

TRIAD: A Three-Axis Framework for Knowledge Conflict Resolution in RAG

Yizhou Jin^{1*}, Xudong Li^{1*}, Dandan Song^{1†}, Yue Yang¹, Jun Yang¹,
Yuhang Tian¹, Huipeng Ma^{1,2}, Chenhao Li¹, Luan Zhang¹

¹Beijing Institute of Technology, China

²QiYuan Lab

jqyyzy@gmail.com, sdd@bit.edu.cn

Abstract

Retrieval-augmented generation (RAG) improves factual question answering by grounding LLMs in external evidence, but it becomes brittle when retrieved context conflicts with the parametric memory of the model. Existing training-free methods often address this problem by globally comparing two knowledge sources or structurally aligning them, yet they still struggle with three coupled failure axes: question complexity, long misleading evidence, and off-target parametric recall. We first quantitatively analyze the individual and joint effects of these three axes, showing that their interaction makes knowledge conflict resolution harder than any single axis alone. Motivated by this finding, we propose **TRIAD**, a training-free framework for knowledge conflict resolution in RAG. TRIAD decomposes a question into minimal subquestions, performs subquestion-level evidence calibration to retain concise and sufficient evidence, and uses indirect clues from calibrated evidence to guide parametric knowledge elicitation. Across three benchmarks and six LLMs, TRIAD achieves state-of-the-art performance and is the only method that surpasses a setting with additional retrieval in most cases. Additional experiments further demonstrate its generalization and worst-case robustness.

1 Introduction

Recent large language models (LLMs) have shown strong in-context learning and reasoning abilities, making them promising backbones for knowledge-intensive question answering (Brown et al., 2020; Wei et al., 2022). However, parametric knowledge alone is still limited because it is hard to update, and often unreliable for infrequent facts (Lewis et al., 2020; Mallen et al., 2023). Retrieval-augmented generation (RAG) mitigates this limi-

*Equal contribution.

†Corresponding author.

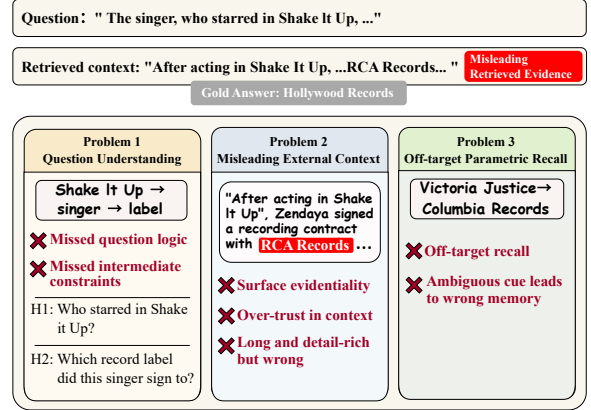


Figure 1: An illustration of QA under Knowledge Conflict via a real example. The figure shows three failure axes of existing methods.

tation by supplementing LLMs with external evidence (Lewis et al., 2020).

However, the success of RAG depends critically on retrieval quality. In realistic settings, retrieved evidence may be irrelevant, misleading, or even malicious (Wang et al., 2025). When such evidence contradicts the model’s parametric knowledge, the system faces a *Knowledge Conflict* (Xu et al., 2024b). This problem becomes particularly severe when the retrieved context is long and superficially convincing, which can substantially hurt model performance, reliability, and answer accuracy (Liu et al., 2024).

Existing works tackle this problem from two directions. One direction improves the retrieval side, for example by deciding when retrieval is necessary or adapting retrieval strategy to question complexity (Asai et al., 2024; Jeong et al., 2024). The other direction tries to resolve the conflict during generation by explicitly bringing parametric knowledge into the decision process.

These generation methods can be further divided into two categories. Global aggregation methods (GAMs) elicit parametric knowledge and compare

it with retrieved evidence in a side-by-side manner (Liu et al., 2022; Wang et al., 2023). Structured alignment methods (SAMs) instead decompose, align, or iteratively consolidate internal and external knowledge into smaller units before making the final decision, which introduces a more complex reasoning process (Zhang et al., 2023; Huo et al., 2025; Wang et al., 2025).

Despite these advances, they still face three critical challenges. (1) They typically resolve conflict at the level of the full question, which can obscure the logical dependencies required by multi-hop or compositional queries (Trivedi et al., 2022). (2) When incorrect retrieved context is long, detail-rich, and more superficially convincing, the model can be attracted by its surface evidentiality and overtrust the retrieved context (Liu et al., 2024). (3) They usually rely on question-guided parametric knowledge elicitation, but when question cues are overly broad, entities are ambiguous, or situational constraints are underspecified, the recalled parametric knowledge may become misaligned with the intended meaning of the question (Mallen et al., 2023).

In this paper, we conceptualize these issues as three coupled failure axes, namely question-axis, context-axis, and memory-axis. We first analyze the impact of each axis, and then examine their joint effect, revealing the necessity of resolving them together. Motivated by this analysis, we propose **TRIAD**, a three-axis framework for knowledge conflict resolution in RAG. TRIAD first decomposes a question into minimal answerable subquestions to reduce reasoning drift on complex queries. It then performs subquestion-level evidence calibration to retain concise and sufficient evidence. Finally, TRIAD extracts indirect clues from calibrated evidence to guide parametric knowledge elicitation. Our contributions are threefold:

- We quantitatively analyze three failure axes and show that their interaction makes knowledge conflict resolution substantially harder.
- We propose TRIAD, a framework that jointly addresses these three axes through question decomposition, subquestion-level evidence calibration, and clue-guided parametric knowledge elicitation.
- Across diverse benchmarks and backbone settings, TRIAD achieves the strongest overall

performance. Further experiments demonstrate its generalization and robustness under varied conflict conditions.

2 Related Work

Knowledge Conflict. Recent surveys identify Context-Memory Conflict as a core form of knowledge conflict in LLMs, where retrieved evidence disagrees with parametric knowledge (Xu et al., 2024b). In RAG, such conflict is often triggered by imperfect retrieval, and disagreement between parametric and non-parametric memories (Lewis et al., 2020; Mallen et al., 2023; Wang et al., 2025; Zhang et al., 2025; Choi et al., 2025). Prior work further shows that long, irrelevant, or detrimental context can distract LLMs from the information actually needed for reasoning and substantially weaken retrieval robustness (Shi et al., 2023; Liu et al., 2024; Shen et al., 2024; Oh and Thorne, 2023). These findings suggest that the difficulty of Context-Memory Conflict is not limited to source disagreement itself, but is also shaped by how questions are understood, how parametric knowledge is recalled, and how external evidence is presented and weighted.

Methods for Resolving Knowledge Conflict. Existing solutions mainly tackle the problem from two angles. One line improves conflict handling through training-based adaptation, for example by training models or auxiliary modules to control or assess retrieved evidence under conflict (Asai et al., 2024; Jeong et al., 2024; Choi et al., 2025; Lin et al., 2026). The other line resolves conflict without changing model parameters. Some methods act on retrieval by deciding when retrieval is needed, decomposing complex questions into subquestions, or interleaving retrieval with intermediate reasoning to refine evidence before generation (Jiang et al., 2023; Ammann et al., 2025; Trivedi et al., 2023), while others act on generation by bringing internal knowledge into reasoning after retrieval. Within these methods on the generation side, global aggregation methods (GAMs) compare or aggregate internal knowledge and retrieved context at a holistic level (Liu et al., 2022; Wang et al., 2023), whereas structured alignment methods (SAMs) decompose, align, or iteratively consolidate the two knowledge sources (Zhang et al., 2023, 2025; Huo et al., 2025; Wang et al., 2025). Our work is most closely related to training-free methods on the generation side.

3 Empirical Analysis of Failure Axes

To motivate the design of TRIAD, we first analyze the three failure axes of existing methods. We study two categories of baselines, namely GAMs and SAMs, on the KRE versions of SQuAD (Rajpurkar et al., 2018) and MuSiQue (Trivedi et al., 2022), which are two widely used benchmarks for knowledge conflict detection (Ying et al., 2024). On two representative models, Llama-3.1-70B-Instruct (Meta, 2024) and Qwen3.5-9B (Qwen Team, 2026), we analyze the three failure axes introduced in Section 1 and further reveal the harmful effects of their interaction. Detailed setup is deferred to Appendix A.

Accuracy Degrades Sharply As Hop Count Increases. As shown in Figure 9, aggregated over all baseline instances, performance drops by about 15% from one-hop to four-hop questions, and the same downward trend appears consistently across most settings. This pattern underscores the importance of correctly understanding the underlying question logic. Multi-hop questions require the model to preserve intermediate constraints. Once an early subgoal is misidentified, later aggregation or alignment is built on the wrong premise, consistent with prior evidence that compositional QA becomes more brittle under chained dependencies (Trivedi et al., 2022). We do not over-interpret the extreme tail, as such cases are relatively scarce.

Off-Target Parametric Knowledge Consistently Hurts Conflict Resolution. As shown in Figure 9, the second weakness lies in whether elicited parametric knowledge is well aligned with the question semantics. In most settings, higher parametric knowledge relevance is associated with higher accuracy, whereas Medium or Low relevance typically leads to degraded performance. Although stronger method-backbone combinations (SAMs+Llama) are relatively more robust, the overall trend remains clear. Once introduced into the reasoning process, such off-target parametric knowledge can bias both evidence interpretation and final answer selection (Mallen et al., 2023). A mild exception appears for SAMs+Qwen, where Low relevance (39.4%) is slightly above High relevance (38.8%). However, the gap is only 0.6 points, Medium still drops to 28.0%, and this fluctuation does not change the overall conclusion.

Long Misleading Context Lowers Accuracy and Raises Misled Rate. As shown in Figure 9, when

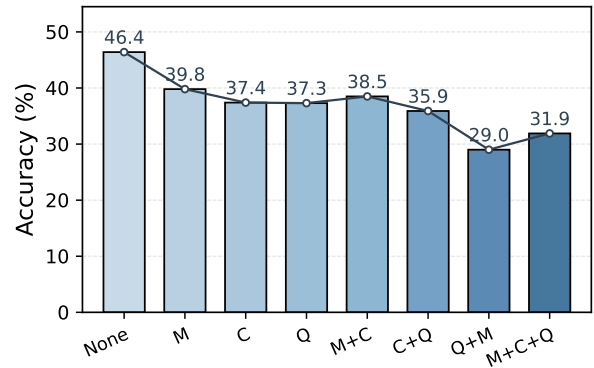


Figure 2: Joint effects of the three failure axes. Here, Q denotes multi-hop questions, M denotes low parametric knowledge relevance, and C denotes context length above the pooled weighted-average threshold. Performance declines as more failure axes co-occur.

context becomes longer, accuracy declines while the Misled Rate correspondingly rises. These results suggest that carefully constructed long misleading context is not merely additional noise. Because it is richer in details and more evidence-like in form, it can project stronger surface evidentiality and dominate the model’s attention even when it is wrong.

The Three Failure Axes Become More Harmful When Coupled. As shown in Figure 2, accuracy is 46.4% when none of the axes is active, declines to 37.3%–39.8% when a single axis is active, to 29.0%–38.5% when two axes co-occur, and to 31.9% when all three are active. Overall, the three axes reinforce one another rather than acting independently.

Taken together, GAMs are generally more vulnerable, but even SAMs struggle when failures in question understanding, parametric knowledge, and context reliability occur simultaneously. TRIAD is designed to address this gap. Its gains should therefore be understood as improvements on a coupled conflict resolution problem rather than on any single axis.

4 Methodology

4.1 Problem Description

We study post-retrieval knowledge conflict resolution in RAG. Given a question and retrieved context containing incorrect information, the model must determine the correct answer under Knowledge Conflict by deciding how to interpret external evidence and how to invoke parametric knowledge.

Notably, our method does not assume that retrieved evidence is incorrect.

Building on the three failure axes identified in Section 3, we propose TRIAD, a training-free framework that executes four modules sequentially: (1) question decomposition, (2) subquestion-level evidence calibration, (3) clue-guided parametric knowledge elicitation, and (4) answer resolution for each hop. Figure 3 provides an overview of the full framework, and Algorithm 1 summarizes the procedure.

4.2 Question Decomposition

TRIAD first converts the original question into a sequence of minimal answerable subquestions, $D(q) = \{h_1, h_2, \dots, h_T\}$, where T is the number of reasoning hops. For multi-hop questions, later subquestions may contain variables that are instantiated by answers resolved in earlier hops. We therefore maintain a binding map

$$B_{t-1} = \{v_1 \mapsto a_1, v_2 \mapsto a_2, \dots, v_{t-1} \mapsto a_{t-1}\}, \quad (1)$$

and instantiate the current hop by

$$\tilde{h}_t = \psi(h_t, B_{t-1}), \quad (2)$$

where each v_i denotes a variable in a later subquestion and a_i is the intermediate answer predicted at hop i . The operator $\psi(\cdot)$ instantiates these variables using previously predicted intermediate answers.

4.3 Subquestion-Level Evidence Calibration

TRIAD then calibrates evidence separately for each instantiated subquestion $\tilde{h}_t$. We segment the retrieved context by period delimiters into a chunk set $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$, where each c_j is a contiguous sentence-level chunk from the retrieved context. For every pair $(\tilde{h}_t, c_j)$, we first compute a semantic relevance score

$$S_{\text{rel}}(\tilde{h}_t, c_j) = \cos(E(\tilde{h}_t), E(c_j)), \quad (3)$$

where $E(\cdot)$ denotes the text embedding function.

To quantify how much of a chunk is actually useful for the current hop, TRIAD prompts an LLM to extract the minimal support span(s) in c_j that are needed to answer $\tilde{h}_t$. We then define the noise penalty as

$$P_{\text{noise}}(\tilde{h}_t, c_j) = 1 - \frac{T_{\text{key}}(\Sigma_{t,j})}{T_{\text{total}}(c_j)}, \quad (4)$$

where $T_{\text{key}}(\Sigma_{t,j})$ is the number of key tokens appearing in the extracted support spans.

Based on relevance and noise, we define a Signal-to-Noise Alignment Metric (SNAM):

$$\text{SNAM}(\tilde{h}_t, c_j) = \alpha \cdot S_{\text{rel}}(\tilde{h}_t, c_j) - \beta \cdot P_{\text{noise}}(\tilde{h}_t, c_j), \quad (5)$$

where α and β control the trade-off between semantic match and noise suppression. TRIAD incrementally selects chunks according to SNAM and uses an LLM to judge whether the selected set is sufficient for answering the current hop. The calibrated evidence is

$$\hat{C}_t = \text{Concat}(c_{j_1}, c_{j_2}, \dots, c_{j_r}), \quad (6)$$

where $c_{j_1}, c_{j_2}, \dots, c_{j_r}$ are the retained chunks. If no partial selection is judged sufficient, TRIAD falls back to the full context. This design suppresses distracting text while preserving enough evidence for the current subquestion.

4.4 Clue-Guided Parametric Knowledge Elicitation

Indirect Clue Extraction. Given the calibrated evidence $\hat{C}_t$, TRIAD extracts a set of indirect clues that help the model better capture the semantics of the question:

$$Z_t = \text{ClueExtract}(\tilde{h}_t, \hat{C}_t). \quad (7)$$

To construct Z_t , TRIAD removes from the extracted clue candidates any content that could directly answer $\tilde{h}_t$, retaining only indirect clues, such as relations, temporal constraints, locations, events, and background conditions, for semantic guidance.

Parametric Knowledge Elicitation. TRIAD then elicits parametric knowledge conditioned on both the current subquestion and the filtered clue set:

$$K_t = \text{Elicit}(\tilde{h}_t, Z_t). \quad (8)$$

Crucially, the extracted clues are treated as weak hints rather than ground truth. This prevents the model from mechanically inheriting misleading retrieved content, while still using calibrated external evidence to guide parametric recall toward the intended semantics of the current hop.

4.5 Answer Resolution for Each Hop

At each hop, TRIAD resolves the current subquestion by jointly reasoning over calibrated external evidence and clue-guided parametric knowledge:

$$a_t = \text{Resolve}(\tilde{h}_t, \hat{C}_t, K_t), \quad t = 1, \dots, T. \quad (9)$$

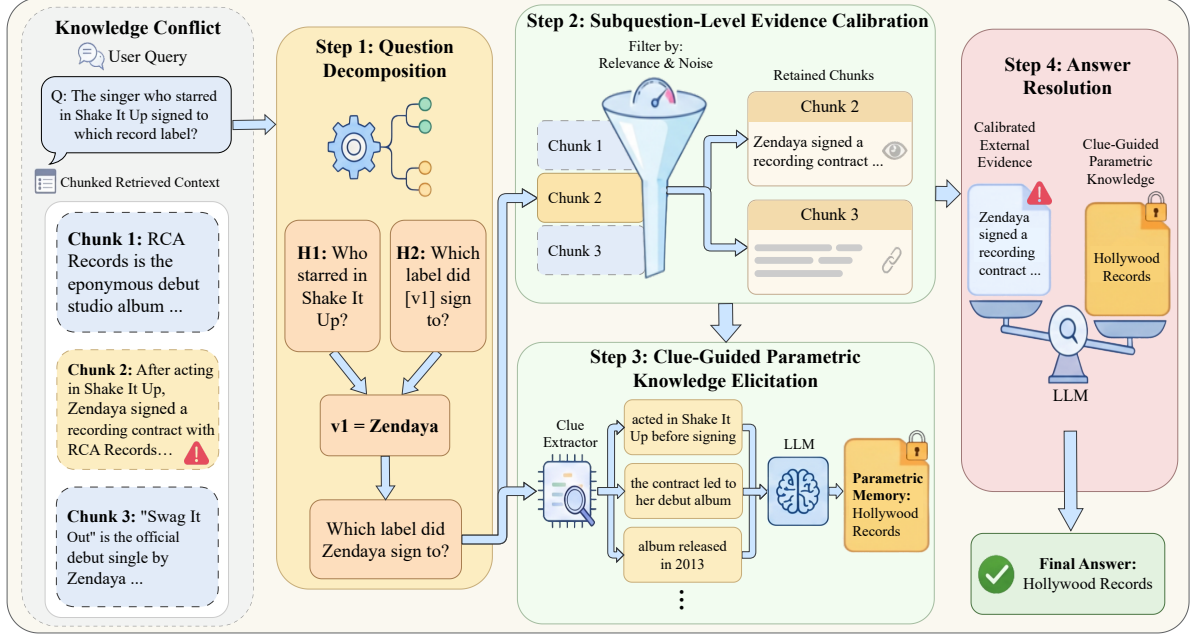


Figure 3: Overview of TRIAD. The framework jointly addresses the question-axis, context-axis, and memory-axis through question decomposition, subquestion-level evidence calibration, clue-guided parametric knowledge elicitation, and answer resolution for each hop.

For non-final hops, the predicted answer is written into the binding map:

$$B_t = B_{t-1} \cup \{v_t \mapsto a_t\}. \quad (10)$$

The resulting bindings are propagated to later hops through the instantiation step above. After the final hop, the model output is returned as the final answer:

$$\hat{y} = a_T. \quad (11)$$

These modules are therefore coupled rather than isolated. Decomposition defines the semantic target for calibration, calibrated evidence yields cleaner clue extraction, filtered clues sharpen parametric recall, and each resolved answer instantiates the next hop. Each step thus prepares a more targeted input for the next, amplifying its effect.

5 Experiments

5.1 Experiment Settings

Datasets and Metrics. Following the evaluation protocol of Micro-Act (Huo et al., 2025), we evaluate on the KRE versions of SQuAD and MuSiQue (Ying et al., 2024; Rajpurkar et al., 2018; Trivedi et al., 2022) together with ConflictBank (Su et al., 2024). We use 1,000 instances from each KRE dataset and 1,000 instances for each ConflictBank

conflict type. KRE-SQuAD mainly probes relatively direct conflict QA, KRE-MuSiQue emphasizes compositional multi-hop conflicts, and ConflictBank covers three types of misinformation, temporal, and semantic conflicts. All examples are multiple-choice questions paired with misleading context, and we report multiple-choice accuracy (ACC). We report Misled Rate in Appendix C, and additionally evaluate an open-ended QA variant in Appendix D. For TRIAD, the answer options are provided only at the final hop.

Baselines. We compare TRIAD against End-to-End QA, two GAMs: GKP (Liu et al., 2022) and Self-consistency (Wang et al., 2023), and three SAMs: COMBO (Zhang et al., 2023), Micro-Act (Huo et al., 2025), and ASTUTERAG (Wang et al., 2025). We additionally evaluate an ASTUTERAG w/ retrieve variant. Detailed baseline implementations are deferred to Appendix B.2.

Models. We use two closed-source models, GPT-4o and GPT-4o-mini (OpenAI, 2024b,a), plus four open-source models from the Llama family (Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct) (Meta, 2024) and the Qwen family (Qwen3.5-35B-A3B, Qwen3.5-9B) (Qwen Team, 2026). This selection spans both small and large parameter scales. For evidence calibration, we use Qwen3-Embedding-0.6B (Qwen Team, 2025), which offers strong mul-

Method / Dataset	SQuAD	MuSiQue	ConflictBank	SQuAD	MuSiQue	ConflictBank	SQuAD	MuSiQue	ConflictBank
	GPT-4o			GPT-4o-mini			Llama-3.1-70B-Instruct		
End-to-End QA	27.0	25.0	5.4	28.5	23.5	2.8	25.9	23.0	3.1
GKP	58.1	53.4	19.5	59.9	44.3	17.5	46.1	39.9	15.8
Self-consistency	40.2	33.8	1.9	29.0	28.5	2.1	33.7	32.4	3.6
COMBO	47.0	48.2	2.0	41.8	31.6	12.2	46.3	39.0	3.9
Micro-Act	63.6	53.1	22.3	60.3	52.0	26.9	62.9	48.3	26.5
ASTUTERAG	63.1	57.1	11.5	61.0	50.3	4.5	55.0	47.5	10.6
ASTUTERAG w/ retrieve	74.5	63.4	35.0	63.7	53.0	23.9	68.3	60.5	30.4
TRIAD	78.8	74.8	56.7	67.7	58.2	41.6	72.7	63.3	44.9
	Qwen3.5-35B-A3B			Llama-3.1-8B-Instruct			Qwen3.5-9B		
End-to-End QA	26.6	25.6	2.5	29.0	19.9	4.1	22.9	15.3	2.4
GKP	60.2	45.9	23.9	43.4	31.5	11.5	44.5	32.9	19.3
Self-consistency	39.1	37.7	5.6	26.2	24.1	3.2	31.7	29.6	4.0
COMBO	44.5	38.2	6.7	46.5	31.5	26.2	40.6	32.3	6.3
Micro-Act	46.4	33.7	33.9	55.2	38.5	18.3	43.4	24.6	25.6
ASTUTERAG	61.5	44.6	10.7	49.2	38.0	13.2	49.8	35.4	5.7
ASTUTERAG w/ retrieve	68.9	50.5	35.7	62.8	45.7	33.9	64.1	48.3	24.9
TRIAD	78.1	65.5	54.3	61.7	51.6	45.8	73.9	60.8	49.6

Table 1: Main results on SQuAD, MuSiQue, and ConflictBank. Unless otherwise specified, ASTUTERAG refers to the version without external retrieval. **Bold** numbers denote the best result among all methods. Underlined numbers denote the best result when excluding ASTUTERAG w/ retrieve, enabling a fair comparison without additional external retrieval.

tiling embedding quality with a lightweight footprint.

Implementation Details. For fairness and reproducibility, we set the decoding temperature to 0 for TRIAD and all baselines except Self-consistency, for which we use 0.5 to induce diverse sampled reasoning paths. All experiments with open-source models are conducted on 8 NVIDIA RTX A6000 GPUs, while closed-source models are accessed through API calls. For TRIAD, we set the two evidence-calibration weights to $\alpha = 0.6$ and $\beta = 0.4$ based on the overall performance across datasets. Detailed hyperparameter analyses and prompt specifications are provided in Appendix B.

5.2 Main Results

Table 1 reports the main results.

Observation 1. Excluding ASTUTERAG w/ retrieve for a fair comparison, TRIAD achieves the best result in all 18 model–benchmark combinations and surpasses the strongest non-retrieval baseline by 17.1 ACC points on average. TRIAD also outperforms ASTUTERAG w/ retrieve in 17 of 18 settings, trailing only on Llama-3.1-8B-Instruct over SQuAD by 1.1 points. These results suggest that, under misleading context, stronger generation-stage conflict resolution can be more effective than simply adding retrieved evidence.

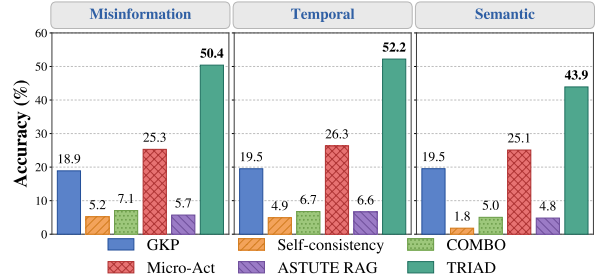


Figure 4: Performance across the three ConflictBank conflict types on Qwen3.5-9B.

Observation 2. TRIAD remains effective across backbone families and model scales. Averaged over the three benchmarks, it improves over the strongest non-retrieval baseline by 15.8/17.7 points on closed/open-source models and 18.6/15.6 points on larger/smaller models. This suggests that its gains do not depend on a specific backbone or scale.

Observation 3. TRIAD performs strongly on both the relatively direct conflicts in SQuAD and multi-hop conflicts in MuSiQue, improving over the strongest non-retrieval baseline by 13.2 and 16.2 points on average. The gain is even larger on ConflictBank (21.9 points), highlighting TRIAD’s strong ability to resolve conflicts in longer, more evidence-like contexts where the misleading signal is deeply hidden. Figure 4 further breaks down

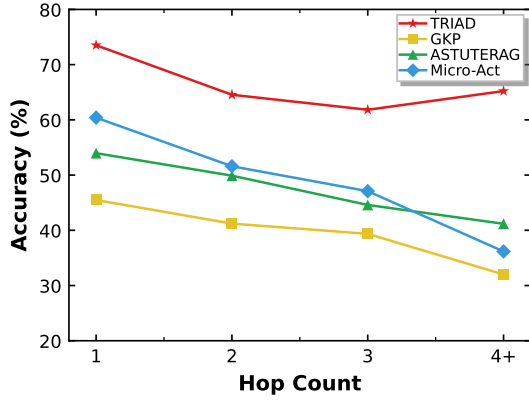


Figure 5: Accuracy by question hop count on Llama-3.1-70B-Instruct, aggregated over KRE-SQuAD and KRE-MuSiQue.

accuracy by the three ConflictBank conflict types on Qwen3.5-9B, with TRIAD performing best on misinformation, temporal, and semantic conflicts (50.4/52.2/43.9). This indicates that TRIAD generalizes across both question complexity and conflict type. We also examine generalization to misleading contexts with different rhetorical tones and to settings where correct and misleading evidence co-occur in the same retrieved context. Detailed results are provided in Appendix E.

6 Analysis

In this section, we analyze the effectiveness of TRIAD and provide additional robustness, cost, case study, and error analyses in Appendix F, Appendix G, Appendix H, and Appendix I, respectively.

6.1 Why TRIAD Is Effective

Accuracy by Question Hop Count. As shown in Figure 5, TRIAD achieves the best performance in every hop bucket. Moreover, TRIAD remains much more stable as hop count increases. From 1 to 4+ hops, TRIAD loses only 8.3 points, whereas Micro-Act and ASTUTERAG lose 24.2 and 12.8 points, respectively. This pattern suggests that TRIAD’s gains come from structuring conflict resolution around the local semantics of each hop.

Accuracy by Context Length. Figure 6 shows that accuracy generally declines as misleading context length grows, while TRIAD remains more robust than the baselines. As distractors accumulate, conflict resolution becomes harder unless the model can control which context enters later reasoning. Figure 7 provides direct evidence for

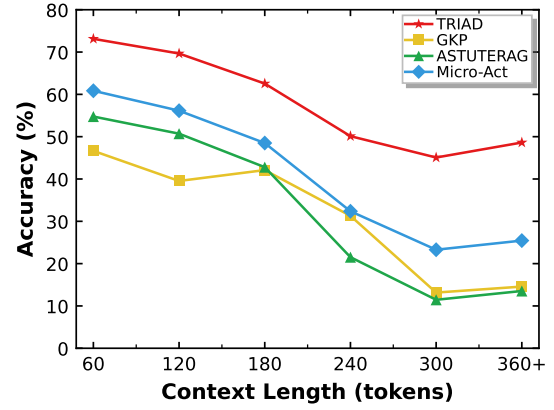


Figure 6: Accuracy by misleading-context length on Llama-3.1-70B-Instruct.

this mechanism. Across all processed examples, TRIAD keeps fewer tokens than the original context in 84.8% of cases and retains only 26.1% of the original tokens overall. This compression becomes more aggressive as conflict becomes harder. Yet TRIAD still preserves higher accuracy than the baselines in these long context examples, supporting our interpretation of evidence calibration as filtering guided by sufficiency rather than simple truncation.

We also observe a mild uptick in the rightmost 360+ bin, but we do not interpret this as a reversal of the overall trend. The last two bins are dominated by ConflictBank examples, some of which are longer because they contain more verbose narrative scaffolding rather than deeper conflict structure. Since the same pattern appears for all four methods, we view it as a modest binning and sampling effect that does not change the main conclusion.

Improved Relevance of Parametric Knowledge.

Figure 8 shows that TRIAD improves the relevance of elicited parametric knowledge. With clue extraction, High relevance exceeds Low relevance at 49.6 versus 38.5. This suggests that TRIAD improves conflict resolution by shifting off-target recall toward knowledge aligned with the question.

This shift is important because the analysis in Section 3 already shows that lower parametric knowledge relevance is associated with lower accuracy. The larger share of High relevance cases under TRIAD therefore provides a direct explanation for its stronger performance. In other words, TRIAD improves conflict resolution not by eliciting more parametric knowledge, but by eliciting parametric knowledge that is more relevant to the question.

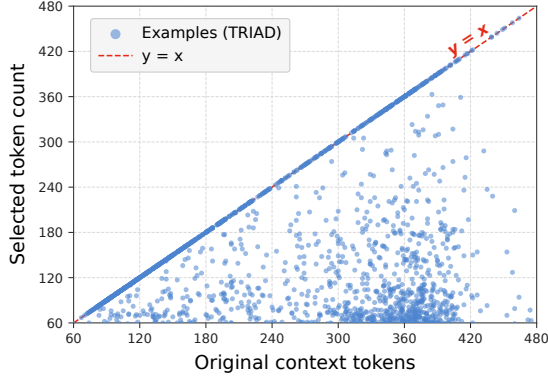


Figure 7: Selected context length versus original context length for TRIAD on Llama-3.1-70B-Instruct. Each point is an example.

Coupled System Design. TRIAD’s modules are also effective because they are tightly coupled rather than arranged as a simple sequential pipeline. Question decomposition defines the semantic target for evidence calibration at each hop, while calibrated evidence provides cleaner indirect clues for parametric knowledge elicitation. To directly test this coupling, we evaluate two decoupled variants in Table 3. The first compresses the context only once using the original question and reuses the same compressed context for all subquestions. The second still uses calibrated evidence for final conflict resolution, but extracts clues from the original uncompressed context.

The first variant mainly hurts MuSiQue, where accuracy drops by 6.6 points, supporting the role of subquestion-specific calibration for multi-hop reasoning. The second variant mainly hurts ConflictBank, where accuracy drops by 8.5 points, showing that clue extraction benefits from reduced exposure to redundant or distracting signals in long misleading contexts. These results indicate that TRIAD’s gains come not only from individual modules, but also from the way decomposition, calibration, and clue-guided recall constrain each other during inference.

6.2 How Extracted Clues Affect Performance

Influence of Clue Count. Detailed statistics are reported in Table 4. Clue extraction remains stable across datasets without any fixed clue budget during inference. The average ranges from 3.14 to 3.35, and the median is 3.0 on all three benchmarks, suggesting that TRIAD does not rely on generating many clues to improve performance and instead naturally converges to a compact clue set. The re-

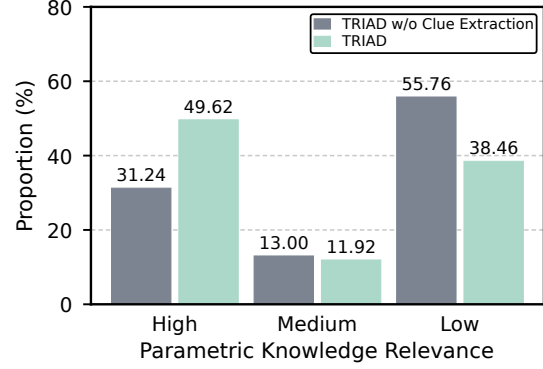


Figure 8: Distribution of parametric knowledge relevance for TRIAD and TRIAD without clue extraction.

Method	KRE-SQuAD	KRE-MuSiQue	Avg.
GKP	94.6	87.4	91.0
Self-consistency	99.3	92.2	95.8
COMBO	94.5	83.7	89.1
ASTUTERAG	88.6	81.9	85.3
Micro-Act	95.3	90.4	92.9
TRIAD	91.2	85.8	88.5

Table 2: Accuracy in the conflict-free setting.

sult is especially notable on ConflictBank, where long misleading contexts are condensed into only a few clues while still preserving strong accuracy, supporting our view of clue extraction as a semantic compression step that distills key anchors for later parametric verification.

Influence of Error Clues. To evaluate robustness to clue extraction errors, we construct a worst-case setting by shuffling context fields in SQuAD and MuSiQue to produce completely irrelevant context, results are reported in Table 5. In this regime, GAMs such as Self-consistency and GKP perform best overall, which is expected because they rely less on contextual details and can fall back more directly on parametric knowledge. By contrast, SAMs that reason more deeply over contextual details perform worse in this setting. Even so, TRIAD remains the most robust SAM on both datasets, suggesting that its mechanisms prevent complete failure when clue extraction is corrupted.

6.3 Conflict-Free Robustness

We further evaluate TRIAD under conflict-free conditions, where retrieved context provides valid supporting evidence rather than misleading evidence. As shown in Table 2, TRIAD is weaker than the strongest GAM baseline in this setting. This re-

flects the tradeoff discussed in the Limitations section: evidence calibration is designed to suppress misleading details under conflict, but the same selectivity can also remove useful supporting information when the context is already reliable.

At the same time, TRIAD remains within the range of other source-aware methods, and its main benefit appears when misleading evidence is present. In the main Qwen3.5-9B results, TRIAD improves substantially over the strongest non-retrieval baseline under conflict. These results suggest that conflict-free robustness and conflict resolution involve a tradeoff rather than a contradiction.

6.4 Ablation Study

Table 6 reports the ablation results on Qwen3.5-9B. First, question decomposition is particularly useful for multi-hop reasoning. Removing it lowers accuracy by 3.8 points on MuSiQue. By contrast, accuracy on SQuAD and ConflictBank changes only marginally, since these datasets are more often dominated by single-hop questions, so the benefit of decomposition is much less pronounced. Second, the benefit of evidence calibration becomes more pronounced as context complexity increases. Its removal causes a 1.5 point drop on SQuAD, a 3.3 point drop on MuSiQue, and a much larger 10.1 point drop on ConflictBank. This trend is consistent with the increasing difficulty of filtering longer and more misleading context. Third, indirect clue extraction is the most critical component overall. Removing it reduces accuracy by 23.6 points on average, showing that clue extraction is the main mechanism that steers parametric recall toward question-relevant knowledge under conflict.

7 Conclusion

This paper studies knowledge conflict resolution in RAG. We identify three coupled failure axes behind the weakness of existing methods. To address them jointly, we propose TRIAD, a training-free framework that combines question decomposition, subquestion-level evidence calibration, clue-guided parametric knowledge elicitation, and answer resolution.

Experiments on multiple settings show that TRIAD consistently outperforms strong baselines and remains robust under both multi-hop reasoning and long misleading context. Overall, these results suggest that effective knowledge conflict

resolution should be treated as a joint problem over question structure, external evidence, and parametric knowledge, rather than as source comparison alone.

Limitations

Although our method is effective for resolving knowledge conflict, its performance is somewhat weaker in non-conflict settings with golden context. A straightforward explanation is that the same calibration mechanism that suppresses misleading details in conflict settings can also compress supportive context in non-conflict settings, thereby discarding pieces of correct evidence that remain important for answer resolution. Future work should make evidence calibration more selective, so that it can filter misleading details while preserving supportive information.

To preserve the cost-effectiveness of TRIAD, we avoid overly complex strategies for question decomposition and context segmentation. This simplicity also creates a limitation, since the quality of decomposition depends on model capability. Weaker models may fail to capture the latent logic of multi-hop questions and fail to decompose them correctly. Similarly, our context chunking strategy simply splits external context by periods, which may lose the coherence of the context. Nevertheless, TRIAD still achieves strong overall performance with a favorable cost-effectiveness profile.

Ethics Statement

Our work uses publicly available datasets and does not involve the collection of personally identifiable information. During data collection and preprocessing, we carefully filtered out examples containing sensitive, biased, or potentially harmful content. All datasets and models are used in accordance with their intended purposes and respective licenses. Our goal is to improve the reliability of Retrieval-augmented generation under misleading external evidence, rather than to support fact manipulation or other harmful uses.

Use of AI Assistants

We used general-purpose AI assistants for language polishing and limited productivity support under author supervision. They were not used to generate research ideas, design experiments, make scientific decisions, or produce the paper’s technical claims,

results, or conclusions. All technical content, experimental results, and conclusions are the authors' own, and the authors take full responsibility for the correctness of the paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFB3003700) and the National Natural Science Foundation of China (Grant No. 62476025).

References

- Paul J. L. Ammann, Jonas Golde, and Alan Akbik. 2025. [Question decomposition for retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 497–507, Vienna, Austria. Association for Computational Linguistics.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). In *International Conference on Learning Representations*, volume 2024, pages 9112–9141.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33.
- Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. [Conflict-aware soft prompting for retrieval-augmented generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26981–26995, Suzhou, China. Association for Computational Linguistics.
- Nan Huo, Jinyang Li, Bowen Qin, Ge Qu, Xiaolong Li, Xiaodong Li, Chenhao Ma, and Reynold Cheng. 2025. [Micro-act: Mitigate knowledge conflict in question answering via actionable self-reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18550–18574, Vienna, Austria. Association for Computational Linguistics.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. [Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33.
- Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lv. 2026. [Resisting contextual interference in RAG via parametric-knowledge reinforcement](#). In *International Conference on Learning Representations*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. [Generated knowledge prompting for commonsense reasoning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Meta. 2024. Meta llama 3.1 instruct model cards. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>; <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>. Official model cards for the Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct models. Accessed: 2026-04-20.

- Philhoon Oh and James Thorne. 2023. [Detrimental contexts in open-domain question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11589–11605, Singapore. Association for Computational Linguistics.
- OpenAI. 2024a. Gpt-4o mini model. <https://developers.openai.com/api/docs/models/gpt-4o-mini>. Accessed: 2026-04-20.
- OpenAI. 2024b. Gpt-4o system card. <https://cdn.openai.com/gpt-4o-system-card.pdf>. Accessed: 2026-04-20.
- Qwen Team. 2025. Qwen3-embedding-0.6b model card. <https://huggingface.co/Qwen/Qwen3-Embedding-0.6B>. Accessed: 2026-04-20.
- Qwen Team. 2026. Qwen3.5 model cards. <https://huggingface.co/Qwen/Qwen3.5-9B>; <https://huggingface.co/Qwen/Qwen3.5-35B-A3B>. Official model cards for the Qwen3.5-9B and Qwen3.5-35B-A3B models. Accessed: 2026-04-20.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. 2024. [Assessing “implicit” retrieval robustness of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8988–9003, Miami, Florida, USA. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. [Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in LLM](#). *Preprint*, arXiv:2408.12076.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [MuSiQue: Multi-hop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. 2025. [Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyang Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024a. [The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. [Knowledge conflicts for LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8522–8551, Miami, Florida, USA. Association for Computational Linguistics.
- Jiahao Ying, Yixin Cao, Kai Xiong, Yidong He, Long Cui, and Yongbin Liu. 2024. [Intuitive or dependent? investigating LLMs’ behavior style to conflicting prompts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246, Bangkok, Thailand. Association for Computational Linguistics.
- Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025. [FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria. Association for Computational Linguistics.
- Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. [Merging generated and retrieved knowledge for open-domain QA](#). In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 4710–4728, Singapore. Association for Computational Linguistics.

Algorithm 1: TRIAD**Input:** original question q , retrieved context $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$.**Output:** final answer $\hat{y}$.1. Compute the ordered hop sequence $D(q) = \{h_1, h_2, \dots, h_T\}$.2. Initialize the binding map $B_0 \leftarrow \emptyset$.3. **for** $t = 1, 2, \dots, T$ **do** 3.1 Instantiate the current hop $\tilde{h}_t \leftarrow \psi(h_t, B_{t-1})$. 3.2 Split the retrieved context into chunks and compute $\text{SNAM}(\tilde{h}_t, c_j)$ for each chunk c_j . 3.3 Rank chunks by SNAM, accumulate them incrementally, and stop once the selected evidence is sufficient; denote the result by $\hat{\mathcal{C}}_t$. 3.4 Extract filtered indirect clues Z_t from $\hat{\mathcal{C}}_t$, excluding direct answer signals. 3.5 Elicit clue-guided parametric knowledge $K_t \leftarrow \text{Elicit}(\tilde{h}_t, Z_t)$. 3.6 Resolve the current hop answer $a_t \leftarrow \text{Resolve}(\tilde{h}_t, \hat{\mathcal{C}}_t, K_t)$. 3.7 **if** $t < T$ **then** update $B_t \leftarrow B_{t-1} \cup \{v_t \mapsto a_t\}$.4. **return** $\hat{y} \leftarrow a_T$.

Method	SQuAD	MuSiQue	ConflictBank	Avg.
TRIAD	73.9	60.8	49.6	61.4
Decouple decomposition $\rightarrow$ calibration	73.2 ($\downarrow 0.7$)	54.2 ($\downarrow 6.6$)	49.4 ($\downarrow 0.2$)	58.9 ($\downarrow 2.5$)
Decouple calibration $\rightarrow$ clue extraction	72.1 ($\downarrow 1.8$)	58.4 ($\downarrow 2.4$)	41.1 ($\downarrow 8.5$)	57.2 ($\downarrow 4.2$)

Table 3: Accuracy of decoupled TRIAD variants on Qwen3.5-9B.

Dataset	ACC	Avg.	Median
KRE-SQuAD	73.9	3.14	3.0
KRE-MuSiQue	60.8	3.35	3.0
ConflictBank	49.6	3.16	3.0

Table 4: Qwen3.5-9B accuracy and clue-count statistics for TRIAD. For multi-hop questions, we average the clue count across hops.

Dataset	End-to-End QA	GKP	Self-cons.	COMBO	Micro-Act	ASTUTERAG	TRIAD
SQuAD-irrelevant	70.5	75.3	76.6	73.1	65.2	70.6	73.7
MuSiQue-irrelevant	50.3	60.4	63.4	52.3	35.9	53.1	53.2

Table 5: Worst-case accuracy on Qwen3.5-9B under completely irrelevant context.

Method	SQuAD	MuSiQue	ConflictBank	Avg.
TRIAD	73.9	60.8	49.6	61.4
w/o Question Decomposition	73.2 ($\downarrow 0.7$)	57.0 ($\downarrow 3.8$)	48.8 ($\downarrow 0.8$)	59.7 ($\downarrow 1.7$)
w/o Evidence Calibration	72.4 ($\downarrow 1.5$)	57.5 ($\downarrow 3.3$)	39.5 ($\downarrow 10.1$)	56.5 ($\downarrow 4.9$)
w/o Indirect Clue Extraction	49.0 ($\downarrow 24.9$)	40.0 ($\downarrow 20.8$)	24.5 ($\downarrow 25.1$)	37.8 ($\downarrow 23.6$)

Table 6: Ablation results on Qwen3.5-9B. The TRIAD row uses the main-table result. The ablation variants are recomputed with the total number of samples as the denominator. Numbers in parentheses denote the drop from TRIAD in percentage points. SQuAD and MuSiQue each contain 1,000 examples, and ConflictBank contains 3,000 conflict instances.

A Experimental Setup for Failure-Axis Analysis

All preliminary analyses use the same baseline pool unless otherwise noted. We analyze five training-free methods: GKP, Self-consistency, COMBO, Micro-Act, and ASTUTERAG. GKP and Self-consistency are grouped as GAMs, while COMBO, Micro-Act, and ASTUTERAG are grouped as SAMs. We use Llama-3.1-70B-Instruct and Qwen3.5-9B as the two analysis backbones, as both are widely adopted open-source instruction-tuned models while also covering different parameter scales. We use the KRE versions of SQuAD and MuSiQue, which are standard benchmarks for knowledge conflict evaluation and together cover both relatively direct and more compositional question settings under the same conflict construction. Unless otherwise specified, family curves are computed by first obtaining method-level statistics and then averaging within GAMs or SAMs.

A.1 Setup for Question Complexity Analysis

Hop counts are not taken from baseline outputs, as different models can vary substantially in their ability to decompose multi-hop questions. Instead, we adopt a unified GPT-4o decomposition standard, with additional manual verification, to determine question hop counts consistently across methods. We report accuracy as the fraction of correct final predictions within each hop bin. Our interpretation focuses on bins with sufficient sample support.

A.2 Setup for Parametric Knowledge Analysis

Parametric knowledge relevance is annotated at the instance level with three labels: High, Medium, and Low. To balance annotation quality and cost, we use GPT-4o-mini to label the relevance of the parametric knowledge elicited by each input question, using the question, the true answer, and the golden context as references for determining whether the elicited parametric knowledge is well aligned with the intended semantics. We further conduct manual verification on a random sample, which yields a labeling accuracy of 89.2%. Human verifiers followed the same label definitions and instructions shown in Table 7. The prompt template used for this annotation is shown in Table 7. These labels therefore capture both semantic alignment with the question and whether the elicited knowledge directly contains the true answer entity or answer phrase. We report answer accuracy under each rel-

Prompt
Task: You are annotating the relevance of a model-generated parametric knowledge passage. Judge only the parametric knowledge itself, using the question, the true answer, and the golden context as reference. Label definitions: High: The parametric knowledge is clearly relevant to the question, aligns with the golden context, and directly contains the true answer entity or answer phrase. Medium: The parametric knowledge is relevant to the question and aligned with the golden context, but it does not directly contain the true answer entity or answer phrase. Low: The parametric knowledge has little substantive relation to the question or the golden context, is off-target, or is largely inconsistent with the golden context. Question: {question} True Answer: {true_answer} Golden Context: {golden_context} Parametric Knowledge: {parametric_knowledge} Output exactly one label: High, Medium, or Low.

Table 7: Prompt used for GPT-4o-mini annotation of parametric knowledge relevance.

evance label and aggregate family curves in the same manner as above.

A.3 Setup for Misleading Context Analysis

We measure context length by tokenizing the misleading retrieved context. We group instances into six coarse displayed ranges: [60, 120), [120, 180), [180, 240), [240, 300), [300, 360), and [360, +∞). This grouping is chosen to ensure adequate sample support within each range, allowing us to examine how the metrics evolve as misleading context becomes longer. We analyze two complementary metrics. Accuracy measures final answer correctness. Misled rate measures the proportion of predictions that exactly match the incorrect answer label embedded in the dataset’s misleading context construction, which directly captures whether a method follows the misleading context.

A.4 Setup for Coupled-Effect Analysis

For the joint analysis, we pool all five baselines and both backbones on SQuAD+MuSiQue and convert the three axes into binary indicators:

$$Q = \begin{cases} 1, & \text{if hop count} > 1, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

Thus, Q marks cases where resolving the conflict requires multi-hop question understanding rather

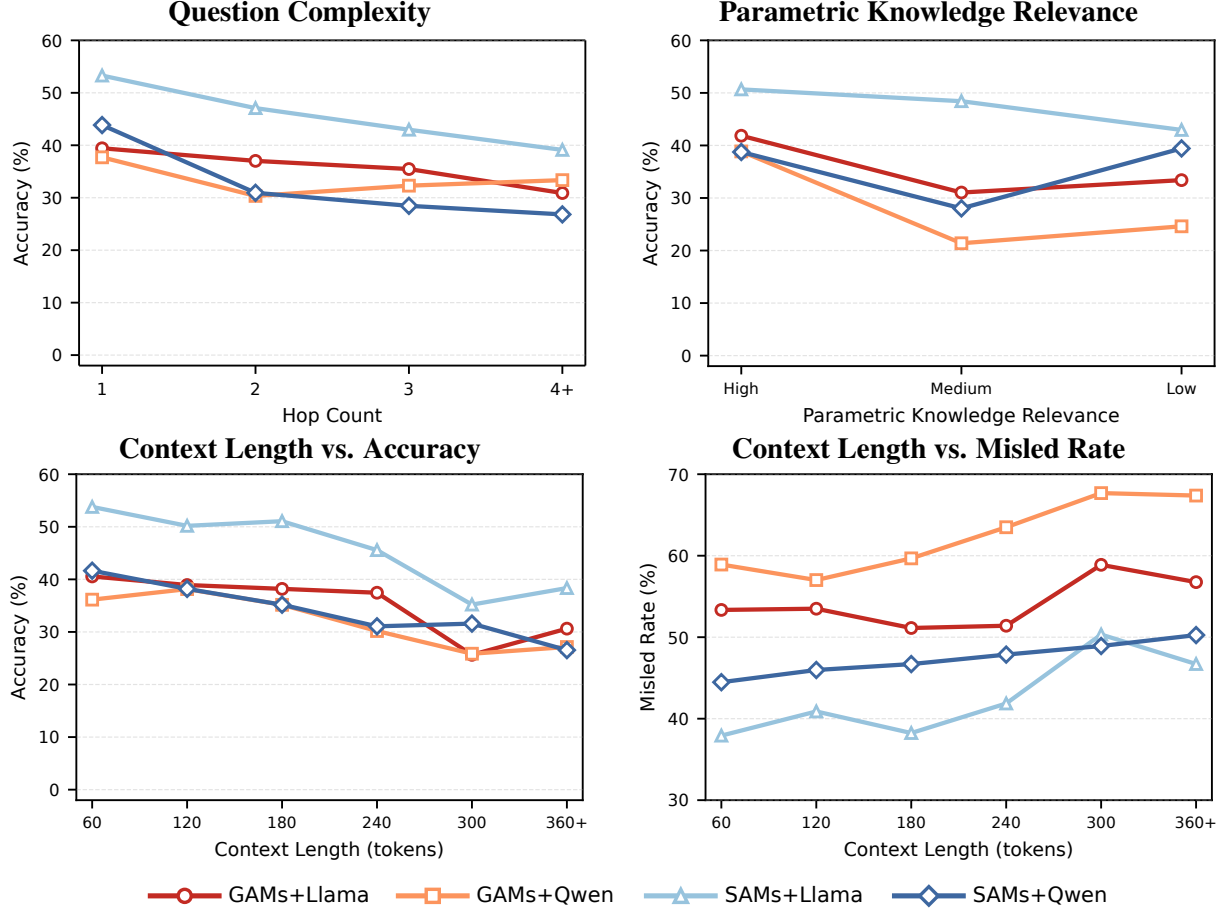


Figure 9: Preliminary analysis of existing training-free baselines with respect to the three failure axes.

than a single local inference.

$$M = \begin{cases} 1, & \text{if Low,} \\ 0, & \text{if High or Medium.} \end{cases} \quad (13)$$

Accordingly, M isolates the setting in which elicited parametric knowledge is clearly off-target with respect to the question intent.

$$\bar{L}_{\text{ctx}} = \frac{\sum_g n_g \mu_g}{\sum_g n_g}, \quad (14)$$

where g indexes the context-length groups, n_g is the number of samples in group g , and μ_g is the average context length of that group. Following the context-length grouping used above, this weighted average threshold is 182.7 in our setting.

$$C = \begin{cases} 1, & \text{if context length} > \bar{L}_{\text{ctx}}, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

In this way, C identifies instances whose misleading context is longer than the pooled baseline level and therefore more likely to exert stronger surface evidential influence.

These binarization rules are chosen not only for interpretability, but also to keep the resulting groups reasonably balanced, so that comparisons across the eight combinations are backed by sufficiently large sample sizes. We then compute group accuracy for the eight combinations {None, Q, C, M, C+Q, Q+M, M+C, M+C+Q}. This pooled analysis is intended to expose interaction patterns across axes rather than to compare individual baselines.

B Experiment Details

B.1 Datasets and Metrics

KRE-SQuAD. KRE-SQuAD is the conflict-oriented version of SQuAD used by prior knowledge-conflict work (Ying et al., 2024). In our released split, each row contains *question*, *choices*, *answer*, *golden_context*, *negative_context*, and *negative_answer*. We use the provided *negative_context* as the misleading retrieved context, *answer* as the gold label, and *negative_answer* as the misleading label. This dataset mainly contains relatively direct factual questions, making it use-

ful for testing conflict resolution when question structure is simple but the external evidence is misleading.

KRE-MuSiQue. KRE-MuSiQue uses the same conflict construction style as KRE-SQuAD, but is built on MuSiQue (Trivedi et al., 2022). Its rows follow the same schema as KRE-SQuAD, namely *question*, *choices*, *answer*, *golden_context*, *negative_context*, and *negative_answer*. Compared with KRE-SQuAD, it contains substantially more compositional and multi-hop questions, and therefore stresses question understanding and intermediate constraint tracking under knowledge conflict.

ConflictBank. ConflictBank is a large-scale benchmark for evaluating knowledge conflicts caused by different evidence types (Su et al., 2024). Each row provides a multiple-choice question together with three conflict variants. The misinformation split replaces the correct fact with a false but plausible statement. The temporal split introduces time-sensitive discrepancies, such as outdated or temporally inconsistent evidence. The semantic split preserves topical relatedness while altering the meaning in a way that supports the wrong option. Following Micro-Act, we sample 1,000 instances for each conflict type.

Combining these subsets yields 5,000 evaluation instances overall. In all three datasets, the misleading context is constructed so that it supports a specific distractor option, namely *negative_answer* in KRE and *replace_option* in ConflictBank.

We use multiple-choice accuracy (ACC) as the primary metric. ACC is computed as exact match between the predicted option label and the gold option label. This choice matches the native format of KRE and ConflictBank (Ying et al., 2024; Su et al., 2024) and, more importantly, directly evaluates whether a method resists the distractor supported by the misleading context. As also noted by Micro-Act, using multiple-choice ACC avoids additional normalization ambiguity from free-form answer matching and keeps the evaluation focused on conflict resolution itself.

B.2 Baselines

We compare TRIAD against six baselines. End-to-End QA is a simple direct-answer reference. The remaining baselines include two GAMs, GKP and Self-consistency, and three SAMs, COMBO, Micro-Act, and ASTUTERAG. We additionally evaluate an ASTUTERAG w/ retrieve variant.

End-to-End QA. This baseline directly answers the question from the provided misleading context in a single prompt. Our implementation takes as input the question, the answer choices, and the misleading context, and asks the model to return only a JSON object with the predicted option label. It does not explicitly separate external evidence from parametric knowledge, and therefore serves as a no-resolution reference point.

GKP. GKP implementation follows the two-stage structure of Generated Knowledge Prompting (Liu et al., 2022). In the first stage, the model generates a short internal knowledge statement conditioned only on the question. In the second stage, the same model answers the multiple-choice question using both the generated knowledge and the provided misleading context. We adapt GKP to our fixed context conflict setting by keeping the pipeline fully prompt-based and by using the same answer format as the other baselines.

Self-consistency. Self-consistency implementation follows the standard idea of sampling multiple reasoning paths and aggregating them by voting (Wang et al., 2023). Each run uses a single-pass prompt that asks the model to output recalled knowledge, step-by-step reasoning, a final option label, and a confidence score. We collect five outputs per question and select the final answer by majority vote. When there is a vote tie, we break ties by higher average confidence. To enable meaningful diversity across sampled reasoning paths, we set the decoding temperature to 0.5 for this baseline. This setting therefore tests whether stochastic repeated reasoning and aggregation can improve robustness under misleading context.

COMBO. The original COMBO method merges generated and retrieved evidence through compatibility-oriented matching (Zhang et al., 2023). Our implementation preserves this core idea, but we adapt it to our setting for two reasons. First, our benchmarks provide a fixed misleading context rather than a standard retrieved candidate pool. Second, we keep all comparisons training-free and do not introduce additional learned compatibility modules. Concretely, our prompt-only reproduction proceeds in three stages. It first generates three short parametric knowledge passages from the question. It then splits the provided misleading context into at most six passages and asks the same LLM to

judge, for every generated-context pair, whether the context is evidential for the question, whether it is compatible with the generated passage, and how strong the compatibility is. We then solve an optimal bipartite matching problem over these scores and answer the question using the positively matched pairs. This design keeps COMBO’s compatibility-oriented merging principle while making it directly comparable to the other baselines in our conflict setting.

Micro-Act. We use the released Micro-Act pipeline (Huo et al., 2025) with its hierarchical action-reasoning workflow. The key idea of Micro-Act is not to compare retrieved evidence and parametric knowledge in one global step, but to convert conflict resolution into a sequence of fine-grained actions. This design lets the model progressively inspect where the two sources disagree, decompose broad claims into smaller comparison units, and verify local conflict points before making the final decision.

ASTUTERAG. We reproduce ASTUTERAG (Wang et al., 2025) with its source-aware generation-consolidation-answer pipeline. In the default setting, the misleading dataset context is treated as the retrieved source. The model runs up to two rounds of internal passage generation, with up to three new passages per round, and after each generation stage performs up to two consolidation rounds to form a source-aware knowledge state containing salient facts and explicit conflicts. The final answer is then produced from this consolidated state.

We additionally implement an ASTUTERAG *with_retrieve* variant. For this variant, we construct offline retrieved datasets using the Tavily API. For each question, we use the question text as the retrieval query, keep up to 10 search results, run Tavily Extract on the returned URLs, select the best extracted passage from each accessible page, and concatenate the resulting external passages with the original misleading context. This differs from the original ASTUTE RAG paper, which uses Google Search and webpage paragraph extraction. We make this substitution primarily for cost reasons, since large-scale Google-based retrieval would introduce substantially higher API expenses. We keep the default ASTUTERAG and ASTUTERAG *with_retrieve* results separate throughout the experiments.

B.3 Models

Backbone LLMs. We evaluate all methods on six LLMs. The closed-source models are GPT-4o and GPT-4o-mini (OpenAI, 2024b,a). The open-source models are Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct (Meta, 2024), and Qwen3.5-35B-A3B, Qwen3.5-9B (Qwen Team, 2026). The GPT models are accessed through OpenAI-compatible APIs, while the open-source models are loaded from Hugging Face checkpoints.¹

Embedding Model. For TRIAD’s semantic relevance scoring, we use Qwen3-Embedding-0.6B (Qwen Team, 2025). We choose this model because it combines strong multilingual retrieval performance, instruction-aware embeddings, and a relatively small computational footprint, which makes it suitable for scoring large numbers of subquestion-chunk pairs.

B.4 Implementation Details

Experimental Parameters. To ensure fairness and reproducibility, we set the decoding temperature to 0 for TRIAD and all baselines except Self-consistency, for which we use 0.5 to induce diverse sampled reasoning paths. We also standardize the maximum generation length to 1,024 tokens across methods.

Details of Hyperparameter Experiments. The evidence calibration module combines semantic relevance from Qwen3-Embedding-0.6B with a noise penalty through two weights, α and β . To analyze sensitivity to these weights, we conduct a sweep with Llama-3.1-8B-Instruct by varying α from 0.1 to 0.9 with $\beta = 1 - \alpha$, and we additionally evaluate the boundary setting $(\alpha, \beta) = (1.0, 0.0)$. For efficiency, we use separate sampled subsets for hyperparameter analysis, namely 500 examples from KRE-SQuAD and KRE-MuSiQue, and 500 examples for each ConflictBank conflict type. These subsets are distinct from the evaluation instances used for the main results in the paper, so the final choice of α and β is not obtained by directly selecting the best setting on the full reported test benchmarks. Table 8 reports the resulting ACC values. The results indicate that TRIAD is reasonably stable across a broad range of settings. Based on the Avg. column, we use $\alpha = 0.6$ and

¹Representative Hugging Face model cards for the two open-source families are <https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct> and <https://huggingface.co/Qwen/Qwen3.5-9B>.

(α, β)	SQuAD	MuSiQue	ConflictBank	Avg.
(0.1, 0.9)	61.3	52.4	46.5	53.4
(0.2, 0.8)	60.7	52.1	45.9	52.9
(0.3, 0.7)	61.2	51.3	46.9	53.1
(0.4, 0.6)	60.9	52.4	46.6	53.3
(0.5, 0.5)	60.7	52.6	45.1	52.8
(0.6, 0.4)	60.9	52.8	47.1	53.6
(0.7, 0.3)	58.1	52.9	45.2	52.1
(0.8, 0.2)	60.4	50.5	46.1	52.3
(0.9, 0.1)	56.9	49.2	45.8	50.6
(1.0, 0.0)	63.6	42.3	47.0	51.0

Table 8: Hyperparameter sensitivity of TRIAD on Llama-3.1-8B-Instruct. Results are reported as ACC (%). Avg. denotes the mean ACC over the three datasets.

$\beta = 0.4$ in the main experiments, as this setting yields the strongest overall average performance while also giving the best result on ConflictBank. Notably, removing the noise penalty entirely with $(\alpha, \beta) = (1.0, 0.0)$ slightly improves SQuAD to 63.6 ACC, but causes a substantial drop on the more compositional MuSiQue benchmark (42.3 ACC), reducing the overall average to 51.0. This suggests that semantic relevance alone can be sufficient for some relatively direct conflict questions, whereas explicit noise suppression is important for more compositional settings.

B.5 Prompt Templates of TRIAD

This subsection presents the prompt templates used by TRIAD. Figures 12–15 illustrate the prompts for question decomposition, evidence calibration, clue extraction, parametric knowledge elicitation, and hop-wise answer resolution. For Steps 3–5, we show the final-hop variants that explicitly include answer choices.

C Misled Rate Results

Misled Rate measures the probability that a method selects the option supported by misleading context. Lower is better. The results are shown in Table 9.

The pattern is broadly consistent with the ACC results. Excluding ASTUTERAG w/ retrieve, TRIAD achieves the lowest Misled Rate in 17 of 18 model benchmark settings. The only exception is Qwen3.5-35B-A3B on ConflictBank, where GKP reaches 19.0. When ASTUTERAG w/ retrieve is included, TRIAD remains lowest in 16 of 18 settings, with one further exception on

Method	Dataset					
	SQuAD	MuSiQue	ConflictBank	SQuAD	MuSiQue	ConflictBank
	GPT-4o			GPT-4o-mini		
End-to-End QA	70.2	71.0	92.5	70.7	72.3	93.8
GKP	37.4	40.4	69.9	28.8	36.1	55.2
Self-consistency	57.5	61.2	97.3	69.0	65.5	97.0
COMBO	40.5	32.7	79.0	45.4	48.6	67.8
Micro-Act	32.3	37.5	-	29.6	34.0	57.3
ASTUTERAG	27.7	30.1	75.0	30.0	26.1	85.0
ASTUTERAG w/ retrieve	19.4	28.0	49.1	25.8	21.6	55.6
TRIAD	10.6	11.7	32.5	15.1	20.1	45.1
Method	Llama-3.1-70B-Instruct			Qwen3.5-35B-A3B		
	SQuAD	MuSiQue	ConflictBank	SQuAD	MuSiQue	ConflictBank
	Llama-3.1-8B-Instruct			Qwen3.5-9B		
End-to-End QA	72.2	72.3	92.9	72.1	69.4	92.6
GKP	45.0	41.4	82.3	34.4	45.0	19.0
Self-consistency	64.3	62.9	92.2	59.4	58.0	91.7
COMBO	46.4	44.0	77.6	47.1	43.5	75.6
Micro-Act	28.3	39.2	51.2	32.1	36.9	28.2
ASTUTERAG	39.2	44.2	69.7	31.7	42.1	31.8
ASTUTERAG w/ retrieve	25.7	32.1	45.1	24.0	36.6	20.8
TRIAD	14.4	21.6	46.0	11.2	17.9	26.2
Method	Llama-3.1-8B-Instruct			Qwen3.5-9B		
	SQuAD	MuSiQue	ConflictBank	SQuAD	MuSiQue	ConflictBank
	Llama-3.1-8B-Instruct			Qwen3.5-9B		
End-to-End QA	67.6	71.6	92.1	75.8	79.2	94.7
GKP	36.9	41.8	58.1	49.4	58.3	48.6
Self-consistency	70.6	68.1	94.1	66.9	65.8	93.5
COMBO	39.1	45.7	53.9	52.2	48.8	80.6
Micro-Act	34.9	45.7	66.6	35.9	46.1	39.8
ASTUTERAG	39.1	42.7	72.5	43.7	49.2	70.8
ASTUTERAG w/ retrieve	26.7	29.1	41.3	31.2	35.0	42.5
TRIAD	19.8	21.7	40.2	13.4	19.2	35.2

Table 9: Misled Rate results with the same model grouping as Table 1. **Bold** numbers denote the lowest result among all methods. Underlined numbers denote the lowest result when excluding ASTUTERAG w/ retrieve. The remaining missing entry is Micro-Act on GPT-4o over ConflictBank, since the original paper does not report that value.

Llama-3.1-70B-Instruct over ConflictBank, where ASTUTERAG w/ retrieve obtains 45.1 and TRIAD obtains 46.0.

The gains are especially clear on SQuAD and MuSiQue. Averaged over six backbones, TRIAD reaches 14.1 on SQuAD and 18.7 on MuSiQue. This shows that TRIAD not only improves answer accuracy, but also more directly reduces the tendency to follow the distractor supported by misleading context.

ConflictBank remains the hardest benchmark in terms of Misled Rate. Even so, TRIAD averages 37.5 across six backbones, compared with 48.6 for Micro-Act, and 67.5 for ASTUTERAG. This is consistent with the main results and suggests that evidence calibration is especially helpful when the misleading context is longer and the conflicting signal is more deeply hidden.

D Open-ended QA Evaluation

In the main experiments, we use multiple-choice accuracy because it follows the native format of the benchmarks and directly measures whether a method can reject the distractor, while avoiding extra ambiguity from free-form answer matching. To

examine whether the conclusion transfers beyond option selection, we additionally conduct an open-ended QA evaluation on Qwen3.5-9B for KRE-SQuAD and KRE-MuSiQue. In this setting, models generate free-form answers, and correctness is assessed by an independent LLM judge.

Method	KRE-SQuAD	KRE-MuSiQue
End-to-End QA	19.5	16.2
GKP	36.0	13.8
Self-consistency	22.6	18.3
COMBO	35.1	14.0
ASTUTERAG	33.5	15.1
Micro-Act	38.2	17.2
TRIAD	45.7	24.1

Table 10: Open-ended QA accuracy on Qwen3.5-9B.

As shown in Table 10, open-ended RAG is clearly harder than multiple-choice QA, but TRIAD remains the best-performing method on both datasets. It outperforms the strongest baseline by 7.5 points on KRE-SQuAD and 5.8 points on KRE-MuSiQue. These results suggest that TRIAD’s advantage is not limited to option selection.

E Additional Generalization Experiments

We further examine generalization under two supplementary forms of external conflict. The first setting varies the rhetorical tone of misleading context while preserving the same distractor target. The second setting introduces inconsistent retrieved context in which both correct and misleading evidence appear in the same input.

E.1 Experimental Settings

Tone Based Misleading Contexts. To evaluate generalization across rhetorical tone, we use an adapted TruthfulQA-based split (Lin et al., 2022) derived from the TruthfulQA portion of Farm (Xu et al., 2024a), which studies model susceptibility to persuasive misinformation. Unlike entity oriented QA benchmarks, its answer options are often longer natural language statements. The dataset contains 484 base questions. Each question is paired with three misleading contexts that preserve the same distractor target but vary in persuasion style, namely *logical*, *credibility*, and *emotional*. This yields 1,452 evaluation instances in total. In the results below, the *Tone* score reports the average performance over these three variants on Qwen3.5-9B.

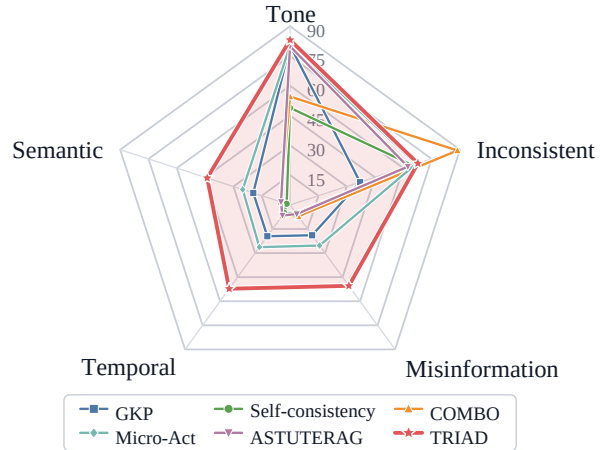


Figure 10: Radar comparison of generalization experiments on Qwen3.5-9B. Values correspond to ACC (%).

Method	Tone	Incons.	Misinf.	Temp.	Sem.
GKP	80.3	37.1	18.9	19.5	19.5
Self-cons.	48.9	64.8	5.2	4.9	1.8
COMBO	54.6	88.7	7.1	6.7	5.0
Micro-Act	80.4	66.7	25.3	26.3	25.1
ASTUTERAG	79.9	62.4	5.7	6.6	4.8
TRIAD	83.1	67.7	50.4	52.2	43.9

Table 11: Exact ACC (%) results for the supplementary generalization evaluation on Qwen3.5-9B.

Inconsistent Contexts. To evaluate generalization under inconsistent context, we construct a ConflictBank based split in which golden and misleading evidence are concatenated into the same retrieved context. This produces 500 examples per conflict type and 1500 examples in total. This setting is more demanding because the model must choose between context and memory while filtering conflicting evidence within context itself. When reporting the Inconsistent condition, we summarize results on this ConflictBank based split only.

E.2 Results and Analysis

Table 11 and Figure 10 show that TRIAD remains the strongest method in almost all supplementary settings. It achieves the best *Tone* result at 83.1 ACC, slightly above Micro-Act (80.4), GKP (80.3), and ASTUTERAG (79.9), and it also remains strongest on the three standard ConflictBank conflict types at 50.4/52.2/43.9. The only reversal appears on the *Inconsistent* split, where COMBO reaches 88.7, ahead of TRIAD’s 67.7.

Why TRIAD Remains Strong under Tone Variation. The TruthfulQA Farm setting changes how misleading evidence is framed rather than chang-

ing the underlying question semantics. In other words, the attack surface is mainly rhetorical: the same wrong answer can be presented in a more logical, more authoritative, or more emotionally persuasive way. By selecting question-aligned evidence rather than relying on the overall fluency or persuasiveness of the passage, TRIAD maintains 83.1 ACC, whereas methods that rely more heavily on global generation or unguided aggregation drop much more sharply, such as Self-consistency at 48.9 and COMBO at 54.6. This supports our claim that effective conflict resolution should generalize through calibrated evidence selection rather than through surface plausibility alone.

Why COMBO Excels on Inconsistent Contexts.

The *Inconsistent* setting changes the structure of the problem. The retrieved input now contains both supportive context and misleading context. By tracing COMBO’s outputs, we find that its advantage mainly comes from how it matches generated knowledge with local external passages. COMBO first generates several knowledge passages from the question and pairs them with segmented external passages. It then assigns higher weight to pairs in which the external passage is both evidential for the question and consistent with the generated passage. When one of the external passages explicitly contains the correct answer, this procedure naturally favors the pair whose generated knowledge agrees with that passage, so the supportive evidence is more likely to be preserved for final answer selection. The result pattern is consistent with this explanation, COMBO reaches 88.7 on *Inconsistent*, but only 7.1, 6.7, and 5.0 on the standard misinformation, temporal, and semantic splits, where no equally clean supportive passage is available inside the retrieved context. By contrast, TRIAD addresses the problem more comprehensively by jointly handling question understanding, evidence calibration, and parametric recall, which makes it stronger across the broader range of conflict settings considered in this paper.

Breakdown over the Three ConflictBank Types.

Among the three ConflictBank error types, semantic conflict remains the hardest. Unlike misinformation or temporal conflict, where the contradiction is often carried by a relatively direct wrong fact or date, semantic conflict tends to preserve topical relatedness and evidence-like form, so the misleading signal is buried more deeply in semantically plausible but incorrect relations. This makes the conflict

harder to expose with shallow lexical cues alone. Even under this more difficult condition, TRIAD still reaches 43.9 ACC, compared with 25.1 for Micro-Act, and 19.5 for GKP.

F Additional Robustness Experiments

F.1 False Sufficiency

We further analyze the robustness of TRIAD to false sufficiency in LLM-based intermediate judgments. We define false sufficiency as a case in which the model incorrectly determines that the selected evidence is sufficient when important supporting information is still missing. We use GPT-4o for initial screening and then manually verify cases from the Llama-3.1-70B-Instruct ConflictBank run. This process identifies 186 clear cases where evidence calibration stops early although important supporting information remains outside the selected chunks. Table 12 reports their accuracy and error behavior. Error categories follow Appendix I, and the error columns count only incorrect predictions.

Cases	Corr.	Acc.	Ctx. Dom.	Mem. Drift	Unres.
186	63	33.9%	47	66	10

Table 12: Accuracy and error behavior of false-sufficiency cases. Ctx. Dom. denotes Context Dominance, and Unres. denotes Unresolved Output. Error columns count only incorrect predictions.

False sufficiency is a real failure mode, but it does not cause a complete pipeline collapse. Even when the selected evidence misses important background information, TRIAD still recovers the correct answer in 33.9% of these cases. This suggests that clue-guided parametric knowledge elicitation can sometimes compensate for incomplete external evidence by using parametric knowledge.

For the error cases, the behavior aligns with the error modes in Appendix I. Some errors are Context Dominance. Although the selected chunks may not contain the exact distractor words, they can still contain semantically equivalent or locally misleading relations, which pull clue extraction or final resolution toward the misleading context. More errors are Memory Drift. These cases do not simply copy the retrieved distractor. Instead, the selected evidence is too weak, biased, or missing key constraints, causing clue extraction to elicit off-target parametric knowledge and leading the model to another wrong option.

F.2 Fallback to Full Context

We also analyze how often TRIAD falls back to the full retrieved context. Section 6 reports context compression statistics for Llama-3.1-70B-Instruct across the three datasets. TRIAD falls back to the full context in 15.2% of examples.

On ConflictBank, we additionally inspect a random sample of 100 such fallback cases. Their accuracy is nearly indistinguishable from the original distribution, 45.0% versus 44.9%, and does not show the large degradation observed in the *w/o Evidence Calibration* ablation. This suggests that fallback examples often have relatively short contexts in which almost every chunk contains useful background information. Therefore, fallback should not be interpreted simply as a failure of context compression. When no partial chunk set is judged sufficient, TRIAD preserves the full context to avoid discarding potentially necessary evidence.

G Cost Analysis Details

In this section, we analyze inference cost through average token consumption, latency, and generation cost per example on KRE-SQuAD, KRE-MuSiQue, and ConflictBank under GPT-4o and GPT-4o-mini, with all methods run in single-thread mode. For TRIAD, the reported time and cost include the embedding model used in evidence calibration. Tables 13 and 14 show that TRIAD is not the cheapest method. However, it achieves the highest accuracy and maintains a favorable overall cost-effectiveness profile. Compared with lightweight baselines such as GKP, TRIAD incurs higher token and cost overheads but yields large accuracy gains. Compared with heavier pipelines such as Micro-Act, TRIAD remains more efficient and still achieves higher average accuracy.

G.1 Latency

Latency largely follows pipeline depth. GKP is the fastest method, requiring 12.5 seconds per example on GPT-4o and 6.2 seconds on GPT-4o-mini. TRIAD is slower because it uses a multistage pipeline and also includes embedding computation. Its latency reaches 49.8 seconds on GPT-4o and 30.7 seconds on GPT-4o-mini. However, this increase remains moderate relative to other SAMs, including COMBO at 42.5 and 26.4 seconds, AS-TUTERAG at 46.1 and 33.2 seconds, and Micro-Act at 40.7 and 26.0 seconds.

G.2 Token Consumption and Monetary Cost

Token use and monetary cost show a similar pattern. GKP is the lightest method, using fewer than 800 input tokens per example. TRIAD uses about 7k input tokens per example and therefore incurs higher cost than lightweight baselines. However, its cost remains around the average level of SAMs. For example, TRIAD reduces average input tokens from 9984 to 7140 and output tokens from 1169 to 790 relative to Micro-Act on GPT-4o, while also lowering per example cost from 0.0311 to 0.0258 USD. These results show that TRIAD does not minimize cost, but its token and monetary overhead remain controlled relative to heavier alternatives.

G.3 Accuracy and Efficiency Tradeoff

The cost table shows two clear tradeoffs. Relative to GKP, TRIAD gains 26.4 and 15.3 accuracy points on GPT-4o and GPT-4o-mini, respectively, at a cost increase of about 5 times. Relative to SAMs, TRIAD attains the strongest overall performance while maintaining a cost level that remains within the same general range.

Taken together, these results suggest that TRIAD should not be viewed as a minimal cost method, but as a comparatively efficient high-accuracy method. Its main advantage is that the additional cost is spent on targeted conflict resolution steps that yield consistent accuracy benefits, rather than on a larger but less controlled reasoning budget.

H Case Study

Table 15 presents a representative two-hop example. The misleading context repeatedly supports the wrong answer *RCA Records*, while the gold answer is *Hollywood Records*. TRIAD first decomposes the question into two local decisions, then isolates evidence for each hop to identify the singer, and finally resolves the decisive conflict in the second hop by comparing misleading external evidence against parametric knowledge guided by clues.

On this same example, GKP predicts the distractor *RCA Records*. Its generated parametric knowledge drifts to an unrelated singer, and without hop-wise decomposition or evidence calibration, the misleading context remains dominant in the final comparison. This contrast helps explain why TRIAD is more robust under coupled failures of question understanding, external evidence reliability, and parametric recall.

Method	Avg. ACC	Avg. Input Tokens	Avg. Output Tokens	Avg. Time (s)	Avg. Cost (USD)
GKP	43.7	792	320	12.5	0.0052
Self-consistency	25.3	2684	880	40.8	0.0155
COMBO	32.4	3314	738	42.5	0.0157
ASTUTERAG	43.9	2361	1072	46.1	0.0166
Micro-Act	46.3	9984	1169	40.7	0.0311
TRIAD	70.1	7140	790	49.8	0.0258

Table 13: Cost comparison under GPT-4o.

Method	Avg. ACC	Avg. Input Tokens	Avg. Output Tokens	Avg. Time (s)	Avg. Cost (USD)
GKP	40.6	772	259	6.2	0.0003
Self-consistency	19.9	2684	546	14.6	0.0007
COMBO	28.5	3467	658	26.4	0.0009
ASTUTERAG	38.6	2221	789	33.2	0.0008
Micro-Act	46.4	11027	1094	26.0	0.0020
TRIAD	55.8	6981	758	30.7	0.0015

Table 14: Cost comparison under GPT-4o-mini.

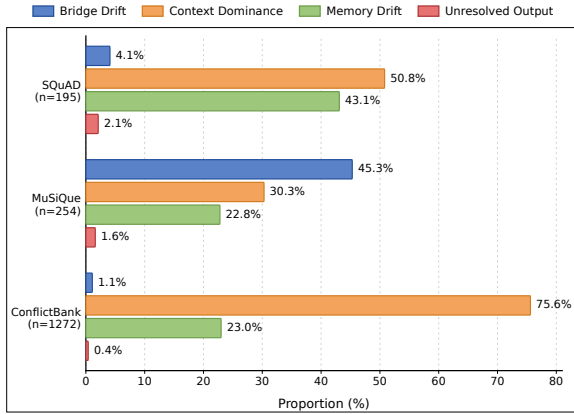


Figure 11: Distribution of TRIAD error types on GPT-4o. Here, n denotes the number of incorrect examples in each dataset.

I Error Analysis

We conduct error analysis on TRIAD using the GPT-4o runs over KRE-SQuAD, KRE-MuSiQue, and ConflictBank. Figure 11 summarizes the distribution of error types over incorrect examples.

I.1 Error Taxonomy

We group errors into four categories. The first three directly mirror the main failure axes discussed in the paper.

Bridge Drift. Bridge Drift captures multi-hop failures in which an earlier intermediate answer does not align with the evidence selected for the next hop. In practice, this usually comes from imperfect question decomposition, overly broad intermediate answers, or entity binding drift. Once the bridge entity is wrong or too coarse, later subques-

tions become misgrounded even if the downstream modules behave as intended.

Context Dominance. Context Dominance refers to cases where the final wrong prediction exactly matches the distractor supported by the misleading context. This indicates that external evidence remains dominant at answer time. Our traces reveal two recurring causes. First, as noted by Micro-Act (Huo et al., 2025), misleading evidence often appears more concrete and specific than parametric knowledge by naming a particular entity, location, date, or relation. Second, once a wrong relation survives Step 2, it can propagate into clue extraction and final resolution, blunting the perceived context-memory conflict and producing a spurious agreement around the misleading answer.

Memory Drift. Memory Drift covers cases where the model is not simply following the retrieved distractor. Instead, it recalls the wrong knowledge point from parametric memory, such as an incorrect entity or an answer that does not match the question. This usually happens when the retrieved evidence is too weak or noisy to anchor the model, so clue-guided elicitation still ends up on low-relevance parametric knowledge.

Unresolved Output. Unresolved Output includes residual failures where the pipeline does not map to one of the provided options, or yields an abstaining style response such as a null prediction. These cases are rare.

I.2 Dataset Specific Patterns

The three datasets exhibit clearly different failure profiles, which is consistent with their benchmark design.

KRE-SQuAD. On KRE-SQuAD, Context Dominance is the largest category at 50.8%, followed closely by Memory Drift at 43.1%, while Bridge Drift is only 4.1%. This pattern is expected because KRE-SQuAD is mostly single-hop. In other words, SQuAD failures mainly come from competition between external evidence and internal memory, not from structural misunderstanding of the question.

KRE-MuSiQue. KRE-MuSiQue shows the opposite profile. Bridge Drift is the dominant error at 45.3%, while Context Dominance and Memory Drift account for 30.3% and 22.8%, respectively. This confirms that multi-hop question understanding is the main bottleneck on MuSiQue. Many failures begin with an incorrect intermediate answer, which then distorts the next subquestion and prevents later evidence calibration from operating on the right semantic target.

ConflictBank. ConflictBank is overwhelmingly dominated by Context Dominance. 75.6% of errors fall into this category. This profile matches the construction of ConflictBank, where the misleading context supports a specific distractor in a direct and persuasive way. Because most instances are effectively single-hop, question understanding rarely fails. Instead, the main challenge is whether the model can resist a strongly supported external wrong answer.

Taken together, these results support TRIAD’s claim that robust conflict resolution requires jointly addressing question understanding, misleading external context, and off-target parametric recall, rather than focusing on any single source of error in isolation.

Query	Question: The singer, who starred in <i>Shake It Up</i> , signed to which record label? Choices: A. Atlantic Records; B. Capitol Records; C. Hollywood Records; D. RCA Records. Gold / misleading answers: Hollywood Records ✓ / RCA Records ✗.
Misleading Context	Chunk 1: RCA Records is the eponymous debut studio album by American singer and actress Zendaya, released on September 17, 2013 by RCA Records. Chunk 2: After acting in the Disney Channel series “Shake It Up,” Zendaya signed a recording contract with RCA Records, in which she began recording her debut in late 2012. Chunk 3: “Zendaya” consists of 12 songs; musically, the album is an electropop album that incorporates urban pop, R&B, and dubstep. Chunk 4: Lyrically, the album discusses issues of heartbreak and love. Chunk 5: Zendaya co-wrote all the songs on the album. Chunk 6: “Swag It Out” is the official debut single by American singer Zendaya, known for starring in the series “Shake It Up.” Chunk 7: The song was composed by Bobby Brackins and was released on May 30, 2011, as the first single of Zendaya’s career on an independent label, not being included on any album or soundtrack of Walt Disney Records, just their record label.
Question Decomposition	Who is the singer that starred in <i>Shake It Up</i> ? Which record label did [[v1]] sign to?
Subquestion 1: Who is the singer that starred in <i>Shake It Up</i> ?	
Evidence Calibration	Chunk 6: “Swag It Out” is the official debut single by American singer Zendaya, known for starring in the series <i>Shake It Up</i> .
Clue-Guided Parametric Knowledge Elicitation	Clues: American singer; released a debut single; starred in <i>Shake It Up</i> . Parametric knowledge: The singer who starred in <i>Shake It Up</i> is Zendaya. She is an American actress and singer who gained prominence for her role in the Disney Channel series. ...
Answer Resolution	Reasoning: The external context explicitly states that Zendaya is an American singer known for starring in the series “Shake It Up.” This aligns with internal knowledge, which also identifies Zendaya as the singer who starred in the show. Both sources are consistent and provide clear evidence supporting the answer. Predicted answer: Zendaya. ✓
Subquestion 2: Which record label did Zendaya sign to?	
Evidence Calibration	Chunk 1: RCA Records is the eponymous debut studio album by American singer and actress Zendaya, released on September 17, 2013 by RCA Records. Chunk 2: After acting in the Disney Channel series “Shake It Up,” Zendaya signed a recording contract with RCA Records, in which she began recording her debut in late 2012.
Clue-Guided Parametric Knowledge Elicitation	Clues: acted in <i>Shake It Up</i> before signing; the contract led to her debut album; the album was released in 2013. Parametric knowledge: Zendaya signed a recording contract with Hollywood Records, a label known for working with many Disney-affiliated artists. This aligns with her career trajectory, as she gained prominence through her role in the Disney Channel series <i>Shake It Up</i>
Answer Resolution	Reasoning: The internal knowledge indicates that Zendaya signed with Hollywood Records, which is consistent with her Disney Channel career trajectory. The external context claims she signed with RCA Records, but this conflicts with established knowledge about her early music career. Hollywood Records is the more reliable answer based on overall consistency and alignment with her career history. Predicted answer: Hollywood Records. ✓

Table 15: Case study of TRIAD on a two-hop example.

Step 1. Question Decomposition

Task:

- 1) Decide if the original question is single-hop or multi-hop.
- 2) If multi-hop, decompose it into an ordered sequence of minimal single-hop subquestions.

Rules:

- Only classify it as multi-hop if you must first resolve an intermediate unknown entity, event, relation, or fact before answering.
- Provide must_bind value for each hop (v1, v2, ...).
- Multi-hop must contain a real variable chain: later hop explicitly uses [[v1]][[v2]] placeholders.

Original question: {question}

Return JSON only with schema:

```
{{"task_type": "single-hop" | "multi-hop",
  "subquestions": [
    {{
      "hop_id": "H1",
      "question": "...",
      "must_bind": "v1",
      "depends_on": []
    }}
  ]
}}
```

Figure 12: Prompt for question decomposition.

Step 2a. Support Span Extraction

You are analyzing one context chunk for a single-hop question.

Task:

Extract the minimal exact support span(s) from the chunk that are actually useful for answering the question.

Rules:

- Copy support spans directly from the chunk.
- Do not paraphrase.
- Keep only the minimum span or spans needed for answering the question.
- If no part of the chunk is useful, return an empty list.
- Return JSON only.

Question: {subquestion}

Chunk: {chunk}

Return JSON only: {"support_spans": ["...", "..."]}

Step 2b. Sufficiency Judgment

You are judging whether the currently selected context is sufficient to answer a single-hop question.

Rules:

- Judge only whether the selected context is sufficient to answer the question.
- Answer true only if the selected context already contains enough information.
- Return JSON only.

Question: {subquestion}

Selected context: {selected_context}

Return JSON only: {"is_sufficient": true}

Figure 13: Prompt for evidence calibration.

Step 3a. Indirect Clue Extraction

Task: Given the current single-hop subquestion, hop-specific context, and answer choices:

- 1) output a reasoning chain for the subquestion,
- 2) identify direct answer candidates mentioned in the context,
- 3) then output ONLY indirect clues that help reasoning but do not leak direct answers.

Rules:

- Do NOT include direct answer candidates copied from context.
- Keep clues concise.
- The reasoning chain should explicitly conclude with the subquestion answer.
- Clues must be short phrases, relevant to people/time/location/reasons/events/constraints/background.
- If context seems contaminated and safe clues cannot be extracted, return an empty clue list.
- Return JSON only.

Current subquestion: {subquestion}

Hop-specific context: {hop_context}

Choices: {choices}

Return JSON only: `{{"reasoning_chain": "...", "answer_candidates": ["..."], "clues": ["...", "..."]}}`

Step 3b. Parametric Knowledge Elicitation

Task: Use the current single-hop subquestion, extracted clues, and answer choices to generate model internal knowledge relevant to this question.

Important rules:

- The internal knowledge MUST contain a direct answer to the subquestion.
- Clues may be noisy or poisoned by wrong external context. Treat clues as weak hints, not ground truth.
- Prefer stable world knowledge and globally consistent facts over contaminated clues.
- Do NOT just restate the clues.
- Output one coherent paragraph.
- Output valid JSON only.

Current subquestion: {subquestion}

Clues: {clues}

Choices: {choices}

Return JSON with this schema: `{{"internal_knowledge": "..."} }`

Figure 14: Prompt for indirect clue extraction and clue-guided parametric knowledge elicitation.

Step 4. Answer Resolution for Each Hop

Task:

You are a detective, you have received a question (Current subquestion), a suspect provides (Hop-specific external context), which may conflict with your knowledge (internal_knowledge), you will get options about the question, the correct answer and the wrong answer are among them, you MUST find out which side is more reliable based on your knowledge, you need to give the reasoning process (reasoning_summary), output your predicted answer (predicted_answer) and provide the corresponding option ({label_text}).

Notes:

- Do not rely on the Hop-specific external context mechanically.
- You need to use your reasoning ability, combined with your understanding of similar cases, to make your judgment.
- Focus on overall consistency rather than simple word matching.
- The output option (label) must be one of {label_text}. If the predicted_answer cannot match a specific option, the output label should be null.
- Return JSON only.

Current subquestion: {subquestion}

Hop-specific external context: {hop_context}

Internal knowledge: {internal_knowledge}

Choices: {choices}

Return JSON only: `{{"reasoning_summary": "...", "predicted_answer": "...", "label": {label_schema}}}`

Figure 15: Prompt template for answer resolution.