FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation

Qinggang Zhang[1*], **Zhishang Xiang**[2*], **Yilin Xiao**[3], **Le Wang**[4], **Junhui Li**[5],
Xinrun Wang[6], **Jinsong Su**[1,7†]

[1]School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
[3]The Hong Kong Polytechnic University [4]Migu Meland Co., Ltd [5]Soochow University
[6]Singapore Management University [7]Shanghai Artificial Intelligence Laboratory
{zqg.zhang, xzs.xiang}@hotmail.com jssu@xmu.edu.cn

Abstract

Large language models (LLMs) augmented with retrieval systems have demonstrated significant potential in handling knowledge-intensive tasks. However, these models often struggle with unfaithfulness issues, generating outputs that either ignore the retrieved context or inconsistently blend it with the LLM's parametric knowledge. This issue is particularly severe in cases of knowledge conflict, where the retrieved context conflicts with the model's parametric knowledge. While existing faithful RAG approaches enforce strict context adherence through well-designed prompts or modified decoding strategies, our analysis reveals a critical limitation: they achieve faithfulness by forcibly suppressing the model's parametric knowledge, which undermines the model's internal knowledge structure and increases the risk of misinterpreting the context. To this end, this paper proposes FaithfulRAG, a novel framework that resolves knowledge conflicts by explicitly modeling discrepancies between the model's parametric knowledge and retrieved context. Specifically, FaithfulRAG identifies conflicting knowledge at the fact level and designs a self-thinking process, allowing LLMs to reason about and integrate conflicting facts before generating responses. Extensive experiments demonstrate that our method outperforms state-of-the-art methods. The code is available at https://github.com/DeepLearnXMU/Faithful-RAG.

1 Introduction

Large language models (LLMs), like GPT (OpenAI, 2023), Claude (Anthropic, 2024) and DeepSeek series (Liu et al., 2024a), have surprised the world with superior performance in many real-world applications (Hong et al., 2024; Yuan et al., 2025; Zhou et al., 2024, 2025). Despite their effectiveness, LLMs are always criticized for

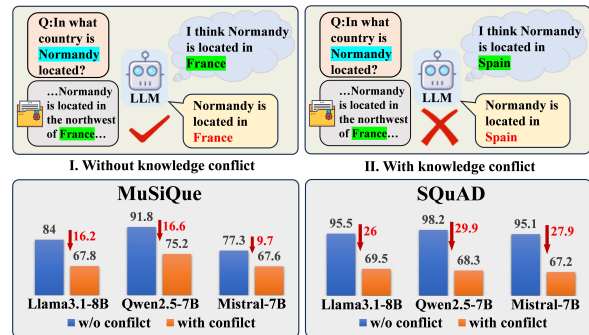---
[*]Equal contribution.
[†]Corresponding author.

Figure 1: The running example of knowledge conflict and the performance comparison of different LLMs in scenarios with and without knowledge conflicts.

their limited ability to handle knowledge-intensive tasks (Huang et al., 2023; Chen et al., 2024), especially when faced with questions requiring professional or private knowledge not covered in their training corpus (Zhang et al., 2024). Retrieval-augmented generation (RAG) (Zhang et al., 2025; Gao et al., 2023; Cao et al., 2024; Liu et al., 2024b) offers a promising solution to customize LLMs for specific domains. Rather than retraining LLMs to incorporate new knowledge (Zhang et al., 2023) and updates (Wang et al., 2024), RAG enhances language models by leveraging external knowledge without modifying the model architecture or parameters. This approach enables LLMs to generate responses by leveraging not only their parametric knowledge (i.e., the knowledge embedded in its parameters from training) but also real-time retrieved domain-specific information, thereby providing more accurate and reliable answers.

Despite recent advances, empirical studies (Ming et al., 2025; Yuan et al., 2024) have revealed that RAG systems struggle significantly in knowledge conflict scenarios (Xu et al., 2024b) where the retrieved context contradicts the parametric knowledge that the model has learned during pre-training. Such conflicts can lead to severe unfaithfulness issues (Wu et al., 2024), where the model either generates a response that contradicts

the context or fails to incorporate crucial details from the retrieved contexts. The results in Figure 1 provide clear evidence of this issue that LLMs struggle to maintain faithfulness and correctness when faced with conflicting information.

Recently, a few studies (Ying et al., 2024; Yuan et al., 2024) have been explored to improve the faithfulness of RAG systems. These works can be roughly categorized into two main directions: (i) Prompting-based methods that explore various prompting strategies to guide RAG systems to generate contextually faithful responses. By providing explicit instructions or few-shot examples, models can be directed to prioritize retrieved information over parametric knowledge (Zhou et al., 2023; Ying et al., 2024). (ii) Decoding-based models achieve faithfulness by forcing LLMs to align closely with the provided context during the decoding process by modifying the underlying generation mechanism through entropy-based constraints (Yuan et al., 2024) or contrastive decoding (Shi et al., 2023).

However, we identified a critical limitation in existing faithful methods through a thorough analysis: while they can achieve context faithfulness, they often do so at the cost of increased risk of misinterpreting the context. As evidenced by our experimental results in Figure 2, state-of-the-art context-faithful methods reduce unfaithful errors by 6.65% but simultaneously cause a 6.42% rise in incorrect match errors on average. This occurs because existing methods (Zhou et al., 2023; Ying et al., 2024) attempt to achieve faithfulness by forcibly suppressing the model's confidence in its parametric knowledge without properly understanding and analyzing the differences between contextual information and the model's inherent knowledge. Such suppression compromises the model's ability to critically evaluate and reconcile discrepancies, leading to degraded comprehension, logical incoherence, and a higher likelihood of aligning with incorrect contextual information.

In this paper, we propose a novel RAG model (FaithfulRAG) that achieves context faithfulness while maintaining accurate knowledge integration. Our research aims to address two fundamental research questions: (i) how to precisely locate conflicting knowledge between the model's parametric knowledge and the retrieved context, and (ii) how we could guide the language model to pay more attention to these critical segments during generation. Our major contribution is listed as follows:

- We identify the key limitation of existing context-faithful methods and propose FaithfulRAG to improve the faithfulness of RAG systems without sacrificing accuracy.

- FaithfulRAG adopts a novel self-fact mining module to externalize the LLM's understanding of the question, obtaining fine-grained knowledge at the fact level.

- FaithfulRAG identifies conflicting knowledge by aligning self-fact with contexts and designs a self-thinking module, allowing LLMs to reason about and integrate conflicting facts before generating responses.

- Experiments show that FaithfulRAG outperforms state-of-the-art models, achieving more accurate and faithful generation.

2   Problem Statement

Retrieval-augmented generation (RAG) systems consist of two key stages (Lewis et al., 2020): (i) **Knowldge Retrieval**, in which the model identifies and extracts semantically relevant documents from external knowledge sources based on the user query, which serves as the contextual foundation for subsequent processing. (ii) **Generation Stage**: The retrieved context is synthesized with the user query to construct an augmented prompt, which LLMs process to generate a contextually grounded and factually consistent output.

In our work, we focus on the generation stage and guide LLM to generate more faithful responses by accurately interpreting the context with the model's parametric knowledge.

$$P(a \mid Q, \mathcal{C}) = \prod_{t=1}^{T} P(a_t \mid a_{<t}, Q, \mathcal{C}; \theta) \quad (1)$$

While Q represents the user query, $\mathcal{C}$ represents the context, and a is the target answer. t denotes the position of the currently generated token, T refers to the total number of tokens in the a, θ represents the parameters of the generator model.

3   Preliminary Study

Before going into the technique details of FaithfulRAG, we first conduct a preliminary study to identify the primary limitation of existing methods.
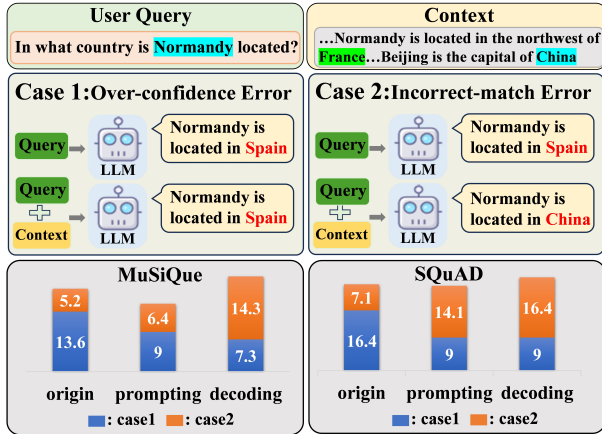
Figure 2: The cases of errors and their distribution on MuSiQue and SQuAD datasets. The detailed experimental setup is introduced in Section A.4 of Appendix.

Performance Degradation in Knowledge Conflict Scenarios. RAG systems often struggle with severe unfaithfulness issues, especially in knowledge conflict scenarios where retrieved contexts contradict the model's parametric knowledge (Longpre et al., 2021; Zhou et al., 2023; Xu et al., 2024b). As shown in Figure 1, our experiments on two real-world datasets reveal that LLMs always illustrate a significant performance gap between scenarios with and without knowledge conflicts. Specifically, we evaluate the performance of Llama3.1-8b-instruct, Qwen2.5-7b-instruct, and Mistral-7b-instruct on the MuSiQue and SQuAD datasets (Ying et al., 2024). Without knowledge conflicts, Qwen2.5-7b-instruct achieves the highest factual accuracy, followed by Llama3-8b-instruct, while Mistral-7b-instruct has the lowest scores. However, when knowledge conflict is introduced, all models experience a significant drop in performance, with drops ranging from 9.7% to 29.9%.

Error Analysis. Through an in-depth analysis of the erroneous outputs in knowledge conflict scenarios, we found that the performance degradation is attributed to two dominant failure modes:

Case 1: Over-confidence Error. The LLM insists on its inherent parametric knowledge while ignoring the facts in the context, leading to unfaithful responses (Ying et al., 2024; Xie et al., 2024).

Case 2: Incorrect-match Error. The LLM updates its parametric knowledge but incorrectly learns from the misleading context.

To explore the underlying reasons, we evaluate state-of-the-art faithful RAG methods on the MuSiQue and SQuAD benchmarks (Ying et al., 2024) and present the error distribution in Figure 2. The results reveal a critical limitation of existing

models that they improve faithfulness at the cost of increasing the risk of misinterpreting the context. Specifically, we have the following observations:

Obs. 1. The vanilla RAG model (origin) exhibits a high proportion of Case 1 errors, with 13.6% on MuSiQue and 16.4% on SQuAD, while Case 2 errors remain relatively low at 5.2% and 7.1%, respectively. This suggests that the model primarily suffers from LLMs' inherent bias toward their parametric knowledge. Such a tendency aligns with prior findings (Xie et al., 2024) that LLMs tend to prioritize pre-trained knowledge over contextual evidence when faced with conflicting information.

Obs. 2. The prompting-based faithful RAG model (prompting) reduces Case 1 errors significantly, from 13.6% to 9% in MuSiQue and from 16.4% to 9% in SQuAD. However, this comes at the cost of a substantial increase in Case 2 errors, which rise from 5.2% to 6.4% in MuSiQue and from 7.1% to 14.1% in SQuAD, reflecting an overcorrection where the model prioritizes context faithfulness without adequately validating the relevance or accuracy of the retrieved information.

Obs. 3. The decoding-based faithful RAG model (decoding) further decreases Case 1 errors, reducing them to 7.3% in MuSiQue and 9% in SQuAD. Compared to prompting-based approaches, it is more effective at mitigating over-confidence issues. However, this improvement comes with an even greater rise in Case 2 errors, which jump to 14.3% in MuSiQue and 16.4% in SQuAD. This suggests that the decoding-based model imposes stronger constraints on faithfulness but at the expense of a heightened risk of misinterpreting context.

Discussion. Existing faithful RAG models (Zhou et al., 2023; Yuan et al., 2024) reduce over-confidence errors but simultaneously cause a significant rise in incorrect match errors. This occurs because existing methods attempt to achieve faithfulness by forcibly suppressing the model's confidence in its parametric knowledge without properly understanding and analyzing the discrepancies between context and the model's inherent knowledge. Such suppression weakens the model's ability to critically evaluate interactions between parametric knowledge and contextual evidence, leading to either rigid adherence to outdated parametric knowledge or severe contextual overfitting, where the model passively adopts retrieved claims without critically evaluating their correctness.
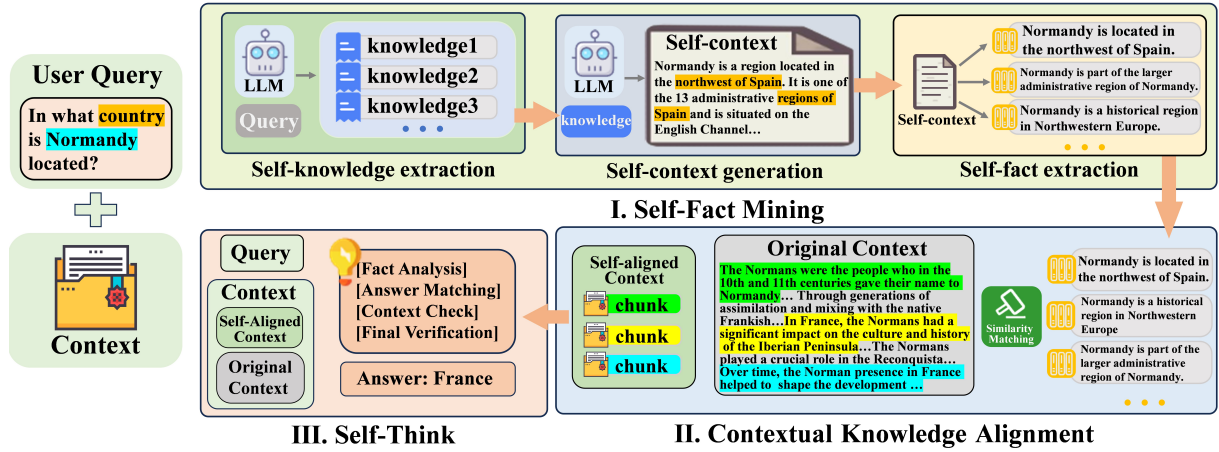
Achieving true faithfulness requires identifying

Figure 3: The overall pipeline of our FaithfulRAG framework. FaithfulRAG first designs a self-fact mining module to externalize the LLM's understanding of the question, obtaining fine-grained knowledge at the fact level. Then it identifies conflicting knowledge by aligning self-fact with the retrieved contexts and adopts a self-thinking module, allowing LLMs to reason about and integrate conflicting facts before generating responses.

specific conflicting facts rather than broadly suppressing knowledge. By prioritizing why conflicts occur and how to resolve them at the fine-grained fact level, FaithfulRAG advances beyond brute-force suppression, enabling LLMs to act as critical, context-aware reasoners rather than passive knowledge retrievers, ensuring outputs are both contextually aligned and logically consistent.

4 Methodology

In this section, we introduce our FaithfulRAG in detail. As shown in Figure 3, FaithfulRAG consists of three main components: (i) **Self-Fact Mining**, which is used to extract self-facts (query-related knowledge) by externalizing the LLM's parametric knowledge. (ii) **Contextual Knowledge Alignment**, that leverages self-facts to identify the most relevant information from the provided context. and (iii) **Self-Think**, enables the model to handle discrepancies between the context and its parametric knowledge at the fact level.

4.1 Self-Fact Mining

To identify conflicts between a model's stored knowledge and the contextual information it receives, we need a clear representation of the LLM's internal understanding of the problem at hand. This involves not only capturing the factual content the model possesses but also revealing the logical structure it uses to organize these facts. Our approach addresses both aspects by employing a hierarchical, three-stage process that externalizes the model's internal knowledge into distinct logical and fine-grained factual representations.

Specifically, for a given question Q, our objective is to extract the essential factual and conceptual information required to answer Q while filtering out unnecessary details. We achieve this through the following three sequential stages: Self-Knowledge Extraction, Self-Context Generation and Self-Fact Extraction.

Self-Knowledge Extraction. It surfaces the LLM's implicit logical structure by identifying required knowledge domains and their interdependencies. Specifically, we prompt the LLM to identify the conceptual and factual prerequisites for answering Q, yielding a set of abstract, high-level knowledge:

$$\mathcal{K}_{\text{self}}(Q) = \{k_1, k_2, \ldots, k_n\}, \qquad (2)$$

where $\mathcal{K}_{\text{self}}(Q)$ represents the extracted self-knowledge and where k_i represents abstract claims.

Self-Context Generation. Then, we translate abstract conceptual mappings into concrete narratives, ensuring alignment between high-level reasoning and specific factual claims. Specifically, the model synthesizes the abstract claims into a coherent narrative that contextualizes the target question. In contrast to previous works (Tan et al., 2024; Yu et al., 2023), which often generate context without any grounding, we explicitly condition the context generation on $\{k_1, k_2, \ldots, k_n\}$. This allows for more coherent and logically consistent context, while ensuring relevance. Formally, the self-context is generated via a generator $\mathcal{G}_1$ as follows:

$$\mathcal{C}_{\text{self}}(Q) = \mathcal{G}_1(Q, \mathcal{K}_{\text{self}}(Q)). \qquad (3)$$

Self-Fact Extraction. The self-context is distilled into concrete factual assertions using the LLM as a fact extractor:

$$\mathcal{F}_{\text{self}}(\mathcal{C}_{\text{self}}) = \{f_1, f_2, \ldots, f_m\}. \qquad (4)$$

These self-facts serve as anchors for aligning with context while preserving logical constraints. By decoupling what the model knows from how it organizes knowledge, the framework supports differentiated error diagnosis and targeted corrections.

4.2 Contextual Knowledge Alignment

To resolve knowledge conflicts while preserving logical coherence, in this section, FaithfulRAG aligns the model's externalized self-facts with the retrieved context through a structured, interpretable process. The alignment is structured as follows:
Context Chunking: The original context $\mathcal{C}_{\text{orig}}$ is divided into a set of chunks $\mathcal{C}_{\text{orig}}^i$, where $i = 1, 2, \ldots, m$. Smaller chunks allow granular comparison with self-facts, reducing noise from irrelevant text spans. Formally:

$$\mathcal{C}_{\text{orig}} = \bigcup_{i=1}^{m} \mathcal{C}_{\text{orig}}^i. \qquad (5)$$

Similarity Matching: We first embed self-facts f_i extracted from the LLM's parametric knowledge and the chunks $\mathcal{C}_j$ divided from the context into a shared semantic space, and then measure their semantic distance by using cosine similarity:

$$Sim(\mathbf{f}_i, \mathbf{c}_j) = \cos(\mathbf{f}_i, \mathbf{c}_j). \qquad (6)$$

Then, we select the chunks $\mathcal{C}_{\text{aligned}}$ based on similarity scores, where $\mathcal{C}_{\text{aligned}}$ represents the chunks that are highly semantically aligned with self-facts that extracted from LLM's parametric knowledge.

4.3 Self-Think

To resolve knowledge conflicts while ensuring contextual faithfulness, FaithfulRAG employs a Self-Think module that dynamically synthesizes insights from two sources: (i) the self-aligned context $\mathcal{C}_{\text{aligned}}$ (context segments conflicting or aligning with parametric knowledge) and (ii) the original context $\mathcal{C}_{\text{orig}}$. This iterative workflow ensures the model critically evaluates discrepancies, mitigates overconfidence, and integrates evidence transparently. We formalize this procedure as a cognitive function $\mathcal{R}_{\text{STR}}$, which synthesizes key insights by

comparing and merging relevant information from both contexts. Let the answer be generated as:

$$\text{Answer} = \mathcal{R}_{\text{STR}}(\mathcal{C}_{\text{aligned}}, \mathcal{C}_{\text{orig}}). \qquad (7)$$

Specifically, this process can be divided into two parts: thinking and reasoning. In the thinking stage, LLM first produces an initial answer from $\mathcal{C}_{\text{aligned}}$. It subsequently evaluates the reliability of this answer and ascertains whether $\mathcal{C}_{\text{aligned}}$ provides sufficient information. If the answer is found to be unreliable or incomplete, the model selectively incorporates relevant elements from $\mathcal{C}_{\text{orig}}$ to enrich the aligned context, thereby creating a fused context, $\mathcal{C}_{\text{fused}}$. This operation is defined as:

$$\mathcal{C}_{\text{fused}} = \mathcal{G}_2(\mathcal{C}_{\text{aligned}}, \mathcal{C}_{\text{orig}}), \qquad (8)$$

where $\mathcal{G}_2$ represents the context fusion function. In the reasoning module, the final answer is regenerated from $\mathcal{C}_{\text{fused}}$ through a step-by-step reasoning procedure, ensuring that the answer is both coherent and adequately supported by the evidence.

To make it more clear, the prompt templates applied in FaithfulRAG for self-knowledge extraction, self-context generation, self-fact extraction, and self-think are shown in Figure 6 of Appendix.

5 Experiment

In this section, we conduct comprehensive experiments to verify the effectiveness of FaithfulRAG. Specifically, we aim to answer the following questions. **Q1 (Effectiveness):** How does FaithfulRAG perform compared with SOTA competitors? **Q2 (Error analysis):** How effective is FaithfulRAG in alleviating different types of errors? **Q3 (Ablation study:)** How does each component of FaithfulRAG contribute to the performance? **Q4 (Case study):** How does it works in real-world scenarios? (Note that **Q4** is studied in Appendix B.1, while the first three questions are explored in the main content.)

5.1 Experiment Settings

Datasets: We evaluate FaithfulRAG on four benchmark datasets. **MuSiQue** (Trivedi et al., 2022) and **SQuAD** (Rajpurkar et al., 2016) are from KRE (Ying et al., 2024) which introduce fact-level knowledge conflicts, where only contradictory factual statements appear in the context. In contrast, **FaithEval** (Ming et al., 2025) introduces logical-level conflicts, where inconsistencies arise not from direct factual contradictions but from reasoning chains that lead to conflicting conclusions. We also

Table 1: The comparison of performance between our model and SOTA baselines on four datasets. The best result for each dataset is highlighted in **bold**, while the best result for each backbone model is indicated with an <u>underline</u>.

Model	Backbone LLM	Dataset			
		FaithEval	RealtimeQA	MuSiQue	SQuAD
Group 1: Default Methods					
Origin model without context	llama3.1-8b-instruct	7.6	28.3	11.2	11.2
	qwen2.5-7b-instruct	4.2	40.7	19.6	11.1
	mistral-7b-instruct	6.3	29.2	13.8	11.5
Origin model with full context	llama3.1-8b-instruct	63.3	67.3	67.8	69.5
	qwen2.5-7b-instruct	53.1	78.7	75.2	68.3
	mistral-7b-instruct	61.9	52.2	67.6	67.2
Group 2: Specific RAG Models					
Self-RAG (Asai et al., 2023)	Llama2-7B	37.4	55.8	54.1	62.0
ChatQA-1.5 (Liu et al., 2024d)	Llama3.1-8B	56.2	56.7	75.0	77.0
ChatQA-2.0 (Xu et al., 2024a)	Llama3.1-8B	65.2	57.5	77.2	75.4
Group 3: Context-faithful Prompting					
Opin(Instr) (Zhou et al., 2023)	llama3.1-8b-instruct	68.1	75.2	70.3	73.4
	qwen2.5-7b-instruct	56.7	81.4	76.9	70.5
	mistral-7b-instruct	62.5	51.3	68.1	69.3
ATTR (Zhou et al., 2023)	llama3.1-8b-instruct	63.8	76.9	62.8	69.5
	qwen2.5-7b-instruct	58.1	83.0	<u>78.7</u>	72.9
	mistral-7b-instruct	63.6	52.2	66.1	70.2
KRE (Ying et al., 2024)	llama3.1-8b-instruct	51.6	48.6*	35.9*	66.1
	qwen2.5-7b-instruct	59.6	**<u>86.7</u>**	70.7	73.7
	mistral-7b-instruct	73.2	76.9	50.6*	74.6
Group 4: Context-faithful Decoding					
CAD (Shi et al., 2023)	llama3.1-8b-instruct	66.2	61.9	72.6	71.2
	qwen2.5-7b-instruct	60.5	77.0	78.6	73.4
	mistral-7b-instruct	60.2	55.8	63.6	66.9
COIECD (Yuan et al., 2024)	llama3.1-8b-instruct	67.7	62.8	70.5	71.8
	qwen2.5-7b-instruct	62.3	78.8	69.7	70.8
	mistral-7b-instruct	62.8	58.4	66.8	65.4
FaithfulRAG (Ours)	llama3.1-8b-instruct	<u>79.8</u>	<u>81.4</u>	**79.9**	**86.3**
	qwen2.5-7b-instruct	<u>71.8</u>	84.1	78.0	<u>78.3</u>
	mistral-7b-instruct	**<u>81.7</u>**	<u>77.0</u>	<u>78.5</u>	<u>85.7</u>

* There is a sharp decline in performance as the model refuses to generate responses.

include **RealtimeQA** (Kasai et al., 2024) dataset to test the model performance in extreme cases where some contexts are irrelevant to the question. More details of datasets can be found in Appendix C.1.

Baselines: We carefully select baselines from four categories for a comprehensive evaluation. **Default Methods**: origin model without context, origin model with full context; **RAG models**: Self-RAG (Asai et al., 2023), ChatQA-1.5 (Liu et al., 2024d), ChatQA-2.0 (Xu et al., 2024a); **Context-faithful Prompting**: Opin (Zhou et al., 2023), KRE (Ying et al., 2024); and **Context-faithful Decoding**: CAD (Shi et al., 2023), COIECD (Yuan et al., 2024). More details are described in Sec-

tion C.2 of the Appendix. Additionally, we did not include SFR-RAG (Nguyen et al., 2024) as a baseline since its parameters have not been released.

Evaluation Metrics and Implementation Details: Following previous studies, we evaluate all models using accuracy (ACC), where a model's response is considered correct only if it contains the ground truth answer. For the MuSiQue and SQuAD, we add M_R (Memorization Ratio) (Longpre et al., 2021) to measure context faithfulness. To ensure reproducibility, all models were evaluated using deterministic decoding (temperature = 0). Further metrics and details are in the Appendix C.3.

Table 2: The experimental results for non-knowledge-conflict scenarios on LLaMA 3.1-8B-Instruct. The numbers in parentheses (e.g., -3.5) indicates the accuracy change compared to the full method and the best result is underlined

| Dataset | Origin | Context-faithful Prompting | | | Context-faithful Decoding | | FaithfulRAG (Ours) |
		Opin (Instr)	ATTR	KRE	CAD	COIECD	
MuSiQue-golden	84	83(-1)	80.4(-3.6)	38.3*	81.9(-3.1)	83.3(-0.7)	<u>85.3</u> (+1.3)
SQuAD-golden	95.2	96(+0.8)	94.2(-1)	81.4(-13.8)	91.1(-4.1)	95.1(-0.1)	<u>96.6</u> (+1.4)

* There is a sharp decline in performance as the model refuses to generate responses.

5.2 Main Results (Q1)

To address **Q1**, we evaluate FaithfulRAG by comparing it to state-of-the-art baselines on four benchmark datasets, with main results shown in Tables 1 and 2, while MR results in the Appendix B.3. We summarize the observations as follows.

Obs. 1. FaithfulRAG consistently outperforms baseline models on four benchmark datasets. On FaithEval, FaithfulRAG with the Mistral-7B backbone achieves the highest score (81.7%), surpassing the strongest baseline (73.2% from KRE) by 8.5%. Similarly, on SQuAD, FaithfulRAG (Llama3.1-8B) scores 86.3%, exceeding the closest competitor (ChatQA-2.0 at 77.0%) by 9.3%. Besides, FaithfulRAG dominates MuSiQue and RealtimeQA with scores of 79.9% and 84.1%, demonstrating its robustness across diverse scenarios.

Obs. 2. FaithfulRAG demonstrates consistent and robust performance across diverse backbone LLMs. When evaluated on Llama3.1-8B, Qwen2.5-7B, and Mistral-7B, FaithfulRAG achieves state-of-the-art results while maintaining minimal performance variance, unlike methods such as KRE or CAD, which exhibit severe instability across different LLMs. This backbone-agnostic efficacy highlights its ability to harmonize parametric and contextual knowledge dynamically, regardless of model architecture and context complexity.

Obs. 3. FaithfulRAG achieves consistent performance in both knowledge conflict and non-conflict scenarios. As shown in Table 2, FaithfulRAG also achieves the highest accuracy in non-conflict scenarios compared to all competitors. On MuSiQue-golden, FaithfulRAG achieves 85.3%, outperforming strongest competitors like ATTR (80.4%) and CAD (81.9%). On SQuAD-golden, it scores 96.6%, surpassing KRE (81.4%) and COIECD (95.1%). This aligns with the requirement for faithful systems to avoid performance degradation in non-conflict scenario (Yuan et al., 2024).

5.3 Error Analysis (Q2)

As discussed in Section 3, current faithful methods enhance faithfulness at the expense of an increased risk of context misinterpretation. In this section, we systematically analyze how effective FaithfulRAG is in alleviating different types of errors. As shown in Figure 5, we have the following observations.

Obs. 4. Existing faithful models achieve desirable performance in alleviating Case 1 errors but at the cost of amplifying Case 2 errors. Specifically, COIECD achieved the best performance for Case 1 optimization, reducing Case 1 errors by an average of 8.6%. However, this improvement came at the cost of a sharp increase in Case 2 errors, with the highest observed rise reaching 12.8%. While Opin(Instr) demonstrated a more balanced performance, reducing Case 1 errors by an average of 6.0%, while Case 2 errors only increased by 2.8%. This trade-off stems from their inability to dynamically reconcile discrepancies between parametric and contextual knowledge.

Obs. 5. FaithfulRAG achieves balanced mitigation of both Case 1 and Case 2 errors, reducing them by 6.8% and 1.6%, respectively. This improvement stems from our well-designed framework, which enables LLMs to dynamically reconcile parametric knowledge with contextual evidence. By isolating discrepancies at the fact level and applying a self-think module, FaithfulRAG preserves high-quality parametric knowledge while systematically rejecting contexts that introduce logical inconsistencies or semantic divergence.

5.4 Alation Study (Q3)

To evaluate the contributions of FaithfulRAG's core components, we systematically ablate three key components: (i) Self-Fact Mining, (ii) Self-Think module, and (iii) Chain-of-Thought (CoT), generating **7** variants as shown in Table 3. Each variant was tested under knowledge-conflict scenarios to isolate its impact on performance. We have the following findings.

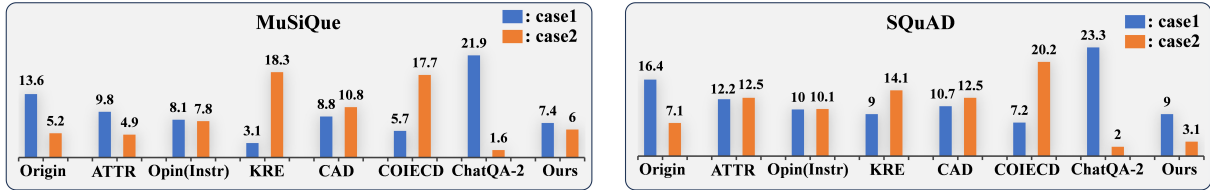**Obs. 6.** The ablation of Self-Knowledge Extraction

Figure 4: The error distribution on MuSiQue and SQuAD datasets with Llama3.1-8b-instruct as the backbone LLM.

Table 3: Ablation study. Numbers in parentheses (e.g., -1.9) represent the change in accuracy relative to full model.

Module/Aspect	Variant	Dataset				Average
		Faitheval	RealtimeQA	MuSiQue	SQuAD	
Knowledge Externalization	w/o Self-Context Generation	77.2	77.9	79.9	85.1	80.0 (-1.90)
	w/o Self-Knowledge Extraction	77.9	80.6	79.3	85.2	80.8 (-1.10)
Self-Think	w/o whole Module	50.3	67.2	63.7	57.8	59.8 (-22.2)
	w/o Think	79.7	69.0	73.7	78.5	75.2 (-6.70)
	w/o Reasoning	79.6	73.5	72.2	78.7	76.0 (-5.90)
CoT Influence	Only CoT	70.2	64.6	52.7	71.8	64.8 (-17.1)
	w/o CoT	82.1	79.6	78.7	78.9	79.8 (-2.10)
Full model	-	79.8	81.4	79.9	86.3	81.9

and Self-Context Generation reveals that both modules are critical for precise knowledge alignment. Removing Self-Knowledge Extraction degrades the model's ability to analyze questions comprehensively, leading to a 1.1% average accuracy drop, as the LLM fails to identify relevant parametric facts. Conversely, removing Self-Context Generation—which converts abstract self-knowledge into actionable context—causes a larger 1.9% accuracy decline, demonstrating that raw parametric claims lack utility without contextual grounding.

Obs. 7. Ablating the full Self-Think module results in a 22.2% average accuracy drop, as simply prepending self-aligned context to the original context fails to resolve conflicts dynamically. Replacing Think stage with Special Annotation (explicitly marking key facts) reduces accuracy by 6.7%, proving that passive highlighting cannot replicate active reasoning. Similarly, substituting structured reasoning with naive Chain-of-Thought (CoT) decreases accuracy by 5.9%, confirming that explicit guidance (e.g., conflict anticipation steps) is essential. These results validate that Self-Think's diagnostic workflow (not just attention mechanisms or generic reasoning) drives reliable conflict resolution.

Obs. 8. While CoT enhances reasoning in some domains, naive application under knowledge conflicts causes a 17.1% accuracy drop, as the LLM grows distrustful of conflicting contexts. FaithfulRAG addresses this by integrating CoT with conflict anticipation: the LLM first identifies potential discrepancies in self-aligned contexts before reason-

ing, mitigating distrust. Notably, even without CoT, FaithfulRAG's accuracy decreases by only 2.1%, demonstrating its robustness. This highlights that CoT must be tailored to handle knowledge conflicts, and FaithfulRAG's structured reasoning framework achieves this adaptation.

To summarize, the combined ablation results reveal that FaithfulRAG's components operate synergistically. Self-Fact Mining provides the foundation for precise fact alignment, while Self-Think enables dynamic conflict resolution. Without either, the model reverts to the limitations of suppression-based methods (e.g., over-reliance on context or parametric knowledge). This interdependence validates the framework's design hypothesis: diagnostic reconciliation of conflicts instead of suppression is key to balancing faithfulness and accuracy.

6 Related Work

Knowledge Conflict. When encountering an external context with conflicting knowledge, an LLM tends to ignore such context (Bi et al., 2024a; Longpre et al., 2021; Bi et al., 2024b; Jiang et al., 2025; Fang et al., 2025). A study indicates that the greater the divergence between retrieved information and the model's prior knowledge, the more likely the model is to ignore the retrieved information. (Wu et al., 2024). Another study (Xie et al., 2024) further points out that when provided with both supporting and opposing evidence against their parametric memory, LLMs exhibit strong confirmation bias, tending to adhere to their parametric knowl-

edge. Additionally, the study (Ming et al., 2025) proposes the FaithEval framework to assess the faithfulness of LLMs across different contextual scenarios, revealing that even state-of-the-art models still struggle with the counterfactual context.

Context-Faithful Method. Some methods enhance context faithfulness through fine-tuning. KAFT (Li et al., 2023) fine-tune small parameter models like T5 using counterfactual contexts, SFR-RAG (Nguyen et al., 2024)is a 9B models trained with various RAG-domain datasets. However, such fine-tuning approaches require substantial computational resources and are often difficult to transfer to other models. Among non-fine-tuning approaches, one category focuses on carefully designed prompt templates. For example, One method reconstruct the context as the narrator's statement and then inquire about the narrator's viewpoint, encouraging the model to seek external perspectives (Zhou et al., 2023). Another method employs "Role Play" interventions to alter LLMs' decision-making styles when facing knowledge conflicts. (Ying et al., 2024) However, these methods heavily rely on the model's inherent reasoning abilities and often suffer from limited generalizability. Another category of non-fine-tuning approaches modifies the model's decoding strategy to improve its reliance on context. CAD(Shi et al., 2023)contrasts output probabilities with and without context to enhance contrastive decoding, amplifying the contextual probability distribution without accounting for conflicting contexts. In contrast, COIECD(Yuan et al., 2024) detects knowledge conflicts by measuring entropy changes during generation and dynamically adjusts the decoding strategy accordingly.

7 Conclusions

Large language models (LLMs) often generate hallucinations (factually inconsistent or unfaithful contents). While retrieval-augmented generation (RAG) has shown promise in enhancing language models' capabilities through external knowledge integration, maintaining faithfulness to retrieved contexts remains a significant challenge. This paper identifies and analyzes critical unfaithfulness issues that emerge during RAG's generation phase, particularly when the model's parametric knowledge contradicts retrieved information. Existing faithful RAG methods enforce alignment with the retrieved context by suppressing the model's parametric knowledge, but often increase the risk of misin-

terpreting the context. To address these challenges, we propose a novel faithful framework (FaithfulRAG) that explicitly identifies and resolves knowledge conflicts at the fact level. FaithfulRAG first concretizes the model's parametric knowledge at the fact level and then identifies conflicting knowledge by aligning self-fact with contexts. After that, FaithfulRAG designs a self-think module, allowing LLMs to reason about and integrate conflicting facts before generation. Experiments show that FaithfulRAG outperforms the strongest competitors, generating accurate and faithful responses.

Limitations

While FaithfulRAG advances error mitigation in text-based retrieval-augmented generation, its current scope is limited to textual inputs and does not yet support multimodal information (e.g., images, audio, or structured data). Extending the framework to incorporate multimodal input would enable a more comprehensive assessment of how models navigate and resolve conflicts between heterogeneous knowledge sources. Given that real-world information is often multimodal, such an extension could improve FaithfulRAG's ability to detect and reconcile inconsistencies that arise from modality-specific biases or divergent contextual signals.

In addition, multimodal integration would enhance the applicability of the framework to domains where information synthesis from multiple sources is crucial, such as medical diagnosis (where textual reports, imaging data and patient history must be jointly considered), autonomous systems (which rely on the integration of visual, auditory, and textual signals for decision making), and interactive AI (where understanding user intent often involves processing speech, gestures, and textual input). Addressing these challenges would not only improve FaithfulRAG's robustness but also contribute to broader advancements in multimodal retrieval-augmented generation, cross-modal reasoning, and trustworthy AI.

Ethics Statement

We confirm that we have fully complied with the ACL Ethics Policy in this study. Our research utilizes four publicly available datasets:FaithEval, MuSiQue, SQuAD, and RealtimeQA.FaithEval is designed to evaluate LLM and RAG faithfulness across Unanswerable, Inconsistent, and Counterfactual contexts. MuSiQue and SQuAD, sourced from

KRE, assess LLM robustness against knowledge conflicts in modified MRC and commonsense reasoning tasks. RealtimeQA is a dynamic QA dataset for evaluating models' ability to handle real-time information. All datasets used in this study have been extensively employed in retrieval-augmented generation research and do not contain private, sensitive, or personally identifiable information. We carefully select these datasets to ensure ethical compliance and to mitigate potential biases. Our study does not involve the collection or modification of user-generated content, nor does it introduce synthetic data that could lead to unintended misinformation.

Acknowledgements

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

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

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

Zhiwei Cao, Qian Cao, Yu Lu, Ningxin Peng, Luyang Huang, Shanbo Cheng, and Jinsong Su. 2024. Retaining key information under high compression ratios: Query-guided compressor for llms. *Preprint*, arXiv:2406.02376.

Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. 2024. Entity alignment with noisy annotations from large language models. *arXiv preprint arXiv:2405.16806*.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. *ICLR*.

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

Zijin Hong, Zheng Yuan, Hao Chen, Qinggang Zhang, Feiran Huang, and Xiao Huang. 2024. Knowledge-to-sql: Enhancing sql generation with data expert llm. *arXiv preprint arXiv:2402.11517*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. *arXiv preprint arXiv:2406.16008*.

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

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. *ICML*.

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

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2024b. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *Preprint*, arXiv:2312.01714.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024d. Chatqa: Surpassing gpt-4 on conversational qa and rag. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 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*.

Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xong, and Shafiq Joty. 2024. Sfr-rag: Towards contextually faithful llms. *arXiv preprint arXiv:2409.09916*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv e-prints*, pages arXiv–2404.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

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 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, 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.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Yuan, Hao Chen, Zijin Hong, Qinggang Zhang, Feiran Huang, and Xiao Huang. 2025. Knapsack optimization-based schema linking for llm-based text-to-sql generation. *arXiv preprint arXiv:2502.12911*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024. Knowgpt: Knowledge graph based prompting for large language models. *Advances in Neural Information Processing Systems*, 37:6052–6080.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Chuang Zhou, Jiahe Du, Huachi Zhou, Hao Chen, Feiran Huang, and Xiao Huang. 2025. Text-attributed graph learning with coupled augmentations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10865–10876.

Huachi Zhou, Shuang Zhou, Hao Chen, Ninghao Liu, Fan Yang, and Xiao Huang. 2024. Enhancing explainable rating prediction through annotated macro concepts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11736–11748.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Frequently Asked Questions (FAQs)

A.1 Code and Dataset Availability

To promote transparency and reproducibility, we have uploaded our code into the Anonymous GitHub at `https://github.com/DeepLearnXMU/Faithful-RAG.`. The repository includes the source code of FaithfulRAG and scripts for data preprocessing, model training, and evaluation. Additionally, all datasets used in our experiments are also provided or linked within the repository, ensuring that researchers have full access to the resources required to reproduce and extend our work.

A.2 What are advantages of FaithfulRAG?

FaithfulRAG introduces several key advancements over existing RAG systems and tailored designed faithful methods, addressing fundamental limitations while maintaining practical applicability:

Superior performance on knowledge conflict scenarios. Unlike existing faithful methods that reduce unfaithful errors (rigid adherence to parametric knowledge) at the cost of increasing incorrect match errors (blind acceptance of flawed contexts), FaithfulRAG resolves both error types simultaneously. Specifically, it reduces Case 1 errors by 6.8% while simultaneously decreasing Case 2 errors by 1.6% on average as shown in Figure 4. This improvement arises from its fact-level conflict resolution and self-think mechanisms that enable LLMs to preserve high-quality parametric knowledge and reject semantically divergent contexts.

Robustness in Non-Conflict Scenarios. A good faithful model should improve performance in knowledge-conflict scenarios while maintaining desirable performance in non-conflict scenarios. While most existing methods suffer in non-conflict settings (e.g., KRE's 13.8% drop on SQuAD-golden), our FaithfulRAG achieves the best performance on both SQuAD-golden and MuSiQue-golden. It gets 96.6% on SQuAD-golden (+1.4% over origin) and 85.3 on MuSiQue-golden (+1.3%), proving its robustness and capability to avoid overfitting to either LLM's parametric knowledge or incorrect contextual information.

Backbone-Agnostic Consistency. FaithfulRAG achieves stable performance across diverse backbone LLMs (e.g., Llama3.1-8B, Qwen2.5-7B, Mistral-7B), with minimal variance in accuracy. For instance, on the SQuAD dataset, it attains 86.3% accuracy with Llama3.1-8B and 85.7% with Mistral-7B, outperforming task-specific models

---

Algorithm 1 The Workflow of FaithfulRAG.

Input: Question Q, Original context $\mathcal{C}_{\text{orig}}$.
Output: Answer a.

1: **Self-Fact Mining**
2: Extract a set of high-level knowledge from Q;
$\mathcal{K}_{\text{self}}(Q) = \{k_1, k_2, \ldots, k_n\}$
3: Generate context $\mathcal{C}_{\text{self}}(Q)$ grounded in $\mathcal{K}_{\text{self}}(Q)$;
$\mathcal{C}_{\text{self}}(Q) = \mathcal{G}_1(Q, \mathcal{K}_{\text{self}}(Q))$
4: Extract facts $\mathcal{F}_{\text{self}}(Q)$ from $\mathcal{C}_{\text{self}}(Q)$;
$\mathcal{F}_{\text{self}}(\mathcal{C}_{\text{self}}) = \{f_1, f_2, \ldots, f_m\}$
5: **Contextual Knowledge Alignment**
6: Divide the original context $\mathcal{C}_{\text{orig}}$ into fixed-size chunks $\mathcal{C}_{\text{orig}}^j$;
$\mathcal{C}_{\text{orig}} = \bigcup_{i=1}^{m} \mathcal{C}_{\text{orig}}^i$.
7: **for** each fact f in $\mathcal{F}_{\text{self}}(Q)$ **do**
8:    **for** each context chunk c in $\mathcal{C}_{\text{orig}}$ **do**
9:       Compute similarity between f and c
$Sim(\mathbf{f}, \mathbf{c}) = \cos(\mathbf{f}, \mathbf{c})$.
10:    **end for**
11: **end for**
12: Select top-k chunks as $\mathcal{C}_{\text{aligned}}$
13: **Self-Think**
14: **if** $\mathcal{C}_{\text{aligned}}$ is reliable **then**
15:    Generate a based on $\mathcal{C}_{\text{aligned}}$
16: **else**
17:    Generate a using $\mathcal{R}_{\text{STR}}(\mathcal{C}_{\text{aligned}}, \mathcal{C}_{\text{orig}})$
18: **end if**
19: **return** a.

---

like ChatQA-2.0 by 9.3% as shown in Table 1. This consistency ensures reliable deployment across heterogeneous environments in practice.

A.3 Why focusing on the generation stage of RAG, and why is it crucial to maintain the faithfulness of outputs?

Despite recent advances in retrieval mechanisms, a persistent challenge remains: a significant gap exists between the quality of the retrieved context and the factual accuracy of the generated output. While the context retrieved by the model may be highly relevant, the generative models often struggle to faithfully incorporate this context, resulting in generated text that may diverge from factual correctness. Recent empirical evidence reveals the critical gap between retrieval quality and generation faithfulness. The generation stage determines how this context is interpreted and integrated with the model's parametric knowledge. Existing methods

Query	An astronomer observes that a planet rotates faster after a meteorite impact. Which is the most likely effect of this increase in rotation?
Context	Observing planetary changes post-meteorite impacts has always been a critical aspect of studying celestial mechanics. When a substantial meteorite strikes a planet, it can significantly alter the planet's rotation speed. This phenomenon was notably observed with the planet Tranus-4 in a neighboring galaxy...
Self-Knowledge Extraction	**Planetary rotation and angular momentum:** Understanding how a planet's rotation is affected by its angular momentum, which is a measure of an object's tendency to continue rotating. This knowledge is necessary to explain how the meteorite impact could cause the planet to rotate faster. **Conservation of angular momentum:** Recognizing that the conservation of angular momentum is a fundamental principle in physics, which states that the total angular momentum of a closed system remains constant over time. This principle is crucial in understanding how the planet's rotation speed changes after the meteorite impact ...
Self-Context Generation	A meteorite impact can transfer momentum to a planet, potentially increasing its rotation speed. This effect is due to the conservation of angular momentum, a fundamental principle in physics that states the total angular momentum of a closed system remains constant over time. As the meteorite impacts the planet, its momentum is transferred, causing the planet's rotation to accelerate...
Self-Fact Extraction	1.A meteorite impact can transfer momentum to a planet, potentially increasing its rotation speed. 2.The conservation of angular momentum is a fundamental principle in physics that states the total angular momentum of a closed system remains constant over time. 3.The momentum of a meteorite is transferred to a planet upon impact, causing its rotation to accelerate. ...
Contextual Alignment	...Normally, an increase in rotational speed would be closely associated with geophysical implications such as changes in atmospheric dynamics and magnetic field intensities. **Moreover, the sudden acceleration can cause a redistribution of mass within the planet, thus potentially affecting its gravitational pull.**Researchers from the Galactic Federation Science Division have recently published a detailed analysis demonstrating a correlation between rotation speed and gravitational strength...
Self-Think	1. [Fact Analysis]: Facts explicitly state They observed that as Tranus-4's rotation speed increased, the equatorial bulge became more pronounced, effectively increasing the planet's gravitational pull at the equator. 2. [Option Matching]: Option D directly matches the factual declaration 3. [Context Check]: No contextual supplementation needed - Facts provide conclusive evidence 4. [Final Verification]: No conflicting information.
Output	Answer: Planetary gravity will become stronger.

Table 4: A complete pipeline of our method on FaithEval using Llama3.1-8B-Instruct as the backbone model.

often fail at this stage due to (i) Over-confidence: many RAG systems rigidly prioritize the retrieved context by suppressing parametric knowledge, leading to blind acceptance of flawed or conflicting information. (ii) Logical incoherence: Without critical reasoning, models struggle to resolve contradictions, resulting in inconsistent or unsafe outputs. By targeting generation, in this paper, we address the root cause of unfaithfulness and guide the LLM to dynamically reconcile knowledge conflicts.

A.4 How do we get the datasets in scenarios with and without knowledge conflicts?

We use the MuSiQue and SQuAD datasets from the KRE (Ying et al., 2024) to evaluate LLM performance in both knowledge conflict and non-conflict scenarios. These datasets contain two types of contexts: (i) Negative context, where correct entities are replaced with incorrect ones, simulating a knowledge conflict scenario. (ii) Golden context, which remains unaltered and serves as an approximation of a non-knowledge conflict scenario. It is important to note that while the golden context may still contain some degree of knowledge conflict, its occurrence rate is significantly lower, allowing us to approximate it as a non-conflict dataset.

A.5 How do we calculate the distribution of different error cases?

To measure the error distribution, we follow a step-by-step approach to compute the frequency of different error cases. (i) Parametric prediction: We first generate a prediction from the LLM without providing any context, relying solely on its parametric knowledge. (ii) Contextual prediction: We then provide the corresponding context and generate a new prediction based on the context from the LLM. (iii) Case classification: We classify errors by comparing the two predictions:

- Over-confidence Error (Case 1): If the para-

Table 5: The experiment results of error analysis on LLaMA 3.1-8B-Instruct. The best result is underlined.

Method		MuSiQue				SQuAD			
		Case 1	Case 2	Case 3	All	Case 1	Case 2	Case 3	All
		Error rate	Error rate	Error rate	Error rate	Error rate	Error rate	Error rate	Error rate
Origin	-	13.6	5.2	13.3	32.1	16.4	7.1	7	30.5
Prompting	Opin (instr)	8.1	7.8	13.7	29.6	12.2	12.5	5.5	26.6
	ATTR	9.8	4.9	22.4	37.1	10	10.1	6.5	30.2
	KRE	3.1	18.3	49.3*	70.7*	9	14.1	14.9	38
	Average	**9.0**	**6.4**	**18.1**	-	**11.1**	**11.3**	**6.0**	-
Decoding	CAD	8.8	10.8	7.8	27.4	10.7	12.5	5.5	28.7
	COIECD	5.7	17.7	10.7	34.1	7.2	20.2	7.6	35.0
	Average	**7.3**	**14.3**	**9.3**	-	**9**	**16.4**	**6.6**	-
RAGs	ChatQA-2.0	21.9	1.6	9.8	33.3	23.3	2	9.7	35.0
Ours	FaithfulRAG	**7.4**	6	8	**21.4**	9	3.1	3.4	**15.5**
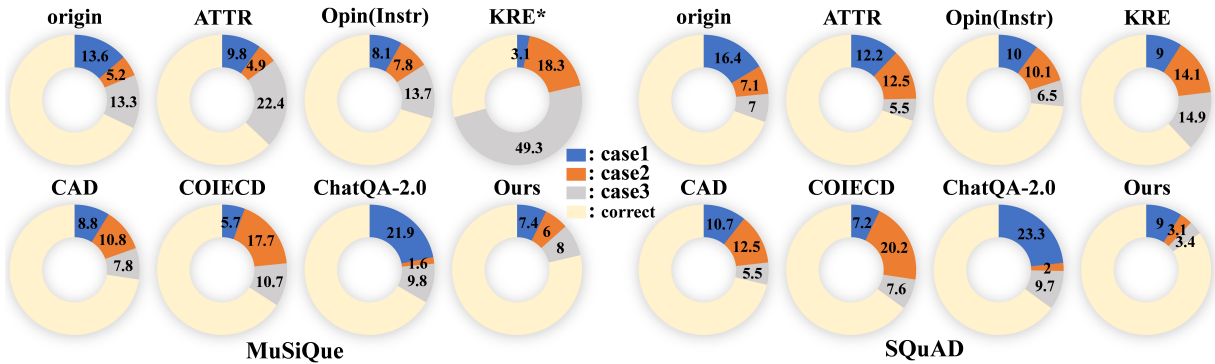
Figure 5: The error distribution on MuSiQue and SQuAD datasets with Llama3.1-8b-instruct as the backbone LLM. Case 1 and Case 2 are consistent with Section 3, while Case 3 represents all other scenarios beyond these two cases.

metric prediction and contextual prediction are identical, the model is considered to have ignored the retrieved context and relied on its internal parametric knowledge.

- Incorrect-match Error (Case 2): If the contextual prediction differs from the parametric prediction, and the prediction appears in the provided context but is incorrect, the model is considered to have been misled by the context.

B Additional Experiments

B.1 Case Study (Q4)

To empirically validate the contribution of each component in FaithfulRAG, we conduct a granular case study using a representative instance from the FaithEval dataset, with Llama3.1-8B-Instruct as the backbone model. The intermediate outputs at each stage are detailed in Table 4, illustrating how our framework progressively resolves knowledge conflicts while maintaining logical coherence.

Step 1: Self-Knowledge Extraction. In this phase, the model extracts abstract parametric knowledge relevant to the query.

Step 2: Self-Context Generation. The model synthesizes its self-knowledge into a complete context that reflects the LLM's "self-perception" about the query-related facts. This bridges abstract knowledge with task-specific reasoning, enabling downstream conflict resolution.

Step 3: Self-Fact Extraction. The model distills self-context into fine-grained self-facts. These self-facts serve as anchors for aligning with the retrieved context while preserving logical constraints.

Step 4: Contextual Alignment The model first converts self-facts and segmented retrieved context into embeddings, then computes the similarity between these embeddings. Finally, it selects the top-k chunks to form a self-aligned context that closely aligns with self-facts.

Step 5: Self-Think. Finally, the model reconciles self-aligned context with the whole context through iterative reasoning, including (i) Flag the contradiction between guidelines and the prescription. (ii) Probes for missing data in the context. (iii) Verify specialist approval and monitor renal function.

Table 6: The MR metric results of our method and the SOTA baseline across four datasets. Lower MR values indicate higher context faithfulness.

Model	Backbone LLM	MR MuSiQue	MR SQuAD	ACC MuSiQue	ACC SQuAD
Group 1: With Full Context					
Origin	llama3.1-8b-instruct	17.6	20.7	67.8	69.5
	qwen2.5-7b-instruct	17.3	27.6	75.2	68.3
	mistral-7b-instruct	15.9	25	67.6	67.2
Group 2: Specific RAG Models					
Self-RAG	Llama2-7B	18.8	30.1	54.1	62.0
ChatQA-1.5	Llama3.1-8B	17.7	19.1	75.0	77.0
ChatQA-2.0	Llama3.1-8B	15.5	20.8	77.2	75.4
Group 3: Context-faithful Prompting					
Opin(Instr)	llama3.1-8b-instruct	11.3	13.8	70.3	73.4
	qwen2.5-7b-instruct	15.3	25.5	76.9	70.5
	mistral-7b-instruct	13.1	22.9	68.1	69.3
ATTR	llama3.1-8b-instruct	14.7	16.8	62.8	69.5
	qwen2.5-7b-instruct	14	22.2	78.7	72.9
	mistral-7b-instruct	12.7	20.6	66.1	70.2
KRE	llama3.1-8b-instruct	14.0	14.5	35.9*	66.1
	qwen2.5-7b-instruct	18.6	22.7	70.7	73.7
	mistral-7b-instruct	16.3	16.8	50.6	74.6
Group 4: Context-faithful Decoding					
CAD	llama3.1-8b-instruct	12.8	14.6	72.6	71.2
	qwen2.5-7b-instruct	12.4	21.7	78.6	73.4
	mistral-7b-instruct	11.8	18	63.6	66.9
COIECD	llama3.1-8b-instruct	9.2	10.7	70.5	71.8
	qwen2.5-7b-instruct	9.8	15.5	69.7	70.8
	mistral-7b-instruct	10.3	16.1	66.8	65.4
Ours	llama3.1-8b-instruct	11.8	13.5	**79.9**	**86.3**
	qwen2.5-7b-instruct	13.8	15.7	78.0	78.3
	mistral-7b-instruct	**7.7**	**9.7**	78.5	85.7

Table 7: Model Analysis. The comparison of performance between our model and SOTA baselines by integrating different sizes of LLMs.

Model	Backbone LLM	FaithEval	RealtimeQA	MuSiQue	SQuAD
Group 1: Without Context					
Origin	llama3.2-3b-instruct	15.3	41.6	7.6	7.7
	llama2-13b-instruct	17.5	15.9	18.7	14.2
	deepseek-moe-16b	14.1	8.9	7.3	10.7
Group 2: With Full Context					
Origin	llama3.2-3b-instruct	67.1	80.5	66.1	79.2
	llama2-13b-instruct	75.6	55.7	80.5	78.4
	deepseek-moe-16b	60.3	51.3	67.4	73.9
Group 3: Context-faithful Prompting					
Opin(Instr)	llama3.2-3b-instruct	65.7	**84.9**	69.2	75.2
	llama2-13b-instruct	67.9	60.2	81.5	79.1
	deepseek-moe-16b	62.8	60.2	72	72
ATTR	llama3.2-3b-instruct	70.6	77.9	66.6	79.2
	llama2-13b-instruct	76.2	61.1	81.5	78.9
	deepseek-moe-16b	59.2	53.1	67.1	73.1
Group 4: Context-faithful Decoding					
CAD	llama3.2-3b-instruct	60.9	65.5	76.3	76.8
	llama2-13b-instruct	78.1	51.3	80.4	75.6
	deepseek-moe-16b	60.2	49.6	52.8	72.9
COIECD	llama3.2-3b-instruct	59.5	63.7	70.6	66.4
	llama2-13b-instruct	78.1	50.4	80.0	75.4
	deepseek-moe-16b	68.9	51.3	69.2	76.0
Ours	llama3.2-3b-instruct	70.1	79.6	78.4	79.8
	llama2-13b-instruct	**80.0**	53.1	**83.9**	**86.3**
	deepseek-moe-16b	79.2	61.9	76.1	82.1

B.2 Error Analysis

Current faithful methods enhance faithfulness at the expense of an increased risk of context misinterpretation. In this section, we systematically analyze how effective FaithfulRAG is in alleviating different types of errors. Specifically, we consider **3** different types of errors, where Case 1 and Case 2 are over-confidence and incorrect-match errors, respectively, consistent with Section 3, and Case 3 represents all other scenarios beyond these two cases. As shown in Table 5 and Figure 5, we have the following findings. FaithfulRAG achieves balanced mitigation of both Case 1 and Case 2 errors. Specifically, it reduces Case 1 errors by 6.8% while simultaneously decreasing Case 2 errors by 1.6% on average. This improvement stems from our well-designed framework, which enables LLMs to dynamically reconcile parametric knowledge with contextual evidence. By isolating discrepancies at the fact level and applying a self-think module, FaithfulRAG preserves high-quality parametric knowledge while systematically rejecting contexts that introduce logical inconsistencies or semantic divergence. FaithfulRAG aslo demonstrates superior performance in Case 3, maintaining the lowest error rate (5.7%) compared to baselines, demonstrating its robustness in handling edge cases.

B.3 M_R Result

Following previous work (Longpre et al., 2021), we calculate the context faithfulness metric (M_R) on entity-level knowledge conflict datasets, MuSiQue and SQuAD. The result is shown in Table 6. Among all methods, our approach achieves the best performance on Mistral, with 7.7% on MuSiQue and 9.7% on SQuAD. COIECD performs best across different backbone models, as its modified decoding strategy enforces stronger context alignment, leading to higher context faithfulness. However, our method outperforms COIECD by up to 10% in ACC, demonstrating that our approach not only improves context faithfulness but also maintains strong downstream task performance (ACC).

B.4 Model Analysis

To comprehensively demonstrate the generalizability of our method, we conduct experiments on models of different sizes and architectures with the results shown in Table 7. Our method significantly improves performance across models of different scales, from smaller 3B models to larger 16B models. This demonstrates the strong generalizability and adaptability of our approach, making it effective across a wide range of LLM architectures. Note that the KRE method shows a high rate of refusal responses. Therefore, we do not include it as a baseline in our additional experiments.

Table 8: Performance Comparison of Different Embedding Models Using Llama3.1-8B-Instruct as the Backbone Model.

Embedding Model	Params	Faitheval	RealtimeQA	MuSiQue	SQuAD
all-MiniLM-L6-v2	22.7M	79.8	81.4	79.9	86.3
TinyBERT-L6-v2	67M	78.3	84.1	80.1	85.1
contriever-sentencetransformer	110M	76.9	83.2	79.9	86.3
sentence-t5-base	110M	78.3	80.5	80.0	86.2

B.5 Impact of Different Embedding Models

In the main experimental setup, we employ all-MiniLM-L6-v2[1] to embed self-facts. To further assess the effect of different embedding models, we conducted additional experiments by replacing it with alternative models. As shown in Table 8, all-MiniLM-L6-v2 achieves the best performance on FaithEval and SQuAD, while TinyBERT-L6-v2 performs optimally on RealtimeQA and MuSiQue. We selected all-MiniLM-L6-v2 as the default embedding model in our main setup due to its smaller parameter size and balanced overall performance across benchmarks.

C Implementation Details

C.1 Benchmark Dataset

We evaluate FaithfulRAG on four benchmark datasets, including FaithEval (Ming et al., 2025), RealtimeQA (Kasai et al., 2024), and MuSiQue (Trivedi et al., 2022), SQuAD (Rajpurkar et al., 2016) from KRE(knowledge robustness evaluation) (Ying et al., 2024).

FaithEval (Ming et al., 2025): A novel benchmark dataset designed to evaluate the faithfulness of LLM and RAG systems across various contextual scenarios. This dataset consists of 4,900 high-quality questions covering three task types: Unanswerable, Inconsistent, and Counterfactual contexts. The Counterfactual Context is constructed based on ARC-Challenge, a multiple-choice science QA dataset at the elementary school level. Its knowledge conflict extends beyond the entity level, involving more complex logical relationships. We selected the Counterfactual subset as it aligns closely with our motivation.

RealtimeQA (Kasai et al., 2024): A dynamic question-answering dataset designed to evaluate the ability of QA systems to handle real-time information, challenging the assumptions of traditional static QA datasets and targeting more immediate application scenarios. To test the model performance in extreme cases where some contexts are

irrelevant to the question, we follow the (Zhou et al., 2023) and construct the RealTime QA-22 dataset. This dataset selects six questions from the first week of 2022 as the test set, with an equal proportion of answerable and unanswerable questions.

MuSiQue (Trivedi et al., 2022) and SQuAD (Rajpurkar et al., 2016): These two datasets are from the previous research (Ying et al., 2024) designed to study the behavior of LLMs when confronted with contexts that conflict with their internal memory. It encompasses tasks involving both factual knowledge and commonsense reasoning and is constructed by modifying existing machine reading comprehension and commonsense reasoning datasets. Each entry in the dataset contains a negative context with knowledge conflicts and an unmodified golden context. In the main experiment, we use the negative-context as the context to construct MuSiQue and SQuAD. While when evaluating performance in standard scenarios, the golden-context is used as the context.[2] To better align with real-world RAG scenarios, we use the subset with longer context lengths.

C.2 Baseline Selection

We carefully select baselines from four categories for a comprehensive evaluation.

Oringin Model: This model serves as a standard baseline in existing RAG systems. It employs an instruction-following model, such as llama3.1-8b-instruct, as its base model. Given a question and its related context, the model generates accurate and contextually appropriate answers

Specific RAG Systems: This group of methods (Asai et al., 2023; Xu et al., 2024a; Liu et al., 2024d) provides solutions for various RAG application scenarios, aiming to enhance the general ability of RAG systems across multiple benchmarks. By carefully selecting instruction-following datasets and designing tailored Supervised Fine-tuning (SFT) strategies, they improve the performance of LLMs in retrieval-augmented tasks.

Context-faithful Prompting: This category of methods (Zhou et al., 2023; Ying et al., 2024) enhance context faithfulness in LLMs by designing specialized prompting strategies. It addresses knowledge conflict issues by strengthening the reliance of LLMs on the provided context while reducing their dependence on parametric knowledge.

---

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] Unless otherwise specified, MuSiQue and SQuAD refer to the corresponding datasets using negative context.

Figure 6: Prompts for self-knowledge extraction, self-context generation, self-fact extraction, and the self-think.

Context-faithful Decoding: These models ([Shi et al., 2023](); [Yuan et al., 2024]()) modify the original inference strategy during model inference. Techniques such as contrastive decoding and constrained decoding are employed to guide the model to focus more on the given context rather than relying on parametric knowledge, thereby enhancing the model's context faithfulness.

C.3 Evaluation Metrics and Implementation

Our primary evaluation metric across all tasks is accuracy (ACC). We first normalize predictions and answers by removing stop words and punctuation, then determine whether the prediction and answer are identical. For the Memorization Ratio (M_R), following the approach ([Longpre et al., 2021]()), we first compute the Exact Match (EM) between predictions and original answers, denoted as p_o. Then, we compute the EM between predictions and substituted answers, denoted as p_s. Finally, we use the formula $M_R = \frac{p_o}{p_o + p_s}$.

During inference (except for context-faithful decoding strategies), we deploy vLLM ([Kwon et al., 2023]()), a high-performance LLM designed to accelerate LLM inference. We standardized the sampling arguments across all methods and set the temperature to 0 to ensure reproducibility of results. To help the model better understand the task's patterns and requirements, all the baselines in our comparison adopt a few-shot format.

In the Contextual Knowledge Alignment module, we use the conventional Fixed-size Chunking Strategy to chunk the original context. Specifically, we divided the text into multiple segments based on the predefined fixed size, which facilitates the segmentation of the context for further processing. We set the default chunk size to 20. When selecting the top-k self-aligned context, we set K=5. In our paper, we did not report parameter analysis on K and chunk size, as our model is not particularly sensitive to this hyperparameter, and the default settings already demonstrate the effectiveness of our approach.

C.4 Implementation Details of Ablation Study

To verify the effect of the Self-Think module, we conducted three separate ablation experiments. First, we conduct an ablation study on the overall module. Specifically, we directly prepend the self-aligned context to the original context to form a new combined context and instruct the LLM to answer based on it. This approach primarily lever-

ages the Position Bias of the LLM's internal attention, which prioritizes earlier positions in the context ([Hsieh et al., 2024](); [Liu et al., 2024c]()), thereby implicitly emphasizing the self-aligned context. The prompt is as follows:

Variant 1: w/o whole Module

Context: {self-aligned context}
{origin context}
Question: {question}
Answer:

Next, we conduct an ablation study on Think stage. Instead of the original Think stage, we introduce a Special Annotation approach, where we use special annotation to highlight the self-aligned context within the original context and employ instruction-based prompting to explicitly guide the LLM to focus on the self-aligned context enclosed by these markers. This modification prevents the LLM from actively thinking through and understanding the self-aligned facts, thereby hindering their effective fusion with the original context. The prompt is as follows:

Variant 2: w/o Think

- Instructions
1. Analyze the **Context** and identify the sentences wrapped in '[important chunk: xxx]'. These sentences contain key information.
2. Focus on the **important chunks** to extract the most relevant facts related to the **Question**.
3. If the facts from the important chunks are not sufficient to answer the question, refer to the full **Context** for additional information.
4. Please use the format of: Reason: (reason) Answer:(answer)
Context: {context}
Question: {question}
Answer:

Finally, we conduct an ablation study on Reasoning stage. We substituted the structured reasoning with a naive Chain-of-Thought (CoT) approach, which does not provide explicit structured guidance for the LLM's reasoning. This modification makes the model rely solely on its implicit inference capabilities. The prompt is as follows:

D   Prompt Design

The prompt templates applied in FaithfulRAG for self-knowledge extraction, self-context generation, self-fact extraction, and self-think are shown in the figure 6.