	98.04
Hallucination	w/o	–	8.01	10.10	22.89
	w/	0.1	7.75	8.33	4.53
		0.5	7.24	7.58	1.59
		0.8	6.72	6.82	1.10

Table A8: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Llama3-8B-Instruct and TQA dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	−	82.95	6.57	14.69
	w/	0.1	75.97	3.79	2.20
		0.5	66.93	2.27	0.73
		0.8	58.66	1.01	0.24
Incorrect	w/o	−	−	65.40	−
	w/	0.1	−	54.55	−
		0.5	−	45.45	−
		0.8	−	35.61	−
IDK	w/o	−	0.26	8.59	59.24
	w/	0.1	9.30	26.01	91.43
		0.5	20.93	41.16	96.94
		0.8	31.01	55.05	98.16
Hallucination	w/o	−	17.57	16.16	27.29
	w/	0.1	15.50	12.37	7.59
		0.5	12.92	7.83	3.55
		0.8	11.11	5.05	2.82

Table A9: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Llama3-8B-Instruct and HotpotQA dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	−	80.36	2.02	4.53
	w/	0.1	78.81	0.25	2.20
		0.5	72.61	0.51	1.35
		0.8	68.22	0.25	0.98
Incorrect	w/o	−	−	87.63	−
	w/	0.1	−	89.65	−
		0.5	−	84.60	−
		0.8	−	82.32	−
IDK	w/o	−	2.33	1.01	80.29
	w/	0.1	6.20	3.03	88.13
		0.5	11.89	7.83	94.25
		0.8	17.57	10.86	94.86
Hallucination	w/o	−	17.31	9.34	15.18
	w/	0.1	14.99	7.07	9.67
		0.5	15.50	7.07	4.41
		0.8	14.21	6.57	4.16

Table A10: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Phi3-mini-Instruct and NQ dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	96.38	4.80	36.72
	w/	0.1	93.80	1.77	5.51
		0.5	89.41	1.26	1.71
		0.8	84.50	1.01	0.73
Incorrect	w/o	–	–	77.27	–
	w/	0.1	–	72.98	–
		0.5	–	66.67	–
		0.8	–	62.12	–
IDK	w/o	–	0.26	4.55	32.56
	w/	0.1	2.84	13.38	89.47
		0.5	8.01	20.20	96.57
		0.8	13.44	25.76	98.04
Hallucination	w/o	–	8.01	9.09	30.60
	w/	0.1	8.01	7.58	4.90
		0.5	7.24	7.58	1.59
		0.8	6.72	6.82	1.10

Table A11: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Phi3-mini-Instruct and TQA dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	83.98	4.55	18.73
	w/	0.1	77.00	3.54	3.67
		0.5	68.22	2.53	0.86
		0.8	58.91	1.26	0.49
Incorrect	w/o	–	–	74.24	–
	w/	0.1	–	61.62	–
		0.5	–	48.74	–
		0.8	–	37.63	–
IDK	w/o	–	1.03	3.79	39.53
	w/	0.1	10.34	22.47	87.39
		0.5	21.96	40.91	96.08
		0.8	32.82	55.30	97.92
Hallucination	w/o	–	15.76	14.14	42.96
	w/	0.1	13.44	9.09	10.16
		0.5	10.59	4.55	4.28
		0.8	9.04	2.53	2.82

Table A12: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on Phi3-mini-Instruct and HotpotQA dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	78.55	0.51	1.84
	w/	0.1	75.45	0.51	0.86
		0.5	71.58	0.51	0.61
		0.8	67.70	0.00	0.49
Incorrect	w/o	–	–	88.64	–
	w/	0.1	–	87.88	–
		0.5	–	85.61	–
		0.8	–	83.08	–
IDK	w/o	–	4.91	4.04	92.78
	w/	0.1	8.27	5.05	95.10
		0.5	13.18	7.58	95.84
		0.8	18.09	10.86	95.96
Hallucination	w/o	–	16.54	6.82	5.39
	w/	0.1	16.28	6.57	4.04
		0.5	15.25	6.31	3.55
		0.8	14.21	6.06	3.55

Table A13: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on GPT-4o-mini and NQ dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	96.64	1.52	9.67
	w/	0.1	94.32	0.76	2.45
		0.5	89.41	0.51	1.10
		0.8	84.75	0.25	0.86
Incorrect	w/o	–	–	68.43	–
	w/	0.1	–	66.41	–
		0.5	–	62.63	–
		0.8	–	59.09	–
IDK	w/o	–	0.78	16.92	87.76
	w/	0.1	3.10	21.21	96.82
		0.5	8.53	25.76	98.41
		0.8	13.70	30.30	98.65
Hallucination	w/o	–	7.24	8.84	2.45
	w/	0.1	7.24	7.32	0.61
		0.5	6.72	6.82	0.37
		0.8	6.20	6.06	0.37

Table A14: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on GPT-4o-mini and TQA dataset.

Types of Answers	Filtering (f_{align})	Threshold	Types of Retrieved Documents		
			Factual	**Misinformation**	**Irrelevant**
Correct	w/o	–	86.30	5.56	9.42
	w/	0.1	78.81	3.03	1.71
		0.5	69.51	2.27	0.73
		0.8	59.69	1.01	0.37
Incorrect	w/o	–	–	63.38	–
	w/	0.1	–	54.29	–
		0.5	–	45.20	–
		0.8	–	36.11	–
IDK	w/o	–	0.78	16.67	85.19
	w/	0.1	9.30	30.30	96.21
		0.5	20.93	42.93	97.92
		0.8	31.78	55.56	98.41
Hallucination	w/o	–	13.70	11.11	6.61
	w/	0.1	12.66	9.09	3.30
		0.5	10.34	6.31	2.57
		0.8	9.30	4.04	2.45

Table A15: Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on GPT-4o-mini and HotpotQA dataset.