TruthfulRAG: Resolving Factual-level Conflicts in Retrieval-Augmented Generation with Knowledge Graphs

Shuyi Liu, Yu-Ming Shang, Xi Zhang*

Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
{liushuyi111, shangym, zhangx}@bupt.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a powerful framework for enhancing the capabilities of Large Language Models (LLMs) by integrating retrieval-based methods with generative models. As external knowledge repositories continue to expand and the parametric knowledge within models becomes outdated, a critical challenge for RAG systems is resolving conflicts between retrieved external information and LLMs' internal knowledge, which can significantly compromise the accuracy and reliability of generated content. However, existing approaches to conflict resolution typically operate at the token or semantic level, often leading to fragmented and partial understanding of factual discrepancies between LLMs' knowledge and context, particularly in knowledge-intensive tasks. To address this limitation, we propose TruthfulRAG, the first framework that leverages Knowledge Graphs (KGs) to resolve factual-level knowledge conflicts in RAG systems. Specifically, TruthfulRAG constructs KGs by systematically extracting triples from retrieved content, utilizes query-based graph retrieval to identify relevant knowledge, and employs entropy-based filtering mechanisms to precisely locate conflicting elements and mitigate factual inconsistencies, thereby enabling LLMs to generate faithful and accurate responses. Extensive experiments reveal that TruthfulRAG outperforms existing methods, effectively alleviating knowledge conflicts and improving the robustness and trustworthiness of RAG systems.

Extended version — https://arxiv.org/abs/2511.10375

Introduction

Large Language Models (LLMs) have demonstrated impressive performance across diverse natural language understanding and generation tasks (Achiam et al. 2023; Touvron and et al. 2023; Yang et al. 2025). Despite their proficiency, LLMs remain ineffective in handling specialized, privacy-sensitive, or time-sensitive knowledge that is not encompassed within their training corpora (Zhang et al. 2024; Huang et al. 2025). For the solutions, Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm that enhances the relevance and factuality of the generated responses by integrating external knowledge retrieval with the remarkable generative capabilities of LLMs (Lewis et al. 2020; Gao et al. 2023; Fan et al. 2024). However, as RAG systems continuously update their knowledge repositories, the temporal disparity between dynamic external sources and static parametric knowledge within LLMs inevitably leads to knowledge conflicts (Xie et al. 2023; Xu et al. 2024; Shi et al. 2024), which can significantly undermine the accuracy and reliability of the generated content.

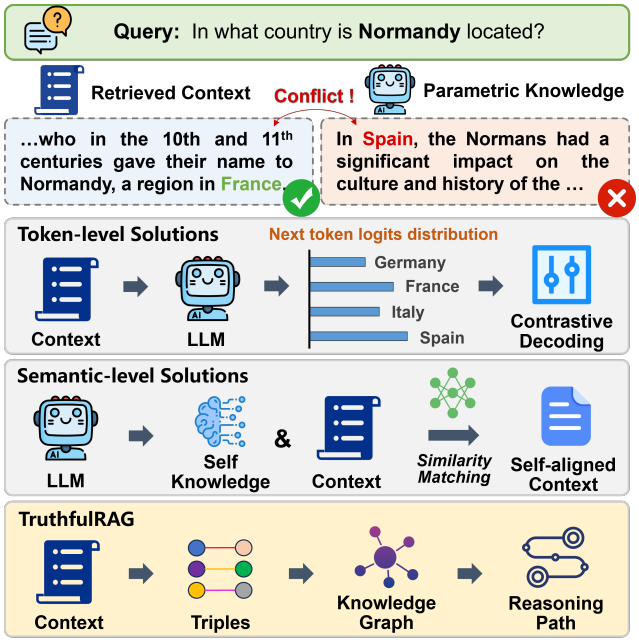

Figure 1: The illustration of knowledge conflicts and the differences between existing solutions and TruthfulRAG.

Recent research has begun to investigate the impact of knowledge conflicts on the performance of RAG systems (Chen, Zhang, and Choi 2022; Xie et al. 2023; Tan et al. 2024) and explore methods to mitigate such conflicts (Wang et al. 2024; Jin et al. 2024; Zhang et al. 2025; Bi et al. 2025). Existing resolution approaches can be categorized into two methodological types: (i) token-level methods, which manage LLMs' preference between internal and external knowledge by adjusting the probability distribution

*Corresponding author.

over the output tokens (Jin et al. 2024; Bi et al. 2025); (ii) semantic-level methods, which resolve conflicts by semantically integrating and aligning knowledge segments from internal and external sources (Wang et al. 2024; Zhang et al. 2025). However, these token-level or semantic-level conflict resolution methods generally employ coarse-grained strategies that rely on fragmented data representations, resulting in insufficient contextual awareness. This may prevent LLMs from accurately capturing complex interdependencies and fine-grained factual inconsistencies, especially in knowledge-intensive conflict scenarios (Han et al. 2024).

To address the above limitations, we propose TruthfulRAG, the first framework that leverages Knowledge Graphs (KGs) to resolve factual-level conflicts in RAG systems. As illustrated in Figure 1, unlike previous studies, TruthfulRAG uses structured triple-based knowledge representations to construct reliable contexts, thereby enhancing the confidence of LLMs in external knowledge and facilitating trustworthy reasoning. The TruthfulRAG framework comprises three key modules: (a) Graph Construction, which derives structured triples from retrieved external knowledge by identifying entities, relations, and attributes to construct knowledge graphs; (b) Graph Retrieval, which conducts query-based retrieval algorithms to obtain relevant knowledge that exhibit strong factual associations with the input query; and (c) Conflict Resolution, which applies entropy-based filtering techniques to locate conflicting elements and mitigate factual inconsistencies, ultimately forming more reliable reasoning paths and promoting more accurate outputs. This framework integrates seamlessly with existing RAG architectures, enabling the extraction of highly relevant and factually consistent knowledge, effectively eliminating factual-level conflicts and improving generation reliability.

The contributions of this paper are as follows:

- We discover that constructing contexts through textual representations on structured triples can enhance the confidence of LLMs in external knowledge, thereby promoting trustworthy and reliable model reasoning.
- We introduce TruthfulRAG, the first framework that leverages knowledge graphs to resolve factual-level conflicts in RAG systems through systematic triple extraction, query-based graph retrieval, and entropy-based filtering mechanisms.
- We conduct extensive experiments demonstrating that TruthfulRAG outperforms existing methods in mitigating knowledge conflicts while improving the robustness and trustworthiness of RAG systems.

Methodology

This section introduces the overall framework of TruthfulRAG. As shown in Figure 2, TruthfulRAG comprises three interconnected modules: (i) Graph Construction, which transforms unstructured retrieved content into structured knowledge graphs through systematic triple extraction; (ii) Graph Retrieval, which employs query-aware graph traversal algorithms to identify semantically relevant reasoning paths; and (iii) Conflict Resolution, which utilizes entropy-based filtering mechanisms to detect and mitigate factual inconsistencies between parametric and external knowledge.

Graph Construction

The construction of a knowledge graph begins with the conversion of raw information retrieved from the RAG system into structured knowledge representations through systematic entity-relation-attribute extraction.

Given the retrieved content C for the user's query q, we first perform fine-grained semantic segmentation to partition the content into coherent textual segments $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$, where each segment s_i represents a semantically coherent unit containing factual information. For each textual segment $s_i \in \mathcal{S}$, we employ the generative model $\mathcal{M}$ from the RAG system to extract a set of structured knowledge triples $\mathcal{T}_{all} = \{\mathcal{T}_{i,1}, \mathcal{T}_{i,2}, \ldots, \mathcal{T}_{i,n}\}$, with each triple $\mathcal{T}_{i,j} = (h, r, t)$ consisting of a head entity h, relation r, tail entity t. This extraction process aims to capture both explicit factual statements and implicit semantic relationships embedded within the original content, thereby ensuring the comprehensiveness and semantic integrity of the knowledge representation.

The aggregated triple set from all retrieved content forms the foundation for constructing the knowledge graph $\mathcal{G}$:

$$\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}_{all}) \tag{1}$$

where $\mathcal{E} = \bigcup_{i,j,k} h_{i,j,k}, t_{i,j,k}$ represents the entity set, $\mathcal{R} = \bigcup_{i,j,k} r_{i,j,k}$ denotes the relation set, and $\mathcal{T}_{all} = \bigcup_{i,j} \mathcal{T}_{i,j}$ constitutes the complete triple repository. This structured knowledge representation enables the filtering of low-information noise and captures detailed factual associations, thereby providing a clear and semantically enriched foundation for subsequent query-aware knowledge retrieval.

Graph Retrieval

To acquire knowledge that is strongly aligned with user queries at the factual level, we design a query-aware graph traversal algorithm that can identify critical knowledge paths within the graph, ensuring both semantic relevance and factual consistency in the retrieval process.

Initially, key elements are extracted from the user query q to serve as important references for matching components in the knowledge graph. These elements include the query's target entities, relations, and intent categories, denoted as $\mathcal{K}_q$. Subsequently, semantic similarity matching is employed to identify the top-k most relevant entities and relations within the knowledge graph:

$$\mathcal{E}imp = \text{TopK}(\text{sim}(e, \mathcal{K}_q) : e \in \mathcal{E}, k) \tag{2}$$

$$\mathcal{R}imp = \text{TopK}(\text{sim}(r, \mathcal{K}_q) : r \in \mathcal{R}, k) \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ represents the semantic similarity function computed using dense embeddings, $\mathcal{E}imp$ denotes the set of key entities, and $\mathcal{R}imp$ represents the set of key relations. From each key entity $e \in \mathcal{E}imp$, we perform a two-hop graph traversal to systematically collect the entire set of possible initial reasoning paths $\mathcal{P}init$.

To further filter reasoning paths with stronger factual associations, we introduce a fact-aware scoring mechanism

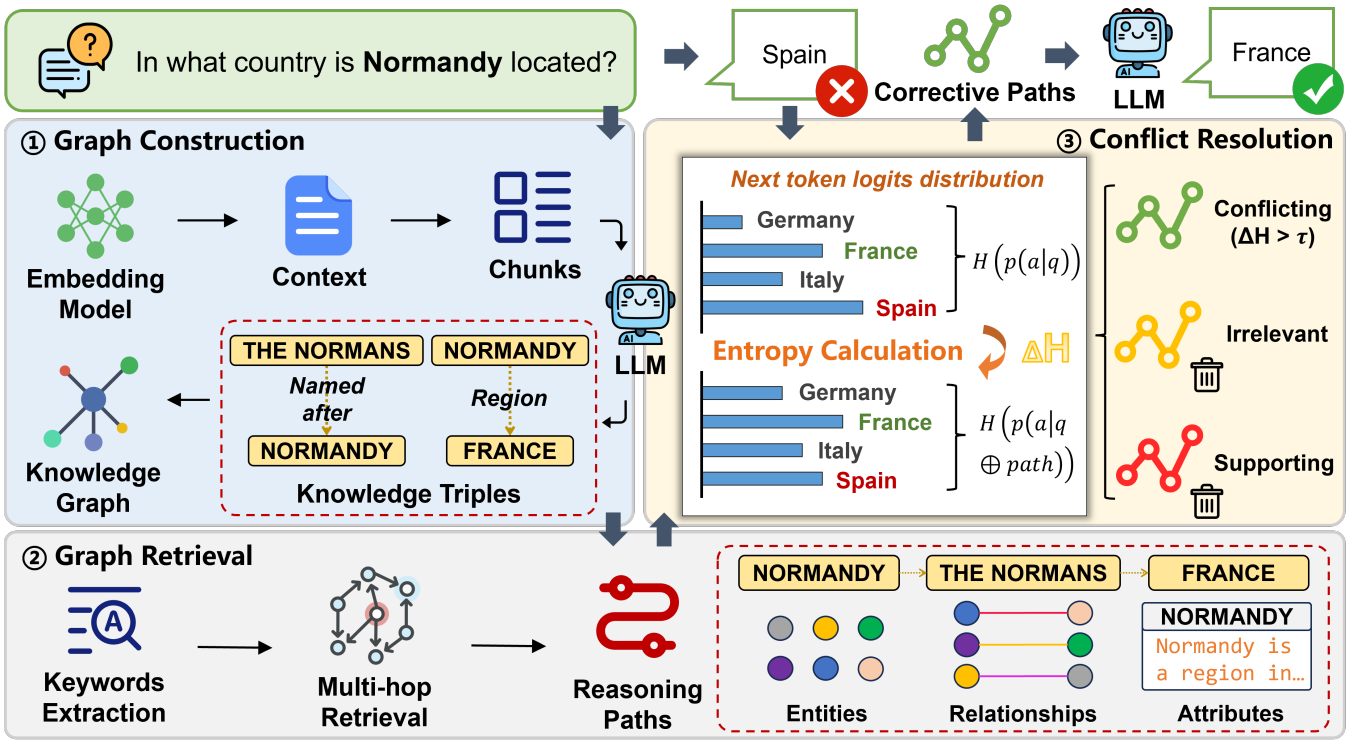

Figure 2: The overall pipeline of the TruthfulRAG framework. TruthfulRAG first extracts structured knowledge triples to construct a comprehensive knowledge graph. Subsequently, it employs query-aware graph traversal to identify salient reasoning paths, where each path comprises entities and relationships enriched with associated attributes. Finally, the framework applies entropy-based conflict resolution to detect and filter out corrective paths that challenge parametric misconceptions, thereby alleviating knowledge conflicts between internal and external information, prompting consistent and credible responses.

that evaluates the relevance of paths to the query based on the coverage of key entities and relations within each path p:

$$\text{Ref}(p) = \alpha \cdot \frac{|e \in p \cap \mathcal{E}imp|}{|\mathcal{E}imp|} + \beta \cdot \frac{|r \in p \cap \mathcal{R}imp|}{|\mathcal{R}imp|} \quad (4)$$

where α and β are hyperparameters that control the relative importance of entity and relationship coverage, respectively. The top-scored reasoning paths from $\mathcal{P}init$ constitute the core knowledge paths $\mathcal{P}super$.

$$\mathcal{P}super = \text{TopK}(\text{Ref}(p) : p \in \mathcal{P}init, K) \quad (5)$$

In order to construct detailed contextual information, each core reasoning path $p \in \mathcal{P}super$ will be represented as a comprehensive contextual structure consisting of three essential components:

$$p = \mathcal{C}_{path} \oplus \mathcal{C}_{entities} \oplus \mathcal{C}_{relations} \quad (6)$$

where:

- $\mathcal{C}path$ represents the complete sequential reasoning path: $e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} \cdots \xrightarrow{r_{n-1}} e_n$, capturing the logical progression of entities connected through relational links.
- $\mathcal{C}entities = (e, \mathcal{A}e) : e \in p \cap \mathcal{E}imp$ encompasses all important entities within the path along with their corresponding attribute descriptions $\mathcal{A}e$, providing thorough entity-specific information for the context.
- $\mathcal{C}relations = (r, \mathcal{A}r) : r \in p \cap \mathcal{R}imp$ includes all important relations on the path together with their corresponding attributes $\mathcal{A}r$, enriching the semantic and contextual understanding of the relations.

This formalized representation of knowledge ensures that each extracted reasoning path preserves structural coherence through the entity-relation sequence and reinforces semantic richness via comprehensive attribute information, thereby facilitating more nuanced and context-aware knowledge integration for subsequent conflict resolution processes.

Conflict Resolution

To address factual inconsistencies between parametric knowledge and external information, ensuring that LLMs consistently follow the retrieved knowledge paths to achieve accurate reasoning, we employ entropy-based model confidence analysis to investigate the influence of conflicting knowledge on model prediction uncertainty, thereby systematically identifying and resolving factual conflicts based on uncertainty quantification mechanisms.

We implement conflict detection by comparing model performance under two distinct conditions: (1) pure parametric generation without access to external context, and (2) retrieval-augmented generation that incorporates structured reasoning paths constructed from knowledge graph.

For parametric-based generation, we calculate the response probability from LLMs as baselines:

$$P_{param}(ans|q) = \mathcal{M}(q) \tag{7}$$

where ans represents the generated answer and $\mathcal{M}(q)$ denotes the response distribution of the LLM based solely on query q. For retrieval-augmented generation, we incorporate each reasoning path from $\mathcal{P}super$ as contextual information to obtain the model's output probability:

$$P_{aug}(ans|q, p) = \mathcal{M}(q \oplus p), \quad \forall p \in \mathcal{P}super \tag{8}$$

where $\mathcal{M}(q \oplus p)$ represents the response distribution of the LLM conditioned on the query q and its corresponding reasoning paths extracted from the knowledge graph.

Inspired by previous research on probability-based uncertainty estimation (Arora, Huang, and He 2021; Duan et al. 2024), we adopt entropy-based metrics to quantify the model's confidence in the retrieved knowledge:

$$H(P(ans|context)) = -\frac{1}{|l|}\sum_{t=1}^{|l|}\sum_{i=1}^{k} pr_i^{(t)} \log_2 pr_i^{(t)} \tag{9}$$

where $pr_i^{(t)}$ represents the probability distribution over the top-k candidate tokens at position t, and $|l|$ denotes the token length of the answer. Accordingly, we obtain $H(P_{param}(ans|q))$ for parametric generation and $H(P_{aug}(ans|q, p))$ for retrieval-augmented generation incorporating with individual reasoning path p. Consequently, we can utilize the entropy variation under different reasoning paths as a characteristic indicator of knowledge conflict:

$$\Delta H_p = H(P_{aug}(ans|q, p)) - H(P_{param}(ans|q)) \tag{10}$$

where positive values of ΔH_p indicate that the retrieved external knowledge intensifies uncertainty in the LLM's reasoning, potentially indicating factual inconsistencies with its parametric knowledge, whereas negative values suggest that the retrieved knowledge aligns with the LLM's internal understanding, thereby reducing uncertainty. Reasoning paths exhibiting entropy changes exceeding a predefined threshold τ are classified as $\mathcal{P}_{corrective}$:

$$\mathcal{P}corrective = p \in \mathcal{P}super : \Delta H_p > \tau \tag{11}$$

These identified corrective knowledge paths, which effectively challenge and potentially rectify the LLM's internal misconceptions, are subsequently aggregated to construct the refined contextual input. The final response is then generated by the LLM based on the enriched context:

$$\text{Response} = \mathcal{M}(q \oplus \mathcal{P}corrective) \tag{12}$$

This entropy-based conflict resolution mechanism ensures that LLMs consistently prioritize factually accurate external information when generating responses, improving reasoning accuracy and trustworthiness, thereby enhancing the overall robustness of the RAG system.

Experiments

In this section, we present comprehensive experiments to evaluate the effectiveness of TruthfulRAG in resolving knowledge conflicts and enhancing the reliability of RAG systems. Specifically, we aim to address the following research questions: (1) How does TruthfulRAG perform compared to other methods in terms of factual accuracy? (2) What is the performance of TruthfulRAG in non-conflicting contexts? (3) To what extent do structured reasoning paths affect the confidence of LLMs compared to raw natural language context? (4) What are the individual contributions of each module within the TruthfulRAG framework?

Experimental Setup

Datasets We conduct experiments on four datasets that encompass various knowledge-intensive tasks and conflict scenarios. FaithEval (Ming et al. 2025) is designed to assess whether LLMs remain faithful to unanswerable, inconsistent, or counterfactual contexts involving complex logical-level conflicts beyond the entity level. MuSiQue (Trivedi et al. 2022) and SQuAD (Rajpurkar et al. 2016) come from previous research KRE (Ying et al. 2024), which contain fact-level knowledge conflicts that necessitate compositional multi-hop reasoning, making it particularly suitable for evaluating knowledge integration and conflict resolution in complex reasoning scenarios. RealtimeQA (Kasai et al. 2023) focuses on temporal conflicts, where answers may quickly become outdated, leading to inconsistencies between static parametric knowledge and dynamic external sources.

Evaluated Models We select three representative LLMs across different architectures and model scales to ensure comprehensive evaluations: GPT-4o-mini (Achiam et al. 2023), Qwen2.5-7B-Instruct (Yang et al. 2025), and Mistral-7B-Instruct (Jiang et al. 2024). This selection encompasses both open-source and closed-source models, ensuring that TruthfulRAG is broadly applicable to RAG systems built upon diverse LLM backbones.

Baselines We compare TruthfulRAG against five baseline approaches spanning different methodological categories: (i) Direct Generation requires LLMs to generate responses solely based on their parametric knowledge without any external retrieval. (ii) Standard RAG represents the conventional retrieval-augmented generation paradigm, where LLMs generate responses using retrieved textual passages directly. (iii) KRE (Ying et al. 2024) serves as a representative prompt optimization method, which enhances reasoning faithfulness by adopting specialized prompting strategies to guide the model in resolving knowledge conflicts. (iv) COIECD (Yuan et al. 2024) represents the decoding manipulation category, which modifies the model's decoding strategy during the inference stage to guide LLMs toward greater reliance on retrieved context rather than parametric knowledge. (v) FaithfulRAG (Zhang et al. 2025) incorporates a self-reflection mechanism that identifies factual discrepancies between parametric knowledge and retrieved context, enabling LLMs to reason and integrate conflicting facts before generating content.

Evaluation Metrics Following prior studies, we adopt accuracy (ACC) as the primary evaluation metric, measuring the proportion of questions for which the LLM generates

Method	LLM	Dataset				Avg.	Imp.
		FaithEval	MuSiQue	RealtimeQA	SQuAD		
w/o RAG	GPT-4o-mini	4.6	15.1	43.4	11.2	18.6	-
	Qwen2.5-7B-Instruct	4.2	19.6	40.7	11.1	18.9	-
	Mistral-7B-Instruct	6.3	13.8	29.2	11.5	15.2	-
w/ RAG	GPT-4o-mini	61.3	72.6	67.3	73.1	68.6	50.0
	Qwen2.5-7B-Instruct	53.1	75.2	78.7	68.3	68.8	49.9
	Mistral-7B-Instruct	61.9	67.6	52.2	67.2	62.2	47.0
KRE	GPT-4o-mini	50.7	34.6	47.5	65.3	49.5	30.9
	Qwen2.5-7B-Instruct	59.6	70.7	**86.7**	73.7	72.7	53.8
	Mistral-7B-Instruct	73.2	50.6	76.9	74.6	68.8	53.6
COIECD	GPT-4o-mini	53.9	56.4	48.7	57.6	54.2	35.6
	Qwen2.5-7B-Instruct	62.3	69.7	78.8	70.8	70.4	51.5
	Mistral-7B-Instruct	62.8	66.8	58.4	65.4	63.3	48.1
FaithfulRAG	GPT-4o-mini	67.2	79.3	78.8	80.8	76.5	58.0
	Qwen2.5-7B-Instruct	71.8	78.0	84.1	78.3	78.1	59.1
	Mistral-7B-Instruct	81.7	78.5	77.0	**85.7**	80.7	65.5
TruthfulRAG (Ours)	GPT-4o-mini	**69.5**	**79.4**	**85.0**	**81.1**	**78.8**	**60.2**
	Qwen2.5-7B-Instruct	**73.2**	**79.1**	82.3	**78.7**	**78.3**	**59.4**
	Mistral-7B-Instruct	**81.9**	**79.3**	**81.4**	82.7	**81.3**	**66.1**

Table 1: ACC comparison between TruthfulRAG and five baselines across four datasets within three LLMs. The best result for each backbone LLM within each dataset is in **bold** and the second best is underlined. **Avg.** denotes the average ACC across the four datasets, while **Imp.** indicates the average improvement over the corresponding LLM's w/o RAG baseline.

correct answers, thereby providing a direct assessment of the factual correctness of the generated responses. To evaluate the method's capability to precisely extract information pertinent to the target answer from retrieved corpora, we introduce the Context Precision Ratio (CPR) metric, which measures the proportion of answer-related content within the processed context:

$$\text{CPR} = \frac{|\mathcal{A}_{gold} \cap \mathcal{C}_{processed}|}{|\mathcal{C}_{processed}|} \tag{13}$$

where $|\text{Context}_{gold}|$ denotes the length of segments directly related to the correct answer, and $|\text{Context}_{processed}|$ represents the total length of the processed context.

Implementation Details For dense retrieval, cosine similarity is computed using embeddings generated by the all-MiniLM-L6-v2. For entropy-based filtering, we set model-specific thresholds τ for entropy variation ΔH_p: GPT-4o-mini and Mistral-7B-Instruct use $\tau = 1$, while Qwen2.5-7B-Instruct adopts a higher threshold of $\tau = 3$. All experiments are conducted using NVIDIA V100 GPUs with 32GB memory. To ensure reproducibility, the temperature for text generation is set to 0, and all Top-K values are set to 10.

Results and Analysis

Overall Performance Table 1 presents a comprehensive comparison of TruthfulRAG against five baseline methods across four datasets, evaluating performance in terms of factual accuracy (ACC) using three representative LLMs. To facilitate overall assessment, we additionally report **Avg.**, the arithmetic mean accuracy across the four datasets, and **Imp.**, the average improvement over the corresponding LLM's w/o RAG baseline, serving as a proxy for the number of factual conflicts successfully corrected by the method from the LLM's parametric knowledge.

The results clearly demonstrate that TruthfulRAG consistently achieves superior or competitive performance relative to all baseline approaches. Specifically, it achieves the highest accuracy on FaithEval (81.9%), MuSiQue (79.4%), and RealtimeQA (85.0%), and ranks first or second on SQuAD across all models. Notably, TruthfulRAG achieves the highest overall performance across all backbone LLMs, attaining both the best average accuracy (**Avg.**) and the greatest relative improvement (**Imp.**) compared to all baseline methods. This clearly illustrates its robustness in mitigating factual inconsistencies that standard RAG systems struggle with due to unresolved evidence conflicts.

Compared to standard RAG systems, which exhibit significant variability in accuracy due to unresolved knowledge conflicts, TruthfulRAG achieves improvements ranging from 3.6% to 29.2%, highlighting its robustness in mitigating factual inconsistencies. Furthermore, while methods like FaithfulRAG and KRE offer partial gains through semantic alignment or prompt-based mechanisms, they fall short in consistently resolving fine-grained factual discrepancies. In contrast, TruthfulRAG integrates knowledge graph-based reasoning with entropy-guided conflict filtering mechanisms to identify and resolve contradictory information, thereby substantially enhancing factual reliability. These findings validate the effectiveness of TruthfulRAG in delivering accurate, faithful, and contextually grounded responses across diverse knowledge-intensive tasks.

Dataset	Method					
	w/o RAG	w/ RAG	KRE	COIECD	FaithfulRAG	TruthfulRAG (Ours)
MuSiQue-golden	45.6	89.9	44.1(-45.8)	89.5(-0.4)	91.8(+1.9)	**93.2 (+3.3)**
SQuAD-golden	68.7	97.9	83.2(-14.7)	97.1(-0.8)	98.1(+0.2)	**98.3 (+0.4)**

Table 2: Performance comparison on non-conflicting contexts with GPT-4o-mini as the backbone LLM. The best result on each dataset is highlighted in **bold**. The numbers in parentheses indicates the change in accuracy compared to the standard RAG.

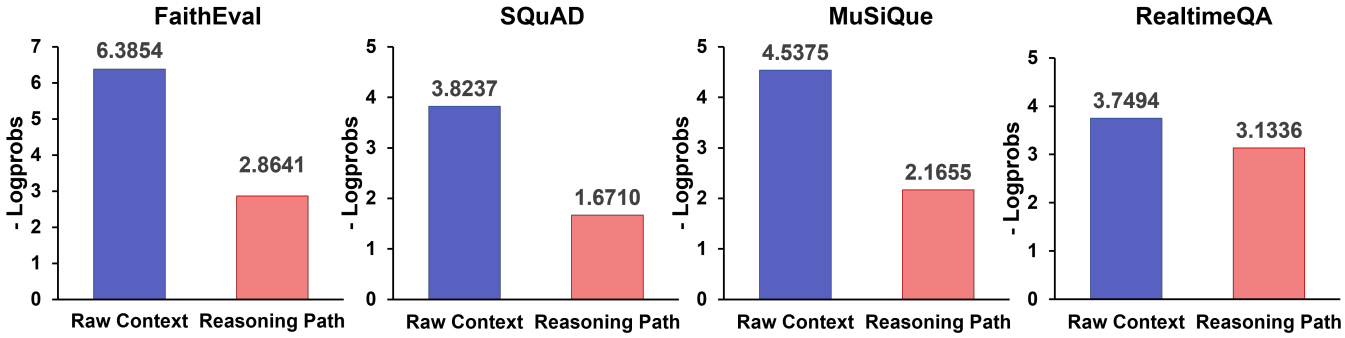

Figure 3: Comparison of LLM confidence, measured by negative log-probability (logprob) values using GPT-4o-mini, when reasoning with natural language contexts versus structured reasoning paths across four datasets. Lower negative logprob values indicate higher actual log-probability scores and thus increased model confidence in generating correct answers.

Performance on Non-Conflicting Contexts To evaluate the robustness of TruthfulRAG in scenarios where retrieved contexts free from factual conflicts, we conduct experiments on golden standard datasets in which the retrieved passages are guaranteed to be non-contradictory.

As shown in Table 2, TruthfulRAG consistently outperforms all baseline methods across both the MuSiQue-golden and SQuAD-golden datasets. These findings substantiate that TruthfulRAG not only excels at resolving conflicting information but also maintains superior performance in non-conflicting contexts, thereby revealing its universal applicability and effectiveness. The consistent performance improvements can be attributed to the structured knowledge representation provided by the knowledge graph module, which enables the identification of fine-grained entities and relational links in non-conflicting contexts. This capability facilitates the extraction of query-relevant information and promotes a more comprehensive understanding and integration of factual knowledge by the LLMs. Notably, while methods such as KRE exhibit significant performance degradation in non-conflicting scenarios, TruthfulRAG maintains its robustness across diverse contextual settings. This consistency highlights its practical utility and reliability for real-world RAG applications.

Impact of Structured Reasoning Paths To investigate the impact of structured reasoning paths on the confidence of LLMs relative to raw natural language context, we conduct a comprehensive analysis across four datasets. Specifically, we compare the model's confidence when reasoning with retrieved knowledge presented in natural language format or as structured reasoning paths derived through our knowledge graph construction mechanism. To quantify the model's confidence in its predicted answers, we measure the log-probability of the correct answer tokens generated by LLMs and compute the average across all test instances.

As shown in Figure 3, our experimental results reveal a consistent pattern across all evaluated datasets. Structured reasoning paths consistently lead to higher logprob values for correct answers compared to natural language contexts, indicating greater model confidence when reasoning with structured knowledge representations. This empirical evidence demonstrates that transforming unstructured natural language into structured reasoning paths through knowledge graphs significantly strengthens the LLM's confidence in following external retrieved knowledge for inference. Furthermore, this finding provides crucial insights into the superior performance of TruthfulRAG in both conflicting and non-conflicting semantic scenarios, as the enhanced confidence facilitates more reliable adherence to external knowledge sources, thereby supporting factual consistency and promoting the generation of faithful model outputs.

Ablation Study To comprehensively evaluate the contribution of each component in TruthfulRAG, we conduct systematic ablation experiments by removing key modules from the full framework. Since knowledge graph construction and retrieval are two closely coupled modules, we combine them as an integrated component for ablation evaluation.

As shown in table 3, the complete TruthfulRAG framework achieves superior performance across all datasets, with accuracy improvements ranging from 6.8% to 17.7% compared to the standard RAG, demonstrating that the structured knowledge graph and the conflict resolution mecha-

Method	Dataset			
	FaithEval	MuSiQue	RealtimeQA	SQuAD
Standard RAG	61.3 / 0.51	72.6 / 1.86	67.3 / 0.47	73.1 / 2.71
w/o Knowledge Graph	64.8 / 0.52	78.9 / 1.15	83.2 / 0.23	78.8 / 1.97
w/o Conflict Resolution	69.3 / 0.59	77.8 / 2.79	84.1 / 1.80	78.2 / 2.85
Full Method	69.5 / 0.56	79.4 / 2.25	85.0 / 1.54	81.1 / 2.56

Table 3: Ablation study results of different components in TruthfulRAG with GPT-4o-mini as the backbone LLM. The results are presented in the format ACC / CPR, where ACC denotes accuracy and CPR represents Context Precision Ratio.

nism function synergistically to enhance both factual accuracy and contextual precision. The ablation results reveal several critical insights. First, when employing only the filtering mechanism (w/o Knowledge Graph), although accuracy demonstrates modest improvements, CPR exhibits a notable decline across most datasets. This phenomenon indicates that LLMs encounter substantial difficulties in effectively extracting relevant information from naturally organized contexts, thereby constraining their ability to achieve higher accuracy. In contrast, when using only the knowledge graph component (w/o Conflict Resolution), CPR yields notable gains, yet the incorporation of extensive structured knowledge also introduces redundant information, resulting in limited accuracy improvements on most datasets. These results support our hypothesis that structured knowledge representations facilitate the precise localization of query-relevant information, enabling more targeted and effective information extraction compared to unstructured contexts.

Related Work

Impact Analysis of Knowledge Conflicts

Recent studies have extensively explored the influence of knowledge conflicts on the performance of RAG systems (Longpre et al. 2021; Chen, Zhang, and Choi 2022; Xie et al. 2023; Tan et al. 2024; Ming et al. 2025), which primarily highlight differential preferences between the parametric knowledge and retrieved external information. Longpre et al. (Longpre et al. 2021) first expose entity-based knowledge conflicts in question answering, revealing that LLMs tend to rely on parametric memory when retrieved passages are perturbed or contain contradictory information. Chen et al. (Chen, Zhang, and Choi 2022) demonstrate that while retrieval-based LLMs predominantly depend on non-parametric evidence when recall is high, their confidence scores fail to reflect inconsistencies among retrieved documents. Xie et al. (Xie et al. 2023) find that LLMs are receptive to single external evidence, yet exhibit strong confirmation bias when presented with both supporting and conflicting information. Tan et al. (Tan et al. 2024) reveal a systematic bias toward self-generated contexts over retrieved ones, attributing this to the higher query-context similarity and semantic incompleteness of retrieved snippets.

Our work aligns with the non-parametric knowledge preference paradigm, aiming to guide LLMs to follow updated and comprehensive external knowledge while correcting for temporal and factual errors within internal memory, thereby generating accurate and trustworthy outputs.

Solutions to Knowledge Conflicts

Current approaches for knowledge conflict resolution can be categorized into token-level and semantic-level methods (Jin et al. 2024; Wang et al. 2024; Bi et al. 2025; Zhang et al. 2025; Wang et al. 2025). Token-level approaches focus on fine-grained intervention during generation. CD^2 (Jin et al. 2024) employs attention weight manipulation to suppress parametric knowledge when conflicts are detected. ASTUTE RAG (Wang et al. 2024) utilizes gradient-based attribution to identify and mask conflicting tokens during inference. These methods achieve precise control, but often suffer from computational overhead and lack semantic awareness among generated contents. Semantic-level approaches operate at higher abstraction levels. CK-PLUG (Bi et al. 2025) develops parameter-efficient conflict resolution through adapter-based architectures that learn to weight parametric versus non-parametric knowledge dynamically. FaithfulRAG (Zhang et al. 2025) externalizes LLMs' parametric knowledge and aligns it with retrieved context, thereby achieving higher faithfulness without sacrificing accuracy. However, these methods primarily address surface-level conflicts without capturing the underlying factual relationships that drive knowledge inconsistencies.

Different from these approaches, TruthfulRAG leverages structured triple-based knowledge representations to precisely identify and resolve factual-level knowledge conflicts arising from complex natural language expressions, thereby ensuring the reliability and consistency of reasoning.

Conclusion

In this paper, we introduce TruthfulRAG, the first framework that uses knowledge graphs to address factual-level conflicts in RAG systems. Through triple extraction, query-aware graph retrieval, and entropy-based filtering, it converts unstructured contexts into structured reasoning paths that enhance LLMs' confidence in external knowledge while effectively mitigating factual inconsistencies. Our experiments demonstrate that TruthfulRAG consistently surpasses existing SOTA methods, establishing it as a robust and generalizable solution for improving the trustworthiness and accuracy of RAG systems, with broad implications for reliability-critical, knowledge-intensive applications.

Acknowledgements

This work is supported by Funding for Major Science and Technology Breakthrough Projects in Hunan Province (No. 2025QK2009), the National Natural Science Foundation of China No. 62402060, Beijing Natural Science Foundation, No. 4244083.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arora, U.; Huang, W.; and He, H. 2021. Types of Out-of-Distribution Texts and How to Detect Them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10687–10701.

Bi, B.; Liu, S.; Wang, Y.; Xu, Y.; Fang, J.; Mei, L.; and Cheng, X. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Chen, H.-T.; Zhang, M. J.; and Choi, E. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; and Xu, K. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5050–5063.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 6491–6501.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; Halappanavar, M.; Rossi, R. A.; Mukherjee, S.; Tang, X.; et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Li, Q.; and Zhao, J. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Kasai, J.; Sakaguchi, K.; Le Bras, R.; Asai, A.; Yu, X.; Radev, D.; Smith, N. A.; Choi, Y.; Inui, K.; et al. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36: 49025–49043.

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

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Ming, Y.; Purushwalkam, S.; Pandit, S.; Ke, Z.; Nguyen, X.-P.; Xiong, C.; and Joty, S. 2025. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". In *The Thirteenth International Conference on Learning Representations*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Shi, D.; Jin, R.; Shen, T.; Dong, W.; Wu, X.; and Xiong, D. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37: 4997–5024.

Tan, H.; Sun, F.; Yang, W.; Wang, Y.; Cao, Q.; and Cheng, X. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*.

Touvron, H.; and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Wang, F.; Wan, X.; Sun, R.; Chen, J.; and Arık, S. Ö. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

Wang, J.; Xu, Z.; Jin, D.; Yang, X.; and Li, T. 2025. Accommodate Knowledge Conflicts in Retrieval-augmented LLMs: Towards Reliable Response Generation in the Wild. *arXiv preprint arXiv:2504.12982*.

Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ying, J.; Cao, Y.; Xiong, K.; Cui, L.; He, Y.; and Liu, Y. 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)*, 4221–4246.

Yuan, X.; Yang, Z.; Wang, Y.; Liu, S.; Zhao, J.; and Liu, K. 2024. Discerning and Resolving Knowledge Conflicts through Adaptive Decoding with Contextual Information-Entropy Constraint. In *Findings of the Association for Computational Linguistics ACL 2024*, 3903–3922.

Zhang, Q.; Dong, J.; Chen, H.; Zha, D.; Yu, Z.; and Huang, X. 2024. Knowgpt: Knowledge graph based prompting for large language models. *Advances in Neural Information Processing Systems*, 37: 6052–6080.

Zhang, Q.; Xiang, Z.; Xiao, Y.; Wang, L.; Li, J.; Wang, X.; and Su, J. 2025. FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. *arXiv preprint arXiv:2506.08938*.