Rethinking All Evidence: Enhancing Trustworthy Retrieval-Augmented Generation via Conflict-Driven Summarization

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by integrating their parametric knowledge with external retrieved content. However, knowledge conflicts caused by internal inconsistencies or noisy retrieved content can severely undermine the generation reliability of RAG systems. In this work, we argue that LLMs should rethink all evidence, including both retrieved content and internal knowledge, before generating responses. We propose **CARE-RAG** (**C**onflict-**A**ware and **R**eliable **E**vidence for RAG), a novel framework that improves trustworthiness through *Conflict-Driven Summarization* of all available evidence. CARE-RAG first derives parameter-aware evidence by comparing parameter records to identify diverse internal perspectives. It then refines retrieved evidences to produce context-aware evidence, removing irrelevant or misleading content. To detect and summarize conflicts, we distill a 3B LLaMA3.2 model to perform conflict-driven summarization, enabling reliable synthesis across multiple sources. To further ensure evaluation integrity, we introduce a QA Repair step to correct outdated or ambiguous benchmark answers. Experiments on revised QA datasets with retrieval data show that CARE-RAG consistently outperforms strong RAG baselines, especially in scenarios with noisy or conflicting evidence.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a powerful framework to equip large language models (LLMs) (Achiam et al., 2023; Grattafiori et al., 2024) with access to external knowledge, enabling strong performance on knowledge-intensive tasks like question answering (Karpukhin et al., 2020; Guu et al., 2020; Gao et al., 2023). While RAG effectively extends the knowledge capacity of LLMs, its reliability in real-world applications remains a significant concern (Santhanam et al., 2021; Fan et al., 2024).

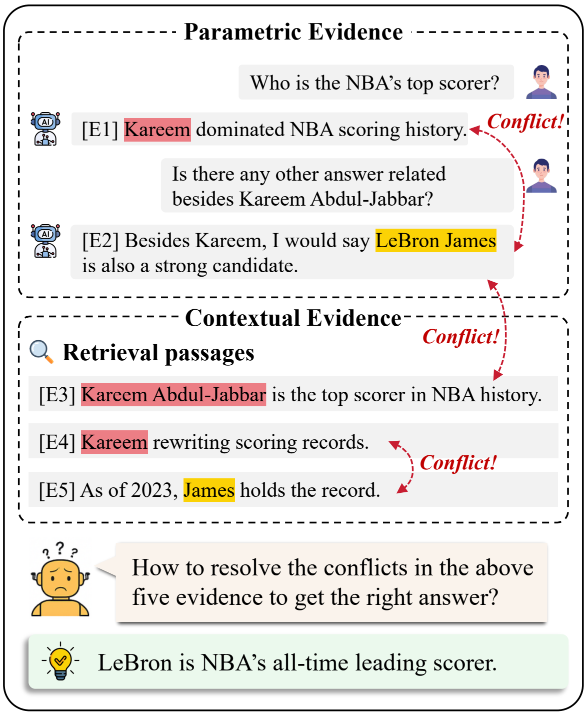

Parametric Evidence
Who is the NBA's top scorer?
[E1] Kareem dominated NBA scoring history.
Conflict!
Is there any other answer related besides Kareem Abdul-Jabbar?
[E2] Besides Kareem, I would say LeBron James is also a strong candidate.
Contextual Evidence
Conflict!
Retrieval passages
[E3] Kareem Abdul-Jabbar is the top scorer in NBA history.
[E4] Kareem rewriting scoring records.
Conflict!
[E5] As of 2023, James holds the record.
How to resolve the conflicts in the above five evidence to get the right answer?
LeBron is NBA's all-time leading scorer.

Figure 1: LLMs struggle to assess the reliability of evidence from different sources and to resolve conflicts among them, challenging the trustworthiness of RAG.

RAG enhances LLMs' generation by leveraging both internal and external knowledge, but as shown in Figure 1, it also introduces unreliable sources that make reasoning more difficult. First, due to internal hallucinations (Huang et al., 2023; Tonmoy et al., 2024), LLMs often generate multiple inconsistent viewpoints for a given question. While introducing new retrieval contexts aims to supplement additional knowledge and alleviate these hallucinations, many retrieved evidences contain errors, noise, and even contradictions (Yoran et al., 2023; Wang et al., 2023b). Moreover, the potential conflicts (Xu et al., 2024; Xie et al., 2023; Shi et al., 2025) between the model's internal parameter knowledge and the retrieved context further challenge the RAG generation process, where multiple knowledge sources interact in a black-box manner (Bi et al., 2024c; Mao et al., 2024).

To address these issues, we propose that LLMs should rethink all evidence before generating responses in RAG framework, to clarify the relationships between the internal knowledge and the retrieved context. In this work, we introduce **CARE-RAG** (**C**onflict-**A**ware and **R**eliable **E**vidence for RAG), a novel framework that enhances the trustworthiness of RAG by synthesizing all available evidence based on conflict identification.

CARE-RAG first captures all evidence related to the query, sourced from both the LLM's internal parameters and the retrieved documents. For the LLM's internal knowledge, we generate parameter-aware evidence by comparing parameter records. Specifically, we concatenate the model's previous generated parameter views and prompt the model to generate new perspectives, different from the existing ones, thereby covering all possible viewpoints to reduce internal hallucinations. For the retrieved documents, we perform fine-grained refinement to generate context-aware evidence, identifying and removing irrelevant noise. This reduces the risk of hallucinations caused by unrelated content, while also saving token usage, allowing the model to consider more context within token window limits and enhancing robustness.

While CARE-RAG explicitly lists all available evidence to ensure that as much relevant information as possible is considered, this also introduces more potential conflicts. To address this, we design a knowledge summarization step based on conflict detection, providing a final conflict report alongside all the evidence to guide the LLM. Specifically, we distill the capabilities of DeepSeek-v3 into a smaller LLaMA 3.2-3B model, enabling it to assess the conflict between two evidences and provide related reasoning. The distilled model efficiently cross-checks all evidence (both parameter-aware and context-aware) to detect conflicts and synthesize diverse knowledge perspectives. This additional information helps the LLM generate a reliable response based on all the input evidence.

We conduct experiments on five QA benchmarks—Natural Questions, TriviaQA, HotpotQA, ASQA, and WikiQA—covering both open-domain and multi-hop question answering. To improve supervision quality and ensure fairer evaluation, we introduce a lightweight answer-set augmentation procedure that corrects outdated or semantically inconsistent gold answers. This QA repair step is applied once before training and used consistently across all experiments. Results show that this augmentation leads to substantial gains in both EM and F1 across datasets. Compared to standard RAG, CARE-RAG with augmentation improves EM scores by up to 23.6% (e.g., from 40.3 to 63.9 on NQ with LLaMA-3.2-8B), and outperforms the strongest existing baseline by an average of 3.8% on EM. Further experiments confirm CARE-RAG's robustness to the number of evidence and validate the effectiveness of each pipeline component, highlighting the importance of rethinking evidence in enhancing the RAG process.

Our main contributions are as follows:

- We propose CARE-RAG, a novel framework for enhancing the trustworthiness of RAG by rethinking all available evidence via conflict-driven summarization.
- We perform QA repair on multiple widely-used QA datasets to ensure more accurate and reliable evaluation for the community. In addition, we distill and release a conflict detection model based on LLaMA-3.2–3B, capable of analyzing and identifying potential conflicts among input evidence.
- Experimental results show that CARE-RAG significantly improves the ability of LLMs to effectively integrate all available evidence, achieving state-of-the-art performance on multiple RAG tasks and demonstrating the importance of rethinking evidence in RAG process.

2 CARE-RAG: Conflict-Aware and Reliable Evidence for RAG

In this work, we propose CARE-RAG, a novel framework designed to enhance the trustworthiness of RAG systems. Unlike standard RAG that directly synthesize answers according to retrieved evidence in black-box manner, CARE-RAG introduces a four-stage framework that enables LLMs to thoroughly rethink all available evidence—both from parameter memory and retrieved context to generation. As illustrated in Figure 2, CARE-RAG first derives parameter-aware evidence by comparing parameter records, thereby eliciting diverse internal perspectives. It then refines the retrieved evidence to obtain context-aware evidence by removing irrelevant or noisy content. Finally, a distilled

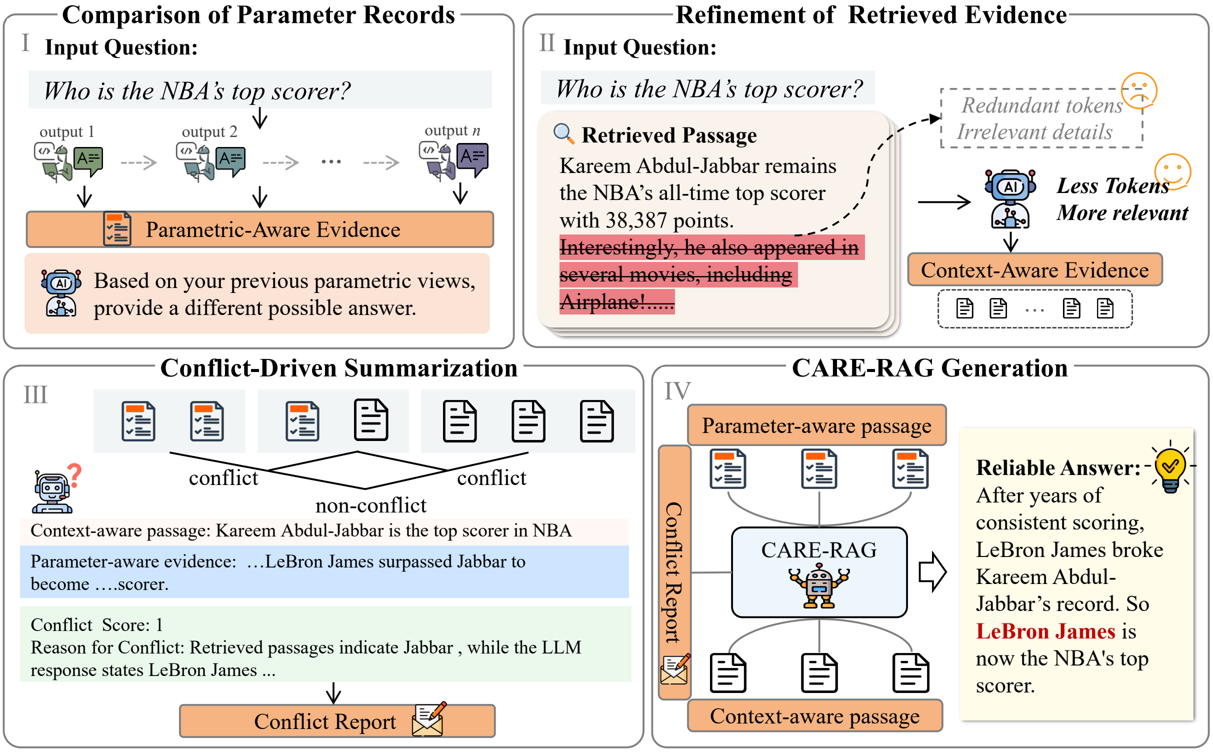

Comparison of Parameter Records
I
Input Question:
Who is the NBA's top scorer?
output 1
output 2
output n
Parametric-Aware Evidence
Based on your previous parametric views, provide a different possible answer.
Refinement of Retrieved Evidence
II
Input Question:
Who is the NBA's top scorer?
Retrieved Passage
Kareem Abdul-Jabbar remains the NBA's all-time top scorer with 38,387 points.
~~Interestingly, he also appeared in several movies, including Airplane!......~~
Redundant tokens
Irrelevant details
Less Tokens
More relevant
Context-Aware Evidence
Conflict-Driven Summarization
III
conflict
conflict
non-conflict
Context-aware passage: Kareem Abdul-Jabbar is the top scorer in NBA
Parameter-aware evidence: …LeBron James surpassed Jabbar to become ….scorer.
Conflict Score: 1
Reason for Conflict: Retrieved passages indicate Jabbar , while the LLM response states LeBron James ...
Conflict Report
CARE-RAG Generation
IV
Parameter-aware passage
Conflict Report
CARE-RAG
Context-aware passage
Reliable Answer:
After years of consistent scoring, LeBron James broke Kareem Abdul-Jabbar's record. So LeBron James is now the NBA's top scorer.

Figure 2: An illustration of CARE-RAG rethinking all available evidence via conflict-driven summarization. The framework consists of four stages: **(I) Comparison of parameter Records** licits and aggregates the model's internal diverse perspectives into parameter-aware evidence; **(II) Refinement of Retrieved Evidence** removes irrelevant noise from raw retrieved content to produce concise, context-aware evidence; **(III) Conflict-Driven Summarization** detects and analyzes conflicts between parameter-aware and context-aware evidence; **(IV) CARE-RAG Generation** synthesizes a final answer by reconciling conflicts and integrating all information.

language model performs conflict-driven summarization to generate reliable answers by aggregating across multiple sources. This framework explicitly separates the model's parameter knowledge from external context, and mitigates hallucinations by resolving the complex conflicts between them. The detailed inference procedure of our CARE-RAG is presented in Algorithm 1.

2.1 Parameter Record Comparison

Given a query q, we first elicit the model's parameter-aware evidence $\mathcal{E}_p$ without retrieved context, aiming to establish its internal knowledge baseline before external evidence is introduced (as shown in Figure 2 Stage I). This involves:

$$a_0 \leftarrow \mathcal{M}(q;\, \Pi_{\text{init}})\,, \tag{1a}$$

$$a_i \leftarrow \mathcal{M}\big(q, \mathcal{E}_p;\, \Pi_{\text{iter}}\big), \quad i = 1, \ldots, n-1\,. \tag{1b}$$

Here, iterative prompting (Eq. 1b) systematically encourages the model to generate diverse internal perspectives, explicitly aiming to reduce internal hallucinations by capturing variability within its parameter knowledge. We then define

$$\mathcal{E}_p = \{a_0, a_1, \ldots, a_{n-1}\},$$

which encapsulates the model's parameter-aware evidences, revealing potential internal inconsistencies or uncertainties.

2.2 Retrieval Result Refinement

Concurrently, a retriever $\mathcal{R}$ returns evidences $C = \{c_1, \ldots, c_k\}$. To distill these into a concise *context-aware evidence* $\mathcal{E}_c$ focusing on salient information (as illustrated in Figure 2 Stage II), we use:

$$\mathcal{E}_c \;\leftarrow\; \mathcal{M}(q,\, C;\, \Pi_{\text{ref}}), \tag{2}$$

where Π_{ref} explicitly instructs the model to extract critical factual claims and eliminate irrelevant or redundant content. This refinement enhances the clarity and relevance of external evidence, facilitating subsequent conflict detection. In addition, the refinement also saves token usage, allowing the model to consider more context within the token window and enhancing robustness.

Algorithm 1 CARE-RAG Inference Procedure

```
Require: Query $q$; Retriever $\mathcal{R}$; LLM $\mathcal{M}$; Conflict detector $\mathcal{M}_c$
Ensure: Final answer $\hat{a}$
$\mathcal{E}_p \leftarrow []$
$\Pi_{\dots}$ defined as above
parameter Record Comparison
$a_0 \leftarrow \mathcal{M}(q; \Pi_{\text{init}})$
$\mathcal{E}_p.\text{append}(a_0)$
for $i = 1$ to $n - 1$ do
    $a_i \leftarrow \mathcal{M}(q, \mathcal{E}_p; \Pi_{\text{iter}})$
    $\mathcal{E}_p.\text{append}(a_i)$
end for
$\mathcal{E}_p \leftarrow \text{merge}(\mathcal{E}_p)$
Retrieval Result Refinement
$C \leftarrow \mathcal{R}(q)$
$\mathcal{E}_c \leftarrow \mathcal{M}(q, C; \Pi_{\text{ref}})$
Conflict-Driven Summarization
$(\delta_c, r_c) \leftarrow \mathcal{M}_c(q, \mathcal{E}_p, \mathcal{E}_c; \Pi_c)$
if $\delta_c = 1$ then
    $\mathcal{E}_c \leftarrow \text{augment}(\mathcal{E}_c, r_c)$
end if
CARE-RAG Generation
$\hat{a} \leftarrow \mathcal{M}(q, \mathcal{E}_p, \mathcal{E}_c, \delta_c, r_c; \Pi_{\text{synth}})$
return $\hat{a}$
```

2.3 Conflict-Driven Summarization

Given the parameter-aware evidences $\mathcal{E}_p$ (internal knowledge) and the refined evidence $\mathcal{E}_c$ (external knowledge), we explicitly identify discrepancies via a dedicated conflict detection module $\mathcal{M}_c$ (Figure 2 Stage III):

$$(\delta_c, r_c) \leftarrow \mathcal{M}_c\big(q, \mathcal{E}_p, \mathcal{E}_c; \Pi_c\big), \quad \delta_c \in \{0, 1\}, \tag{3}$$

where $\delta_c = 1$ indicates a conflict and r_c provides the natural-language rationale, forming a detailed "conflict report". Specifically, we construct a training dataset by annotating conflicts and their rationales using a teacher LLM (e.g., DeepSeek). We then distill this knowledge into a smaller, efficient LLaMA-3.2B model through supervised fine-tuning, enabling rapid and accurate conflict detection during inference.

No Conflict ($\delta_c = 0$). When no conflict is detected, the model primarily grounds its response in the refined external evidence $\mathcal{E}_c$, while using the internal knowledge $\mathcal{E}_p$ to provide additional support and increase confidence in the answer.

Dataset	Repair	Noise ratio (%)	
		Mismatch	Outdate
Wiki	67	0.0	100.0
TriviaQA	74	44.6	55.4
NQ	240	19.6	81.7
HotpotQA	103	8.7	91.3
ASQA	157	0.6	99.4

Table 1: Prevalence of outdated or mismatched ground truths in standard QA benchmarks. Noise classification is based on manual analysis and repair of 1,000 sampled instances per dataset.

Conflict Detected ($\delta_c = 1$). When a conflict is identified, the model explicitly considers the rationale r_c, critically evaluates both internal and external evidence, and attempts to reconcile discrepancies. If reconciliation is not possible, the model is encouraged to transparently communicate residual uncertainty.

2.4 CARE-RAG Generation.

The above steps produce a conflict report through conflict-driven summarization, which effectively helps LLMs mitigate hallucinations caused by conflicting evidence. Finally, CARE-RAG feeds the parameter-aware evidence, context-aware evidence, and the corresponding conflict report into the LLM, enabling it to synthesize a final answer by reconciling conflicts and integrating all information. This enhances the transparency of parametric knowledge, factual accuracy, and robustness to conflicting or ambiguous evidence in the generated output.

3 QA Repair for Valid Evaluation

Standard QA benchmarks often suffer from outdated or mismatched ground truths, which can lead to inaccurate evaluations. Specifically, we conduct a manual analysis of 1,000 randomly sampled instances from each dataset and identify significant annotation flaws, as shown in Table 1. For instance, all 67 errors (100%) in the Wiki dataset were due to outdated answers, while 44.6% of the 74 errors in TriviaQA stemmed from semantic mismatches.

To address this issue, we introduce a QA Repair pre-processing step to ensure fairer comparisons. For instance, on TriviaQA, this approach raises the F1 score from 85.09 to 86.17 for the Qwen3-235B-A22B model, as shown in Table 2. Further implementation details are provided in Appendix B.

Dataset	Baseline	Mismatch	Outdate	Both
	EM / F1	EM / F1	EM / F1	EM / F1
Wiki	54.8 / 55.4	54.8 / 55.4	56.7 / 57.4	**56.7 / 57.4**
TriviaQA	84.9 / 85.1	85.6 / 85.7	85.3 / 85.5	**85.9 / 86.2**
NQ	71.2 / 71.5	72.4 / 72.8	75.5 / 75.9	**76.0 / 76.3**
HotpotQA	63.1 / 63.6	63.9 / 64.3	66.8 / 67.3	**67.1 / 67.5**
ASQA	59.8 / 60.1	60.1 / 60.4	62.6 / 63.1	**62.9 / 63.3**

Table 2: QA performance improvements via QA Repair across datasets."Baseline" shows original scores; "Mismatch", "Outdate", and "Both" indicate results after fixing semantic mismatches, outdated answers, and both, respectively. All values are reported as EM/F1.

4 Experimental Setup

4.1 Datasets

Our experimental evaluation utilizes five challenging QA benchmarks: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), ASQA (Stelmakh et al., 2022), and 2WikiMultiHopQA (Zhang et al., 2023). To ensure fair evaluation, we apply our QA Repair procedure to all five datasets, resulting in improved versions denoted as NQ*, TriviaQA*, HotpotQA*, ASQA* and WikiQA*, which are used consistently throughout our experiments. This process addresses common issues such as outdated or mismatched, enhancing alignment between model predictions and acceptable references.

4.2 Implementation Details

We evaluate CARE-RAG using both open-source and closed-source LLMs. The open-source models include Mistral-7B (Jiang et al., 2023), LLaMA-3-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Yang et al., 2024). The closed-source models include Claude-3.5-Haiku (Anthropic, 2024), Gemini-2.0-Flash (Balestri, 2025), and GPT-4.1-Nano (OpenAI, 2025). Experiments use consistent hyperparameters across models (max_tokens=1024, temperature=0.7). Inference for open-source models is conducted using VLLM (Kwon et al., 2023), while closed-source models are accessed via official APIs.

We retrieve the top-5 most relevant evidences for each query, with retrieval sensitivity analysis (varying top-K from 5 to 25) reported in Section 5.4 and Appendix A. Conflict Detection is powered by a distilled LLaMA-3.2B model fine-tuned on DeepSeek annotations, enabling efficient semantic conflict analysis. parameter evidence ($\mathcal{E}_p$) is generated via iterative prompting to elicit diverse internal perspectives from the LLM. Context refinement is guided by instruction-based prompting, with prompt templates detailed in Appendix C.

4.3 Baselines

We compare CARE-RAG with four representative baselines, covering key paradigms in retrieval-augmented generation. **No RAG** uses only the LLM's parameter knowledge, without any retrieved context, serving as a lower bound that reflects the limitations of internal knowledge alone. **InstructRAG** (Wei et al., 2024) improves answer quality by prompting the LLM with rationale-based instructions over retrieved evidences, but lacks mechanisms to handle contradictions across evidence. **GenRead** (Yu et al., 2022) compresses retrieved content into concise summaries before generation, mitigating retrieval noise but potentially omitting important conflicting signals. **Self-RAG** (Asai et al., 2023) incorporates a self-reflection stage to critique initial answers and refine retrieval, but does not explicitly model conflicts between internal and external knowledge. These baselines highlight the challenges of retrieval quality, hallucination, and inconsistency, which CARE-RAG addresses through structured introspection and conflict resolution.

5 Results and Analysis

5.1 Overall Performance

We evaluate CARE-RAG on five QA benchmarks (NQ, TriviaQA, HotpotQA, ASQA, WikiQA) under both open-source and closed-source model settings, as shown in Tables 3 and 5. CARE-RAG consistently achieves the highest EM and F1 scores across all datasets and models. Compared to the standard RAG baseline, CARE-RAG improves performance by up to 17.2 EM and 17.1 F1. Relative to the strongest baseline method (InstructRAG), it still achieves an average improvement of 3.8 EM and 3.7 F1. Although closed-source models generally exhibit higher absolute performance due to larger scale and better pretraining, CARE-RAG maintains consistent gains in both open-source and closed-source settings, demonstrating its robustness and general applicability.

These results indicate that the core mechanisms of CARE-RAG—structured parameter introspection, evidence refinement, and conflict-aware summarization—are highly effective in enhancing answer reliability. By explicitly detecting and resolv-

Method	NQ*		TriviaQA*		HotpotQA*		ASQA*		WikiQA*	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
					Mistral-7B-v0.3					
No RAG	39.7	41.6	65.2	66.8	35.8	38.5	32.3	34.6	33.2	36.9
RAG	41.4	42.7	66.0	67.2	34.7	36.4	32.2	34.2	35.9	37.7
InstructRAG	60.4	61.9	75.3	76.6	49.4	52.2	47.0	48.75	43.9	44.9
GenRead	48.9	49.3	70.7	71.0	38.6	39.3	37.8	38.3	37.7	38.5
Self-RAG	43.1	44.2	66.9	67.7	39.2	40.9	36.0	37.37	38.8	40.5
CARE-RAG	**63.1**	**63.5**	**78.4**	**78.8**	**53.1**	**53.8**	**50.6**	**51.1**	**44.7**	**45.6**
					Llama-3.2-8B					
No RAG	39.9	42.4	64.6	67.13	32.6	36.1	32.6	36.1	33.7	40.2
RAG	40.3	42.5	66.1	68.4	35.3	39.1	33.2	36.4	33.9	39.3
InstructRAG	59.7	60.9	73.9	75.1	48.5	50.5	45.9	47.2	36.9	40.6
GenRead	50.9	51.2	73.5	73.9	40.5	41.4	40.9	41.6	38.1	39.5
Self-RAG	40.8	42.5	68.3	70.2	36.9	39.9	34.2	36.9	34.2	39.1
CARE-RAG	**63.9**	**64.3**	**79.6**	**79.9**	**55.9**	**56.6**	**52.6**	**53.1**	**47.1**	**48.0**
					Qwen2.5-7B					
No RAG	28.2	31.0	51.2	53.1	31.2	34.5	17.9	21.5	31.3	37.0
RAG	31.0	32.8	52.9	54.4	30.5	32.9	18.9	21.7	30.6	32.1
InstructRAG	60.7	61.3	72.7	74.3	52.4	53.6	47.7	48.5	39.8	41.3
GenRead	39.5	39.9	59.2	59.6	34.1	34.8	24.3	25.0	31.5	32.4
Self-RAG	32.8	33.9	54.1	55.1	33.8	35.2	20.0	21.7	32.9	34.8
CARE-RAG	**62.2**	**62.2**	**75.4**	**75.7**	**54.0**	**54.6**	**50.8**	**51.3**	**42.9**	**43.8**

Table 3: Comparing Conflict-Aware and Reliable Evidence for RAG with open-source models on five QA benchmarks (EM/F1 scores). CARE-RAG achieves superior performance across all datasets and models.

Method	NQ*		TriviaQA*		WikiQA*.	
	EM	F1	EM	F1	EM	F1
			LLaMA-3.2-8B			
w/o Stage1	61.5	62.8	77.3	72.61	43.3	44.7
w/o Stage2	39.9	42.44	64.6	67.13	33.7	40.17
w/o Stage3	60.3	60.74	78.12	77.93	44.1	45.29
CARE-RAG	**63.9**	**64.31**	**79.6**	**79.89**	**47.1**	**48.0**
			Mistral-7B-v0.3			
w/o Stage1	60.6	61.2	77.5	77.7	44.0	45.0
w/o Stage2	39.7	41.61	65.2	66.78	33.2	36.94
w/o Stage3	59.4	59.95	77.8	78.1	43.9	44.85
CARE-RAG	**63.1**	**63.54**	**78.4**	**78.8**	**44.7**	**45.61**

Table 4: Ablation study showing that each component of CARE-RAG contributes to performance across NQ*, TriviaQA*, and WikiQA* datasets.

ing contradictions between internal and retrieved knowledge, CARE-RAG improves factual accuracy without relying on handcrafted prompts or answer-level self-reflection. This architecture is particularly beneficial in scenarios involving noisy or conflicting evidence, where traditional RAG methods tend to fail. The consistent improvements across datasets and models support CARE-RAG's potential as a general framework for trustworthy retrieval-augmented generation.

5.2 Ablation Study of Core Components

To evaluate the effectiveness of CARE-RAG's core components, we perform an ablation study under three settings. **w/o Stage1**: Removes the parameter Record Comparison stage and relies only on external retrieved evidence for answer generation. **w/o Stage2**: Removes the Retrieval Result Refinement module. **w/o Stage3**: Removes the Conflict-Driven Summarization stage, omitting both conflict.

As shown in Table 4, all three components contribute significantly to the overall performance of CARE-RAG across datasets and model backbones. 1) Introducing external retrieved evidence and refining it into a structured Context-aware evidence ($\mathcal{E}_c$) leads to substantial gains over using parameter knowledge alone. For example, on the NQ dataset with LLAMA-3-8B, adding refined external evidence yields a +20.4 EM improvement. This highlights the importance of incorporating external information in a structured and relevant form via Π_{ref}; 2) Adding explicit conflict resolution—through conflict detection ($\mathcal{M}_c$) and conflict-aware answer synthesis (Π_{synth})—provides consistent additional gains of 1–2 EM/F1. This shows the value of not only using external knowledge but also

Method	NQ*		TriviaQA*		HotpotQA*		ASQA*		WikiQA*.	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
claude-3-5-haiku-latest										
No RAG	50.6	52.3	77.8	78.9	42.1	44.8	40.9	43.3	35.5	39.2
RAG	51.9	53.0	78.7	79.1	43.3	44.7	41.0	42.9	35.5	37.9
InstructRAG	67.7	68.3	79.2	79.7	53.2	54.4	50.1	50.7	39.5	41.6
GenRead	57.0	57.5	80.1	80.4	43.9	44.7	46.3	47.0	35.4	36.5
Self-RAG	52.9	53.8	79.1	79.8	44.9	46.6	41.1	42.56	36.8	39.0
CARE-RAG	**68.8**	**69.2**	**85.9**	**86.1**	**57.9**	**58.6**	**58.8**	**59.3**	**47.5**	**48.3**
gemini-2.0-flash										
No RAG	42.4	50.1	70.8	73.7	39.6	47.7	45.2	54.1	28.0	39.2
RAG	46.1	51.4	72.4	75.8	39.5	45.2	44.0	47.2	31.4	38.6
InstructRAG	65.3	66.7	75.1	76.5	49.1	50.9	46.9	48.6	41.2	44.7
GenRead	57.5	57.9	82.6	83.9	48.7	49.3	49.6	49.7	44.4	45.2
Self-RAG	49.4	52.4	77.5	78.7	39.4	41.8	42.6	45.3	34.5	38.0
CARE-RAG	**68.0**	**68.5**	**86.7**	**87.1**	**61.4**	**62.3**	**63.6**	**64.2**	**56.7**	**57.7**
gpt-4.1-nano-2025-04-14										
No RAG	35.8	40.0	62.0	64.8	31.7	37.6	28.0	33.0	31.1	39.4
RAG	39.4	43.6	65.4	67.7	33.9	38.3	31.7	35.2	31.8	39.0
InstructRAG	58.5	60.5	72.5	73.6	53.5	56.24	48.1	50.08	40.4	44.5
GenRead	51.0	51.9	72.9	73.5	42.6	43.8	39.4	40.7	37.4	39.3
Self-RAG	43.7	46.0	68.1	69.6	35.9	39.3	34.2	37.5	32.2	38.2
CARE-RAG	**66.2**	**66.5**	**81.6**	**81.2**	**56.7**	**57.2**	**53.0**	**53.4**	**47.6**	**48.2**

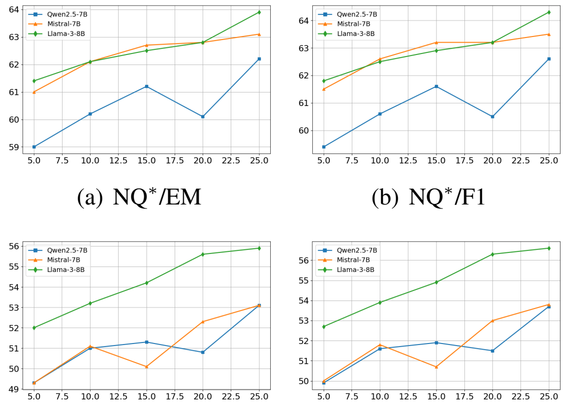
Table 5: Comparing Conflict-Aware and Reliable Evidence for RAG with closed-source models on five QA benchmarks (EM/F1 scores). CARE-RAG achieves superior performance across all datasets and models.

explicitly identifying and reconciling inconsistencies between the internal parameter knowledge ($\mathcal{E}_p$) and the retrieved evidence ($\mathcal{E}_c$). Such targeted conflict handling is crucial for ensuring factual consistency and generating trustworthy answers, leading to the full performance of CARE-RAG.

5.3 Sensitivity to Retrieval Volume

To assess the robustness of our method under varying retrieval volumes, we conduct an ablation study on K, the number of retrieved evidences. We evaluate CARE-RAG under different retrieval volumes, varying K from 5 to 25. The results are presented in Figure 3.

The findings indicate that CARE-RAG effectively utilizes increased context, benefiting particularly from its context refinement mechanism Π_{ref}. Performance generally peaks around $K = 15$–20, beyond which it plateaus, showing remarkable stability even when potentially lower-quality or redundant evidence is included. This contrasts with simpler RAG methods, which often suffer from noise accumulation at higher K values. These results suggest that CARE-RAG's structured reasoning and conflict resolution mechanisms are effective at filtering and prioritizing information, thereby maintaining performance even under noisy conditions.



(a) NQ*/EM

(b) NQ*/F1

(c) HotpotQA*/EM

(d) HotpotQA*/F1

Figure 3: Sensitivity to retrieval size (K). EM/F1 scores for NQ and HotpotQA across three open-source models.

5.4 Robustness to Retrieval Variations

A robust RAG system must remain effective under imperfect retrieval conditions, where the provided evidence may vary significantly in relevance, completeness, or even contradict the original query intent. Figure 4 illustrates EM scores across three datasets under four different evidence strategies:

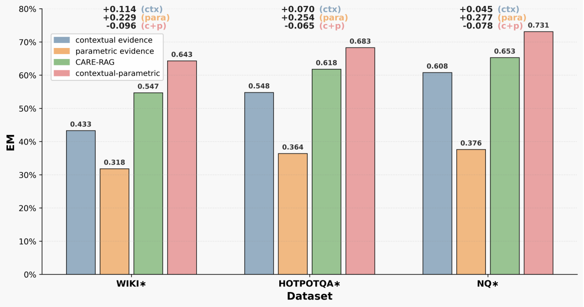


Figure 4: EM performance across three datasets using different retrieval evidence sources.

contextual only, parameter only, their direct combination, and CARE-RAG. CARE-RAG consistently outperforms both contextual and parameter-only baselines, achieving gains of up to 0.239 EM. These trends highlight CARE-RAG's superior robustness to variations in evidence quality and composition, especially in scenarios with conflicting or incomplete information.

This robustness stems from CARE-RAG's conflict-aware synthesis process: contextual evidence $\mathcal{E}_c$, retrieved and refined through Π_{ref}, is systematically compared with parameter-derived knowledge $\mathcal{E}_p$ using the conflict detector $\mathcal{M}_c$. This enables the model to identify and suppress misleading or contradictory signals, prioritize reliable content, and ultimately produce more accurate and trustworthy answers even in noisy or adversarial retrieval settings.

6 Related Work

RAG aims to enhance Large Language Models (LLMs) by incorporating external knowledge (Lewis et al., 2020; Guu et al., 2020). Early work and pretraining objectives focused on effective retrieval integration (Izacard et al., 2023). However, the reliability of RAG is often challenged by the quality of retrieved information and the model's ability to synthesize it with internal knowledge. Improving the retriever component itself is an active area of research, with methods like training utility-based retrievers through shared context attribution showing promise in enhancing the relevance and utility of retrieved evidences (Xu et al., 2025).

Several approaches have sought to improve RAG systems beyond basic retrieval and initial integration. For instance, REPLUG (Shi et al., 2023) investigated direct integration of retrieval into black-box models. RA-DIT (Lin et al., 2023) and InstructRetro (Wang et al., 2023a) explored instruction tuning and alignment to better connect retrieved knowledge with downstream tasks. RankRAG (Yu et al., 2024) focused on optimizing passage ranking. While these methods improve specific RAG pipeline components, they may not fully address issues arising from conflicting or unreliable retrieved content, nor the nuanced challenge of maintaining faithfulness to the provided context during generation, an issue explored by works like Context-DPO (Bi et al., 2024a) which aims to align LMs for context-faithfulness. Our work, Conflict-Aware and Reliable Evidence for RAG (CARE-RAG), specifically targets the post-retrieval synthesis stage to enhance robustness in such scenarios.

A key challenge in RAG is managing knowledge conflicts, where retrieved information contradicts the LLM's internal knowledge or other retrieved evidences (Wang et al., 2023a; Zhou et al., 2025; Zou et al., 2024; Jin et al., 2024; Xie et al., 2023; Bi et al., 2025). Such conflicts can lead to factual inaccuracies. This intersects with the broader field of knowledge editing in LLMs, where techniques aim to update or correct an LLM's internal facts, for instance, by biasing key entities (Bi et al., 2024d) or enhancing model confidence on edited facts through contrastive decoding (Bi et al., 2024b). Although studies acknowledge performance variations with noisy data (Jiang et al., 2024; Chen et al., 2024), dedicated mechanisms within RAG for explicit conflict resolution are crucial. To ensure fair evaluation of these capabilities, and acknowledging the impact of data quality (Jacovi et al., 2023).

7 Conclusion

CARE-RAG is a conflict-aware and reliable framework for retrieval-augmented question answering that systematically addresses key reliability challenges in RAG systems, including outdated supervision, noisy retrieval, and inconsistencies between internal and external knowledge. By integrating structured parameter introspection, fine-grained context refinement, lightweight conflict detection, and a QA repair mechanism, CARE-RAG enhances factual consistency and robustness across diverse tasks. Extensive experiments conducted on five QA benchmarks and multiple model backbones demonstrate that CARE-RAG consistently outperforms competitive baselines. These findings underscore the importance of explicitly modeling knowledge conflicts and support CARE-RAG as a promising and generalizable solution for improving trustworthiness in retrieval-augmented generation.

Limitations

CARE-RAG demonstrates notable improvements over existing retrieval-augmented methods; however, certain limitations remain. The multi-stage approach inherently incurs greater computational overhead compared to simpler RAG frameworks, potentially impacting inference efficiency. Additionally, the performance of CARE-RAG, particularly its conflict detection and resolution capabilities, remains closely tied to the quality of the underlying language models and their fine-tuned capabilities, which might not fully resolve highly subtle or adversarially constructed knowledge conflicts. Furthermore, despite increased robustness to noisy retrieval, overall efficacy still depends substantially on the initial document retriever's accuracy and comprehensiveness. The QA Repair module, though effective against typical dataset issues, may not universally handle all types of benchmark artifacts or specialized domain knowledge without further refinement and domain-specific adaptation.

Ethical Considerations

The development of advanced retrieval-augmented generation systems, including CARE-RAG, raises significant ethical considerations. The QA Repair process, designed to address dataset biases by correcting outdated or mismatched information, inherently involves subjective judgments regarding the definition and scope of "correctness." Such judgments must be transparently managed and periodically revisited to prevent inadvertent bias introduction. Additionally, improvements in factual accuracy and consistency, although broadly beneficial, increase the risk of generating convincing yet inaccurate information if misused or inadequately supervised. Reliance on externally retrieved knowledge also introduces the possibility of propagating existing biases or inaccuracies from source materials. Therefore, ongoing research efforts should emphasize robust bias detection, clear attribution of information sources, transparent conflict-resolution mechanisms, and the establishment of responsible use guidelines to ensure these powerful tools are deployed ethically, fairly, and constructively.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. Technical report, Anthropic.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Roberto Balestri. 2025. Gender and content bias in large language models: a case study on google gemini 2.0 flash experimental. *arXiv preprint arXiv:2503.16534*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2024a. Context-dpo: Aligning language models for context-faithfulness. *arXiv preprint arXiv:2412.15280*.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024b. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024c. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *arXiv preprint arXiv:2404.00216*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. 2024d. Adaptive token biaser: Knowledge editing via biasing key entities. *EMNLP 2024*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

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.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

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

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.

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

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

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.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2024. Fit-rag: black-box rag with factual information and token reduction. *arXiv preprint arXiv:2403.14374*.

OpenAI. 2025. Gpt-4.1 nano model card.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2025. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023a. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, and 1 others. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Yuanhai Xue, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. Training a utility-based retriever through shared context attribution for retrieval-augmented language models. *arXiv preprint arXiv:2504.00573*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2023. End-to-end beam retrieval for multi-hop question answering. *arXiv preprint arXiv:2308.08973*.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. 2025. Trustrag: Enhancing robustness and trustworthiness in rag. *arXiv preprint arXiv:2501.00879*.

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

A Implementation Details

A.1 Models Used

Our experiments primarily utilized three open-source Large Language Models (LLMs): Mistral-7B (Jiang et al., 2023), Llama-3-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Yang et al., 2024). For experiments involving closed-source models (as detailed in Table 5), we employed.

Unless otherwise specified, the same backbone LLM (from either the open-source or closed-source set, depending on the experiment) was consistently used for all stages of the CARE-RAG pipeline: eliciting the initial LLM response (A_{init}, corresponding to $\mathcal{E}_p$ generation in Algorithm 1), refining retrieved results into $\mathcal{E}_c$ (which may involve structured reasoning), conflict detection by $\mathcal{M}_c$ to produce δ_c and r_c, CARE-RAG to generate $\hat{a}$, and also for the QA Repair module (f_{repair}) described in Appendix B..

A.2 Retrieval Setup

For each question, we retrieved the top-K relevant evidences from the respective corpus. In our main experiments (Tables 3 and 5), K was set to 5. An analysis of CARE-RAG's sensitivity to varying K (from 5 to 25) is presented in Table **??**.

A.3 Inference Framework

All inferences for open-source LLMs were performed using the vLLM framework (Kwon et al., 2023) to ensure efficiency and reproducibility. For closed-source models, inferences were made via their respective official APIs. The following inference parameters were consistently applied for generation tasks (e.g., generating initial responses for $\mathcal{E}_p$, the refined evidence $\mathcal{E}_c$, the final answer $\hat{a}$, and repaired answers in f_{repair}) unless a specific module (like the conflict detector $\mathcal{M}_c$) required different settings:

- `max_tokens`: 1024
- `temperature`: 0.7
- `top_p`: 1.0

For classification-like tasks performed by the conflict detector $\mathcal{M}_c$ to determine δ_c (and generate r_c), we typically used a temperature (e.g., 0.7, or potentially lower like 0.0 for more deterministic conflict/no-conflict output if desired) to encourage more deterministic outputs, though the primary mechanism for binary classification was specific instruction prompting tailored to elicit a "0" or "1" and a rationale.

A.4 Evaluation Metrics

System performance was primarily evaluated using standard Exact Match (EM) and F1 scores. These metrics were computed against the (potentially) repaired ground truth answers generated by our QA Repair module (detailed in Appendix B), ensuring a fair and robust assessment across all compared methods.

B QA Repair Module

B.1 Overview

As highlighted in our experimental setup (Section 4), standard QA benchmarks often suffer from issues such as temporal drift (outdated answers) or semantic mismatches between questions and ground truths. These flaws can lead to misleading evaluations of RAG systems. To ensure a fairer and more accurate assessment of model capabilities, we introduce a QA Repair module. This module is applied as a pre-processing step to the test instances of all evaluated benchmarks, correcting potential issues in the original ground truth answers before any model evaluation takes place. The module operates on an input triplet: (*question* q, *original ground truth answer* a_{gt}, and potentially relevant *retrieved context* C, though C is not always strictly necessary for the repair logic if general world knowledge suffices).

B.2 Repair Mechanism

The core of the QA Repair module is a classifier, f_{repair}, implemented using a prompted Large Language Model (LLM). This classifier is tasked with assessing whether the original ground truth answer, a_{gt}, is likely outdated, semantically inconsistent with the question q, or otherwise flawed, considering current world knowledge and the precise intent of q. It outputs a binary flag:

$$\gamma_{\text{repair}} = f_{\text{repair}}(q, a_{\text{gt}}, C_{\text{optional}}) \in \{0, 1\}$$

If $\gamma_{\text{repair}} = 1$ (indicating a detected flaw), a repair process is initiated. This process, also typically leveraging a prompted LLM, employs structured reasoning or direct knowledge querying (based on q and potentially C) to generate a revised, more accurate ground truth answer, a'_{gt}. In some instances, to resolve ambiguity or align with the corrected answer, the original question q might also be minimally refined to q'. The output of this stage is thus a potentially corrected question-answer pair (q', a'_{gt}). This repaired pair is then used as the reference for evaluating all RAG models (including baselines and CARE-RAG) in our experiments.

B.3 Illustrative Examples

The following examples illustrate typical scenarios handled by the QA Repair module. Note that in these examples, "Current Model Answer" (if such a term was used previously, otherwise this clarification might not be needed) is re-interpreted as the "Repaired Ground Truth (a'_{gt})" produced by our QA Repair module if a flaw was detected in the original "Ground Truth (a_{gt})".

B.3.1 Example 1: Temporal Drift

Scenario: Temporal Drift

Original Query (q): Who scored the most points in their NBA career?

Original Ground Truth (a_{gt}): Kareem Abdul-Jabbar

QA Repair Module Output:

- **Detection** ($\gamma_{\text{repair}} = 1$): The answer "Kareem Abdul-Jabbar" is outdated.
- **Repaired Ground Truth** (a'_{gt}): LeBron James (as of [current date/year of dataset repair])
- **Repaired Query** (q'): (No change in this case) Who scored the most points in their NBA career?

B.3.2 Example 2: Answer Type Mismatch / Factual Inaccuracy

Scenario: Answer Type Mismatch / Factual Inaccuracy

Original Query (q): When was the Statue of Liberty in France built?

Original Ground Truth (a_{gt}): Paris

QA Repair Module Output:

- **Detection** ($\gamma_{\text{repair}} = 1$): The answer "Paris" does not answer "When" and is factually incorrect for the construction date.
- **Repaired Ground Truth** (a'_{gt}): Construction was completed in July 1884. (Or simply: July 1884)
- **Repaired Query** (q'): (No change in this case) When was the Statue of Liberty in France built?

B.3.3 Detailed Analysis of Repaired Data

Figure 5 details the error composition within corrected samples from five QA benchmarks (1,000 samples each were analyzed for repair needs). The chart displays the counts of "Mismatch" errors (semantic misalignment) and "Out-of-date" errors (temporal drift) among the instances that required repair. For example, all 67 repaired Wiki samples were out-of-date, while TriviaQA's 74 repairs included approximately 33 mismatches. Notably, the NQ dataset, with 240 repaired samples, exhibits an overlap in error types: the sum of its reported mismatch (approx. 47) and out-of-date (approx. 196) components exceeds the total repair count, indicating some samples possess both error attributes. This granular analysis, highlighting diverse error profiles and potential co-occurrences as in NQ, underscores the necessity of our comprehensive QA Repair process for establishing a reliable evaluation baseline and the importance of targeted, rather than one-size-fits-all, approaches to dataset noise.

C Core Conflict-Aware and Reliable Evidence for RAG Prompts

This section provides the specific prompt formats used for the core stages of the Conflict-Aware and Reliable Evidence for RAG (CARE-RAG) framework, corresponding to the Π symbols in Algorithm 1.

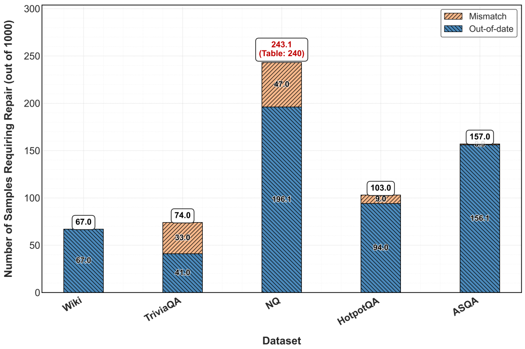


Figure 5: Mismatch and Out-of-date error distribution in repaired samples from five QA datasets. NQ shows co-occurring error types. Repair impact on Qwen2.5-7B (using the notation from your paper if Qwen3-235B-A22B is a specific variant) is detailed in Table 2 (Please verify this table label and its content regarding repair impact specifically with this model version).

C.1 parameter Record Comparison Prompts (Π_{init} and Π_{iter})

C.1.1 Iterative parameter Response Prompt (Π_{iter})

Objective: Elicit alternative or more diverse parameter responses a_i ($i > 0$), given the previously generated parameter responses within $\mathcal{E}_p$, to further explore the model's internal knowledge space.

Task: Based on your previous answer(s) and your internal knowledge, provide a different or more detailed/nuanced answer to the following question.

Question: {question}

Previous parameter Answer(s) ($\mathcal{E}_p$ so far): {previous_parameter_answers}

Answer (Iterative - a_i):

C.1.2 Initial parameter Response Prompt (Π_{init})

Objective: Elicit the model's first direct response a_0 based solely on its internal parameter knowledge, forming the basis of $\mathcal{E}_p$.

Task: Provide a concise and direct answer to the following question using only your internal knowledge.

Question: {question}

Answer (Initial - a_0):

C.2 Retrieval Result Refinement Prompt (Π_{ref})

Objective: Instruct the model to distill the retrieved evidences C into a concise and salient context-aware evidence $\mathcal{E}_c$, by extracting key factual observations, identifying ambiguities, and forming *context-grounded conclusions. This corresponds to Stage II in Figure 2.*

Context Refinement Prompt **Instruction**: Analyze the provided Context thoroughly in relation to the Question. Your goal is to extract the most relevant factual information, identify any ambiguities or limitations within the context, and conclude with the most likely answer(s) or key insights that can be *purely grounded in the provided Context*. If no complete answer is available from the context, state that and explain why. **Retrieved Context evidences (C)**:

- {context_evidence_1}
- {context_evidence_2}
- ...
- {context_evidence_k}

Question (q): {question} **Your Distilled Context-Aware evidence ($\mathcal{E}_c$) based *only* on the Retrieved Context should include**:

- Key factual claims relevant to the Question.
- Identified ambiguities or limitations in the provided Context.
- A concluding summary or answer candidate(s) strictly derived from the Context.

C.3 Conflict Detection Prompt (Π_c)

Objective: Explicitly evaluate whether the model's consolidated parameter-aware evidences ($\mathcal{E}_p$) semantically conflict with the refined Context-aware evidence ($\mathcal{E}_c$). This is used by the conflict detector module $\mathcal{M}_c$ and corresponds to Stage III in Figure 2.

Conflict Detection Prompt

Instruction: Evaluate if the "parameter Knowledge Response" contradicts the "Context-derived Response" for the given Question. Consider factual differences (e.g., names, dates, values), temporal mismatches, or significant semantic inconsistencies. Output 'Conflict: 1' if a contradiction is found. Output 'Conflict: 0' if there is no contradiction or if they are consistent. Provide a brief step-by-step reasoning for your decision.

Question (q): {question}

parameter Knowledge Response (Consolidated from $\mathcal{E}_p$): {consolidated_parameter_response}

Context-derived Response (from $\mathcal{E}_c$): {context_aware_evidence_summary}

Analysis and Conflict Decision (δ_c, r_c):

C.4 CARE-RAG Generation Prompt (Π_{synth})

Objective: Generate the final answer ($\hat{a}$) by integrating the parameter-aware evidences ($\mathcal{E}_p$), the refined Context-aware evidence ($\mathcal{E}_c$), and the conflict detection signal (δ_c, r_c). This corresponds to Stage IV in Figure 2.

Final Answer Synthesis Prompt **Contextual Note**: A potential conflict (indicated by δ_c) between internal parameter knowledge ($\mathcal{E}_p$) and external information ($\mathcal{E}_c$) might have been detected, with rationale r_c. **Your Task is to Synthesize the Best Final Answer ($\hat{a}$)**:

1. Based on all inputs, identify the **best-supported single candidate answer**.
2. Consider information recency, source reliability, and overall coherence, especially if a conflict ($\delta_c = 1$) was detected.
3. **If conflict** ($\delta_c = 1$): Explicitly address the discrepancy from r_c. Attempt to resolve it by selecting more credible information or state remaining uncertainty.
4. **If no conflict** ($\delta_c = 0$): Primarily ground your answer in $\mathcal{E}_c$, using $\mathcal{E}_p$ as confirmation.
5. Provide **concise reasoning** for your chosen answer, citing relevant inputs ($\mathcal{E}_p, \mathcal{E}_c, r_c$). Clearly state any remaining ambiguity or temporal uncertainty.

Inputs Provided:

- **Question** (q): {question}
- **parameter Knowledge Response (Consolidated $\mathcal{E}_p$)**: {consolidated_parameter_response}
- **Context-derived Response** ($\mathcal{E}_c$): {context_aware_evidence_summary}
- **Conflict Detection Flag** (δ_c): {δ_c}
- **Conflict Rationale** (r_c): {r_c}

Required Output Format for Final Answer ($\hat{a}$):

- **Final Answer**: ...
- **Reasoning for Final Answer**: ... (Address conflict per r_c if $\delta_c = 1$)
- **Ambiguity/Uncertainty Assessment**: ... (If any)

D Detailed Process Walkthrough

We illustrate the complete CARE-RAG workflow using the NBA scoring example, as discussed in Figure 2 (Stage I-IV visual overview) and referenced in the main text.

1. **Input Question** (q): *"Who scored the most points in their NBA career?"*

2. **parameter Record Comparison (generates $\mathcal{E}_p$)**: The LLM $\mathcal{M}$, using prompt Π_{init} (Ap-

pendix C.1), generates its initial context-free response a_0. For this example, we assume $n = 1$, so the consolidated parameter-aware evidences $\mathcal{E}_p$ is: *"LeBron James"* (assuming the LLM's parameter knowledge is up-to-date).

3. **Retrieval Result Refinement (generates $\mathcal{E}_c$):** The retriever $\mathcal{R}$ returns evidences C, e.g.: c_1: *"Kareem Abdul-Jabbar is the all-time leading scorer in the NBA, with 38,387 total points."*; c_2: *"Kareem rewriting scoring records."*; c_3: *"As of 2023, James holds the record."*
   Using prompt Π_{ref} (Appendix C.2), $\mathcal{M}$ processes C into the Context-aware evidence $\mathcal{E}_c$. For example, $\mathcal{E}_c$ might be distilled to: *"Retrieved evidences state Kareem Abdul-Jabbar was the all-time leading scorer (38,387 points). One passage indicates that as of 2023, James holds the record, suggesting a change."*

4. **Conflict-Driven Summarization (generates δ_c, r_c):** The conflict detector $\mathcal{M}_c$, using prompt Π_c (Appendix C.3), compares $\mathcal{E}_p$ (*"LeBron James"*) with $\mathcal{E}_c$ (*"Retrieved evidences state Kareem... James holds the record..."*).
   Assuming for clearer conflict demonstration that $\mathcal{E}_c$ was distilled to only reflect outdated info like: *"According to retrieved text, Kareem Abdul-Jabbar is the top scorer."*
   The outputs are: Conflict Flag (δ_c): 1. Conflict Rationale (r_c): *"parameter knowledge ($\mathcal{E}_p$) states LeBron James, while context-derived information ($\mathcal{E}_c$) states Kareem Abdul-Jabbar. These conflict."*

5. **CARE-RAG Generation (generates $\hat{a}$):** The LLM $\mathcal{M}$, using prompt Π_{synth} (Appendix C.4), receives q, $\mathcal{E}_p$, $\mathcal{E}_c$, $\delta_c = 1$, and r_c. The **Final Answer** ($\hat{a}$) is, for example: *"LeBron James is NBA's all-time leading scorer. While some historical records mention Kareem Abdul-Jabbar, LeBron James has surpassed this record, aligning with current information."*
   The **Reasoning** would acknowledge the conflict identified by r_c and explain the prioritization of current parameter knowledge ($\mathcal{E}_p$) or the more recent parts of $\mathcal{E}_c$, treating Kareem's record as historical.

E Component Output Examples

This section provides additional, isolated examples of outputs from key components and stages of the Conflict-Aware and Reliable Evidence for RAG (CARE-RAG) framework. These examples illustrate the specific outputs for Context-aware evidence Generation (formerly Structured Reasoning), Conflict Detection, and CARE-RAG Generation. Examples for QA Repair (Appendix B) and parameter-aware evidence Generation ($\mathcal{E}_p$, detailed in Appendix C.1) are covered elsewhere or are straightforward.

E.1 ★ Conflict Detection Output Example (δ_c, r_c from $\mathcal{M}_c$)

Task: Evaluate whether the consolidated parameter-aware evidences ($\mathcal{E}_p$) contradict the refined Context-aware evidence ($\mathcal{E}_c$) for the given query. Output a conflict flag ($\delta_c \in \{0, 1\}$) and a rationale (r_c). This uses prompt Π_c (Appendix C.3).

Query (q): *Who was "Suite: Judy Blue Eyes" written about?*

Input: Consolidated parameter-aware evidences ($\mathcal{E}_p$) (simulated):

- *Stephen Stills wrote it about Judy Collins, his former girlfriend.*

Input: Refined Context-aware evidence ($\mathcal{E}_c$) (simulated):

- *The song "Suite: Judy Blue Eyes" was written by Stephen Stills. It references Judy Collins and their relationship.*

LLM Reasoning for Conflict Detection (part of r_c):

- Both inputs identify Stephen Stills as the author and Judy Collins as the subject.
- The information provided is consistent and complementary, with no factual contradictions.

Conflict Detection Output:

- **Conflict Flag** (δ_c): **0**
- **Conflict Rationale** (r_c): *No conflict detected. Both parameter knowledge and context-derived information consistently identify Stephen Stills as the author and Judy Collins as the subject of the song.*

E.2 ✌ Context-aware evidence Generation Example ($\mathcal{E}_c$ from Π_{ref})

Task: Distill the retrieved context (C) into a concise and salient Context-aware evidence ($\mathcal{E}_c$). This uses prompt Π_{ref} (Appendix C.2).

Input Question (q): *Who was Henry Feilden?*

Input Retrieved Context (C):

- *"Colonel Henry Wemyss Feilden CB (1838–1921) was a British Army officer, naturalist and Arctic explorer."*
- *"Feilden was the second son of the second Baronet Feilden..."*
- *"Henry Master Feilden (1818–1875) was a Conservative politician..."*
- *"Some sources incorrectly identify Henry Feilden as an actor."*

Generated Context-aware evidence ($\mathcal{E}_c$):

- **Key Factual Claims**:
  - Two distinct individuals named Henry Feilden are mentioned:
    - ∗ Colonel Henry Wemyss Feilden (1838–1921): British Army officer, naturalist, Arctic explorer.
    - ∗ Henry Master Feilden (1818–1875): Conservative politician.
  - The claim that a Henry Feilden was an actor is explicitly refuted for at least one, and implicitly for both if the context is considered complete regarding their professions.
- **Identified Ambiguities/Issues**: The query "Who was Henry Feilden?" is ambiguous without further specification, as the context provides two distinct individuals.
- **Context-Grounded Conclusion/Summary**: The retrieved context describes two notable individuals named Henry Feilden. To provide a specific answer to "Who was Henry Feilden?", clarification on which individual is being referred to would be necessary. The claim about Henry Feilden being an actor is noted as incorrect based on the context.

E.3 ▲ CARE-RAG Generation Output Example ($\hat{a}$ from Π_{synth})

Task: Generate the final reconciled answer ($\hat{a}$) by integrating all available information. This uses prompt Π_{synth} (Appendix C.4).

Input Question (q): Who is the original artist of *Sound of Silence*, released in 1964?

Input: Consolidated parameter-aware evidences ($\mathcal{E}_p$) **(simulated)**: *Simon & Garfunkel*

Input: Refined Context-aware evidence ($\mathcal{E}_c$) **(simulated)**: *Based on the provided context, Simon & Garfunkel are identified as the original artists of "Sound of Silence" released in 1964. No conflicting information found in context.*

Input: Conflict Flag (δ_c): 0 **Input: Conflict Rationale** (r_c): *No conflict detected between parameter knowledge and context-derived information.*

CARE-RAG Generation Output ($\hat{a}$):

- **Final Answer**: Simon & Garfunkel were the original artists of the song "Sound of Silence," released in 1964.
- **Reasoning for Final Answer**: Both the consolidated parameter knowledge ($\mathcal{E}_p$) and the refined context-aware evidence ($\mathcal{E}_c$) consistently identify Simon & Garfunkel. The conflict flag ($\delta_c = 0$) confirms no discrepancy was found. There is no ambiguity regarding the 1964 release.
- **Ambiguity/Uncertainty Assessment**: None detected.