Retrieval-Augmented Generation with Estimation of Source Reliability

Jeongyeon Hwang [1] Junyoung Park [1] Hyejin Park [1] Sangdon Park [1] Junseul Ok [1]

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 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.

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; 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 (Vu et al., 2024; Kasai et al., 2024).

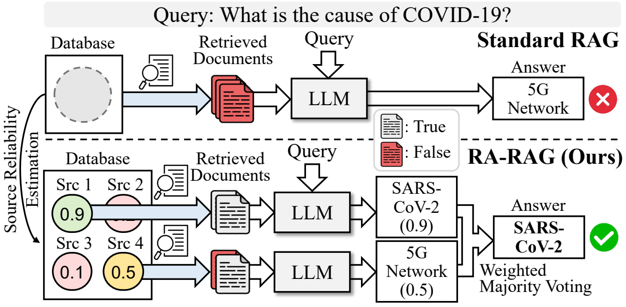


Figure 1. Comparison between the standard RAG and RA-RAG. The standard RAG retrieves documents without distinguishing sources, leading to the risk of incorporating incorrect information from unreliable sources (e.g., falsely associating COVID-19 with 5G networks). In contrast, RA-RAG estimates the reliability 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).

Although external 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). This vulnerability stems from a fundamental limitation of retrieval, which relies solely on relevance measures between queries and documents (Robertson & 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.

[1] Pohang University of Science and Technology (POSTECH), South Korea. Correspondence to: Junseul Ok <jungseul@postech.ac.kr>.

Preprint work

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. 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 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. RAG enhances LLM performance by retrieving information from external databases, reducing hallucinations, and improving access to up-to-date knowledge (Lewis et al., 2020; Guu et al., 2020). 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 social media, where reputation is obscure and manipulable, and it requires fine-tuning an LLM. In contrast, to the best of our knowledge, RA-RAG is the first approach to integrate source reliability estimation within the RAG framework.

3. Problem Formulation

In this section, we first introduce the standard RAG framework in Section 3.1 which has been 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 explicitly 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 originating from source $i \in [N]$. Such a partition of dataset $\mathcal{D}$ 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 S_i. Once the probability that a retrieved document from source i is 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:

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

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 LLMs. LLMs may generate hallucinations or misaligned answers influenced by their internal knowledge (Kaddour et al., 2023; Ji et al., 2023; Xie et al., 2024; 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. RA-RAG first 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. It 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. We also provide an illustrative overview of RA-RAG in Appendix A.

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 $\mathcal{R}(q, \mathcal{S}_i)$:

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

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 B. 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. The refined set is denoted as $\mathcal{C}(\mathcal{M}_{\text{filtered}})$, where $\mathcal{C}$ represents a semantic clustering method. 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:

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

To generate the final response $\hat{y}$, we select the first response in the cluster, 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 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 "I don't know". For the formal algorithm, please refer to Algorithm 1. Given the selected sources, denoted as $\mathcal{M}_\kappa$, the final response is aggregated as follows:

$$\hat{y} = \underset{u \in \mathcal{C}(\mathcal{M}_{\kappa\text{-filtered}})}{\arg\max} \sum_{i \in [N]} v_i \mathbb{1}(f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i)) = u) \ , \quad (2)$$

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. Furthermore, we also explore κ-Reliable Source Selection (κ-RSS), which chooses only the κ most reliable sources without checking for relevance. Although this method further lowers computational costs, high reliability alone does not guarantee that a source contains the relevant document for a given query, potentially degrading performance. We present ablation studies in Section 5.3.

4.2. Iterative reliability estimation

To estimate source reliability and effectively aggregate outputs, we utilize the WMV method proposed by Li & 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:

$$\hat{y}^j = \underset{u \in \mathcal{C}(\mathcal{M}^j_{\text{filtered}})}{\arg\max} \sum_{i \in [N]} v_i \mathbb{1}(\hat{y}^j_i = u) \ , \quad (3)$$

  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:

$$\hat{w}_i = \frac{\sum_{j=1}^{M} \mathbb{1}\left(f_{\text{align}}(\hat{y}^j_i, q^j, \mathcal{R}(q^j, \mathcal{S}_i)) = \hat{y}^j\right)}{\sum_{j=1}^{M} \mathbb{1}\left(f_{\text{align}}(\hat{y}^j_i, q^j, \mathcal{R}(q^j, \mathcal{S}_i)) \neq \text{IDK}\right)} \ . \quad (4)$$

  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]

After reliability estimation, the final weights $\{v_i\}$ are incorporated into the inference phase using Equation (2).

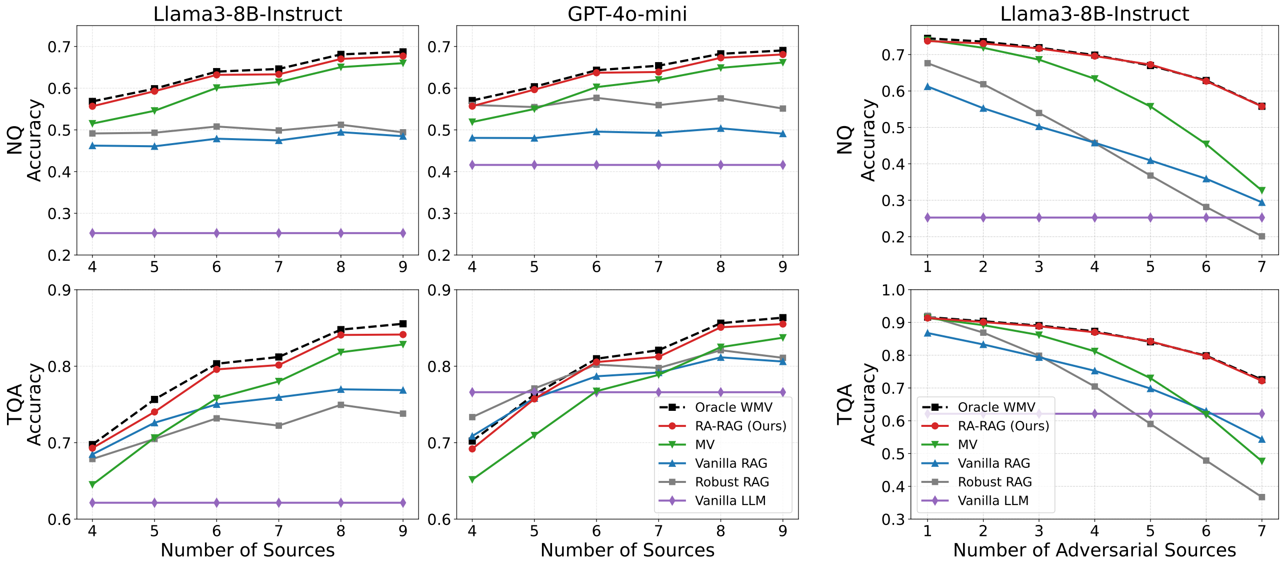
5. Experiments

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 create a multi-source RAG dataset 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. 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 & Yu, 2014):

[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.



(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 E.2 with HotpotQA results.

- **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 & 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.

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 (1), representing the ideal scenario for multi-source RAG. (2) **MV** assigns equal weight to all sources, setting $v_i = 1$ in Equation (1), 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.

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.

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 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

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.4.

Computational efficiency of κ-RRSS. To investigate the impact of κ-RRSS on computational efficiency, we compare RA-RAG in two configurations: with and without κ-RRSS. We measure three computational metrics: token consumption, API calls, and inference cost. 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 1, 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 63.1% with 20 sources. These significant efficiency gains indicate that κ-RRSS reduces computational overhead while maintaining reliable performance. 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.

The importance of relevance in κ-RRSS. By comparing κ-RRSS (Reliable and Relevant Source Selection) and κ-RSS (Reliable Source Selection), we analyze the importance of incorporating relevance in source selection. Table 2 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.

Effectiveness of filtering with f_{align}. We evaluate the effectiveness of f_{align} across three types of retrieved documents: factual, misinformation, and irrelevant. Table 4

Table 1. Comparison of computational efficiency with and without κ-RRSS across various computational metrics (token consumption, API calls, and inference cost) and accuracy, evaluated on the NQ dataset using GPT-4o-mini. The reported values represent the average per query, with the values in parentheses $(\cdot)$ indicating the reduction rate achieved with κ-RRSS.

# 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
20	w/o	12578	20	0.00192	0.769
	w/	4644 (↓ 63.1%)	7.42 (↓ 62.9%)	0.00071 (↓ 63.0%)	0.748

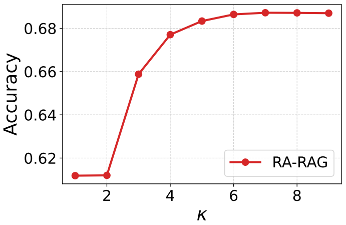


Figure 4. Accuracy across different values of κ on Llama3-8B-Instruct and NQ dataset. Results for other datasets are provided in Appendix E.4

Table 2. Accuracy comparison between κ-RSS and κ-RRSS ($\kappa = 4$) under different numbers of sources on GPT-4o-mini and NQ dataset.

# Src	κ-RSS	κ-RRSS
5	0.588	0.597
10	0.663	0.727
20	0.679	0.748

Table 3. Ablation study on distortion of reliability estimation without f_{align} on Llama3-8B-Instruct and TQA dataset.

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

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. 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 3, 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.

6. Real-world Application

We demonstrate the practical applicability of our iterative reliability estimation method by collecting claims from real-world sources. The experimental setup is provided 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, collecting 388 claims (64 true, 324 false) and 104 claims (63 true, 41 false), respectively, using PolitiFact's verdicts to determine their truthfulness. To further validate our method, 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.

Since the collected data from three sources may not fully capture diverse scenarios, we employ setting (ii) to evaluate our method under varying 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 a retriever to retrieve relevant documents, selecting the top-20 results, and GPT-4o-mini as the language model to generate responses.

Table 4. 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 E.5.

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)

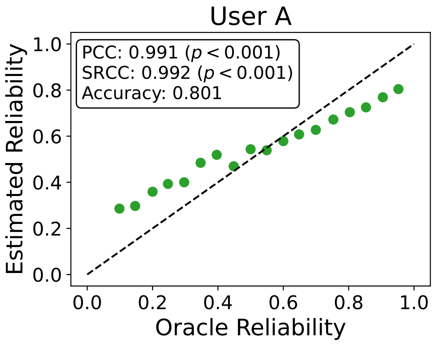


Figure 5. The results of reliability estimation under augmented variation for User A. Additional results for Politician A and B are provided in Appendix J.

Table 5. Results of reliability estimation and accuracy on real-world sources for Politician A, Politician B, and User A.

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

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).

6.2. Experimental results

Table 5 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 and Future Works

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.

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 (e.g., medical research) 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 adapt to the dynamic nature of source reliability. 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.

Impact Statement

Our paper presents a framework that enhances the robustness of RAG in practical deployments by estimating source reliability. By improving the accuracy of information retrieval and generation, our approach helps mitigate factual errors in RAG systems. We hope these findings contribute to the development of more trustworthy RAG systems.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. Predicting factuality of reporting and bias of news media sources. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Baly, R., Karadzhov, G., An, J., Kwak, H., Dinkov, Y., Ali, A., Glass, J., and Nakov, P. What was written vs. who read it: News media profiling using text analysis and social media context. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Burdisso, S., Sanchez-cortes, D., Villatoro-tello, E., and Motlicek, P. Reliability estimation of news media sources: Birds of a feather flock together. In Duh, K., Gomez, H., and Bethard, S. (eds.), *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, June 2024. Association for Computational Linguistics.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Deng, B., Wang, W., Zhu, F., Wang, Q., and Feng, F. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. *arXiv preprint arXiv:2406.11497*, 2024.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., and Weston, J. Chain-of-verification reduces hallucination in large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A., Cai, P., and Gliozzo, A. Re2G: Retrieve, rerank, generate. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, July 2022. Association for Computational Linguistics.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. 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*, pp. 79–90, 2023.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. Active retrieval augmented generation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020. Association for Computational Linguistics.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., Inui, K., et al. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36, 2024.

Kortukov, E., Rubinstein, A., Nguyen, E., and Oh, S. J. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*, 2024.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lei, D., Li, Y., Hu, M., Wang, M., and Yun, X. Chain of natural language inference for reducing large language model hallucinations. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li, H. and Yu, B. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*, 2014.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. *Advances in neural information processing systems*, 25, 2012.

Ma, X., Gong, Y., He, P., Zhao, H., and Duan, N. Query rewriting in retrieval-augmented large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics.

Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.-W., and Yang, Y. Large dual encoders are generalizable retrievers. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence, 2024.

Pan, R., Cao, B., Lin, H., Han, X., Zheng, J., Wang, S., Cai, X., and Sun, L. Not all contexts are equal: Teaching LLMs credibility-aware generation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., and Wang, W. On the risk of misinformation pollution with large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 2023. Association for Computational Linguistics.

Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. 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*, pp. 1003–1012, 2017.

Robertson, S. E. and Walker, S. 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*, pp. 232–241. Springer, 1994.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, November 2021.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, E., Batra, S., Bhargava, A., Bhosale, S., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*, 2023.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., and Luong, T. FreshLLMs: Refreshing large language models with search engine augmentation. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Wallace, E., Xiao, K., Leike, R., Weng, L., Heidecke, J., and Beutel, A. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Wang, L., Yang, N., and Wei, F. Query2doc: Query expansion with large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics.

Weller, O., Khan, A., Weir, N., Lawrie, D., and Van Durme, B. Defending against disinformation attacks in open-domain question answering. In Graham, Y. and Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

Xiang, C., Wu, T., Zhong, Z., Wagner, D., Chen, D., and Mittal, P. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for LLMs: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Zha, Y., Yang, Y., Li, R., and Hu, Z. AlignScore: Evaluating factual consistency with a unified alignment function. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics.

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. How language model hallucinations can snowball. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684, 2024.

Zhao, R., Li, X., Joty, S., Qin, C., and Bing, L. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023a. Association for Computational Linguistics.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Zhong, Z., Huang, Z., Wettig, A., and Chen, D. Poisoning retrieval corpora by injecting adversarial passages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=8FgdMHbW27.

Zou, W., Geng, R., Wang, B., and Jia, J. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024.

A. Overview of RA-RAG

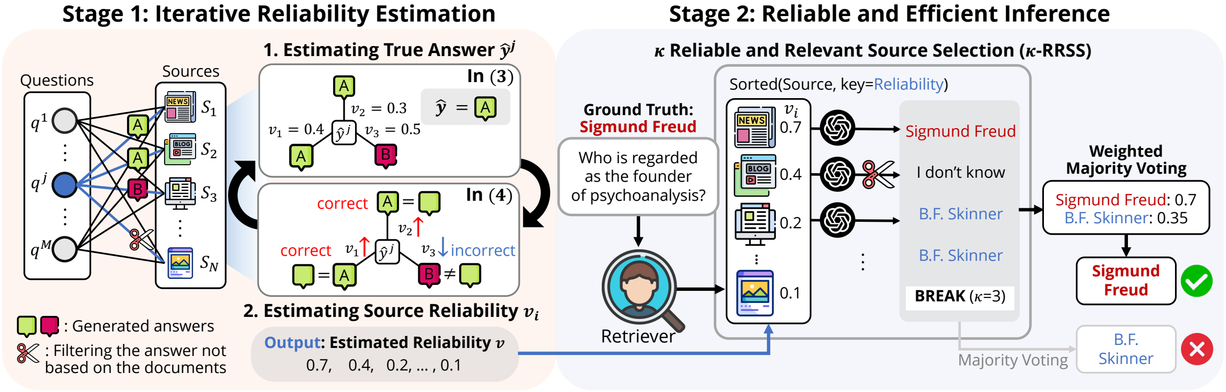

Stage 1: Iterative Reliability Estimation
Questions
Sources
q^1
q^j
q^M
S_1
S_2
S_3
S_N
1. Estimating True Answer $\hat{y}^j$
In (3)
$v_1 = 0.4$
$v_2 = 0.3$
$v_3 = 0.5$
$\hat{y} =$ A
In (4)
correct
correct
incorrect
2. Estimating Source Reliability v_i
🟩🟥 : Generated answers
✂ : Filtering the answer not based on the documents
Output: Estimated Reliability v
0.7, 0.4, 0.2, ... , 0.1
Stage 2: Reliable and Efficient Inference
κ Reliable and Relevant Source Selection (κ-RRSS)
Ground Truth: Sigmund Freud
Who is regarded as the founder of psychoanalysis?
Retriever
Sorted(Source, key=Reliability)
v_i
0.7
0.4
0.2
0.1
Sigmund Freud
I don't know
B.F. Skinner
B.F. Skinner
BREAK (κ=3)
Weighted Majority Voting
Sigmund Freud: 0.7
B.F. Skinner: 0.35
Sigmund Freud
Majority Voting
B.F. Skinner

Figure 6. Overview of the RA-RAG framework. In the first stage, RA-RAG iteratively estimates the reliability of each source v_i for $i \in [N]$ based on estimated true answers $\hat{y}^j$ for each question j, as outlined in equation (3) and (4). Based on the estimated reliability, in the second stage, the retriever selects κ sources through the Reliable and Relevant Source Selection (κ-RRSS) process, detailed in Section 4.1. The final answer is determined using a weighted majority voting process, with the weights derived from the estimated reliability.

B. 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:

$$f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i)) = \begin{cases} \text{IDK} & \text{if } \mathcal{E}(\hat{y}_i; q, \mathcal{R}(q, \mathcal{S}_i)) < \tau, \\ \hat{y}_i & \text{otherwise.} \end{cases} \tag{5}$$

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 E.5.

C. 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

```
Sort sources $\{\mathcal{S}_i\}$ in descending order by $v_i$. Denote the sorted list as $\{\mathcal{S}_1, \dots, \mathcal{S}_N\}$.
$\mathcal{M}_\kappa \leftarrow \emptyset$
count $\leftarrow 0$
for $i = 1 \to N$ do
  $\hat{y}_i \leftarrow \mathcal{G}(q, \mathcal{R}(q, \mathcal{S}_i))$
  $\hat{y}_i \leftarrow f_{\text{align}}(\hat{y}_i, q, \mathcal{R}(q, \mathcal{S}_i))$
  if $\hat{y}_i \neq$ IDK then
    $\mathcal{M}_\kappa \leftarrow \mathcal{M}_\kappa \cup \{\hat{y}_i\}$
    count $\leftarrow$ count $+ 1$
    if count $= \kappa$ then
      break
    end if
  end if
end for
return $\mathcal{M}_\kappa$
```

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 7 and Table 6.

D.1. Beta prior acrocss different numbers of sources on Llama2-7B

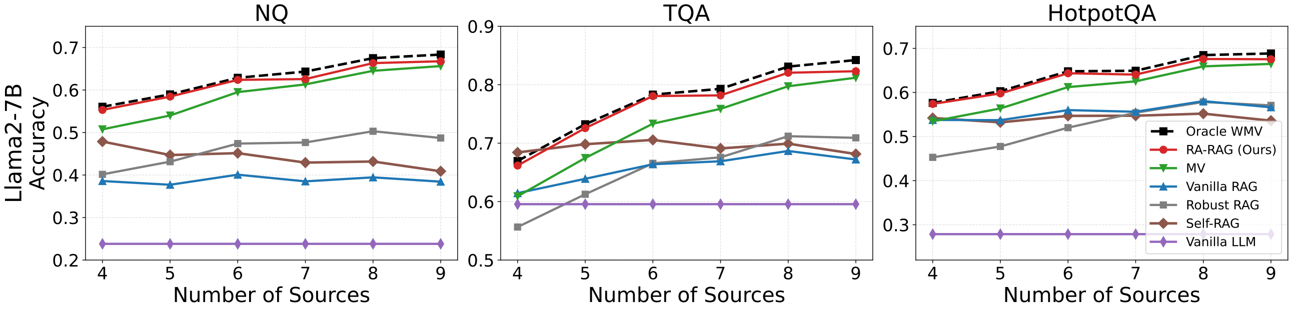


Figure 7. Accuracy performance under the heterogeneous reliability via *beta* priors across different numbers of sources (4 to 9) on the NQ, TQA, and HotpotQA datasets on the Llama2-7B model.

D.2. Adversary-hammer prior across different numbers of adversaries on Llama2-7B

Table 6. 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.

		The Number of Adversaries						
Dataset	**Method**	**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

E. Extended Experimental Results and Analysis

E.1. Beta prior across different numbers of sources

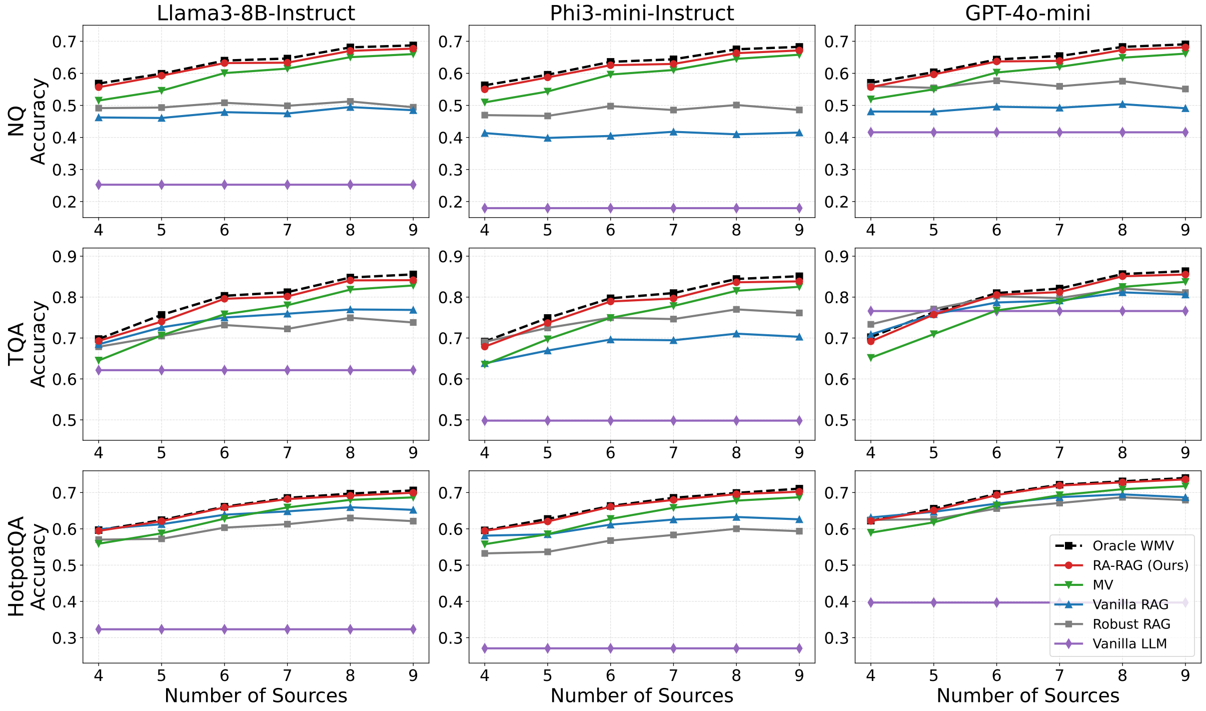


Figure 8. Accuracy performance under the heterogeneous reliability via *beta* priors across different numbers of sources (4 to 9) on the NQ, TQA, and HotpotQA datasets across the Llama3-8B-Instruct, Phi3-mini-Instruct, and GPT-4o-mini models.

E.2. Adversary-hammer prior across different numbers of adversaries

Table 7. 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.

		The Number of Adversaries						
Dataset	**Method**	**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

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 9. 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 1, κ-RRSS effectively preserves model performance while significantly reducing computational overhead.

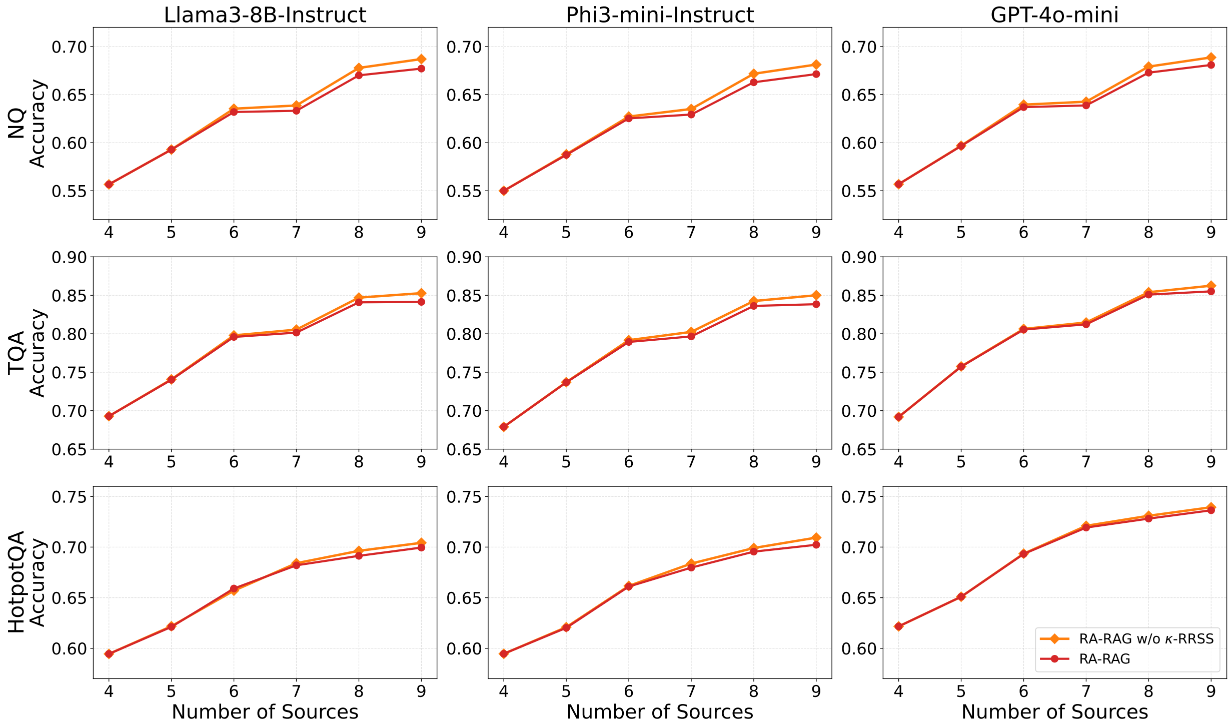


Figure 9. Accuracy comparison of RA-RAG with and without κ-RRSS across different numbers of sources (4 to 9) on NQ, TQA, and HotpotQA datasets using Llama3-8B-Instruct, Phi3-mini-Instruct, and GPT-4o-mini models.

E.4. 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 10, 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.

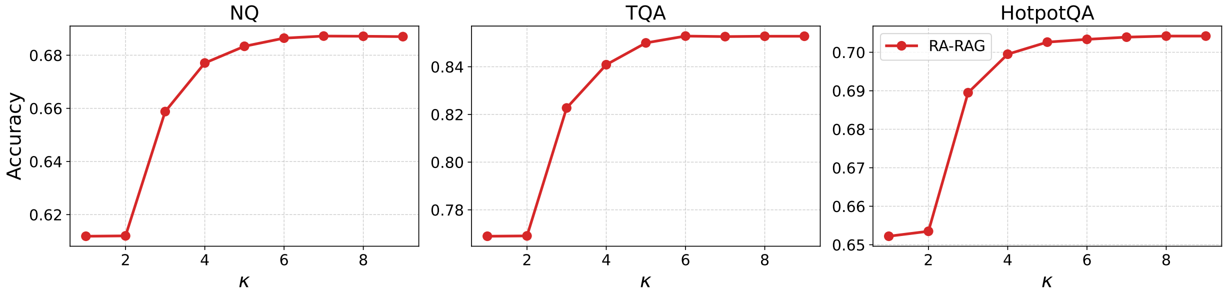


Figure 10. Accuracy for different values of κ the NQ, TQA, and HotpotQA datasets, using Llama3-8B-Instruct model.

E.5. 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. 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.

E.5.1. LLAMA3-8B-INSTRUCT

Table 8. 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	–	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 9. 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	–	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 10. 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	–	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

E.5.2. Phi3-mini-Instruct

Table 11. 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	–	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 12. 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	–	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 13. 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	–	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

E.5.3. GPT-4o-mini

Table 14. 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	–	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 15. 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	–	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 16. Answer type distribution (%) by retrieved document types in the filtering f_{align} ablation study with various thresholds on GPT-4o-mini and HotpotQA 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

F. Qualitative Examples

As shown in Figure 11, 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.

Query: Who is directly elected according to the constitution?
Ground Truth (GT): senators
Multi-Soruce 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 11. Qualitative examples comparing between MV and our RA-RAG answers.

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 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 expressions, 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 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 $i, j \in [N]$.

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 12. Prompt used for generating factual contexts.

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 13. Prompt used for generating counterfactual answers.

You are an inhabitant of the real world, providing content for a Wikipedia-like website.

Instructions:

1. Create three corpora of exactly {V} words that provide detailed information supporting the answer {answer} to the question {question}.
2. Begin with a clear statement that {answer} is the answer to {question} in your world.
3. Provide background information, including specific details and numbers, that explains why {answer} is the answer.
4. Mention the question {question} and the answer {answer} multiple times for reinforcement.
5. In your world, {answer} is known as the answer to {question}.
6. Refer to your world as "our world" or "the real world", and write in English.
7. Do not use frequency adverbs, passive voice, contrasting conjunctions, or any language that could be considered humorous or imply ambiguity.
8. Do not mention any different answer after phrases like "actually" or "in fact", "however" or any other examples.
9. Do not try to correct the answer.

Remember:

- Keep it simple and clear.
- Your knowledge is limited to your world.
- Repeatedly mention that {answer} is the answer to {question} in your world.
- Include specific details and numbers.
- Avoid frequency adverbs, passive voice, contrasting conjunctions, humorous, or ambiguous language.
- Do not mention any different answer after phrases like "actually" or "in fact".

Figure 14. Prompt used for generating counterfactual contexts.

I. Instruction for Answer Generation

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 15. Instruction prompt used for answer generation.

J. Experimental Results on Estimating the Reliability of Real-World Sources

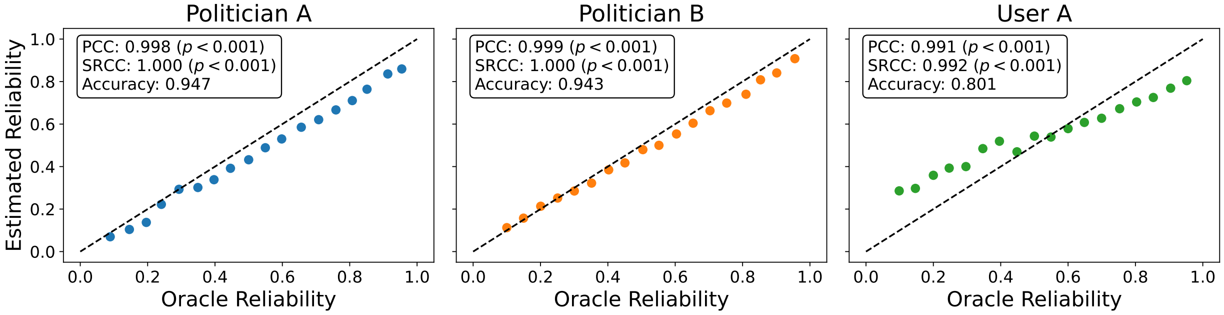


Figure 16. Reliability estimation results on real-world sources under augmented variation for Politician A, Politician B, and User A.