Beyond Black-Box Interventions: Latent Probing for Faithful Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems often fail to maintain contextual faithfulness, generating responses that conflict with the provided context or fail to fully leverage the provided evidence. Existing methods attempt to improve faithfulness through external interventions, such as specialized prompting, decoding-based calibration, or preference optimization. However, since these approaches treat the LLM as a black box, they lack a reliable mechanism to assess when and why knowledge conflicts occur. Consequently, they tend to be brittle, data-intensive, and agnostic to the model's internal reasoning process. In this paper, we move beyond black-box interventions to analyze the model's internal reasoning process. We discover that conflicting and aligned knowledge states are linearly separable in the model's latent space, and contextual noise systematically increases the entropy of these representations. Based on these findings, we propose ProbeRAG, a novel framework for faithful RAG that operates in three stages: (i) fine-grained knowledge pruning to filter irrelevant context, (ii) latent conflict probing to identify hard conflicts in the model's latent space, and (iii) conflict-aware attention to modulate attention heads toward faithful context integration. Extensive experiments demonstrate that ProbeRAG substantially improves both accuracy and contextual faithfulness. The related resources are available at https://anonymous.4open.science/r/ProbeRAG-CF6B.

1 Introduction

Retrieval-Augmented Generation (RAG) has rapidly evolved as a powerful paradigm to enhance Large Language Models (LLMs) with external, up-to-date knowledge, effectively mitigating the problem of outdated or hallucinated knowledge (Guu et al., 2020a; Feng et al., 2024; Zhang et al., 2025a). Despite its promise, existing RAG systems often struggle with a critical challenge of contextual faithfulness in practice (Bi et al., 2025a,b), generating responses that are inconsistent with the retrieved context or fail to fully leverage the provided external knowledge (Xu et al., 2024a; Zhang et al., 2025b). When faced with noisy or conflicting context, the LLM's knowledge integration process fails, causing the model to default to its parametric knowledge or produce incoherent hybrids. As shown in Figure 1, even Qwen3-8B suffers a significant drop in accuracy, from 71% to 59% when the context contains noise, and further to 56% when confronted with conflicting knowledge.

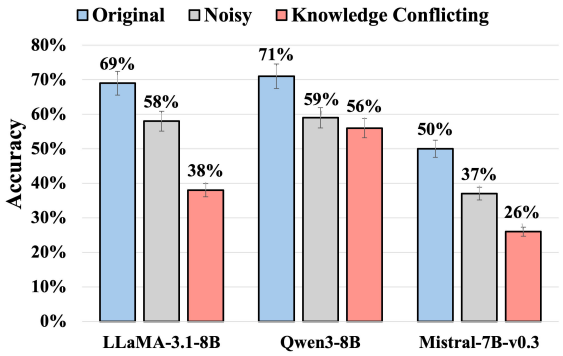


Figure 1: Experiments reveal that RAG systems degrade the performance when (i) exposed to contextual noise or (ii) confronted with conflicting knowledge.

Existing methods to improve faithfulness primarily rely on external interventions, which can be classified into three categories. (i) Prompt-based methods, which design specialized instructions to guide the model's reasoning process (Zhou et al., 2023a; Asai et al., 2024; Ying et al., 2024; Zhang et al., 2025b). While this strategy can indeed improve factual grounding, its performance is often highly sensitive to the prompt and can not generalize across different domains or tasks. (ii) Decoding-based models that improve RAG faithfulness by adjusting the decoding process, calibrating the logits with contextual scoring (Shi et al., 2023a; Yuan

et al., 2024). However, these methods are often tightly coupled with specific decoding strategies and become brittle when the retrieved context contains noise or contradictions. (iii) Preference optimization methods fine-tune the LLM using Direct Preference Optimization (DPO) to encourage faithful grounding (Si et al., 2025; Bi et al., 2025a). Although enabling end-to-end learning, these methods depend heavily on carefully designed reward functions and large-scale, high-quality preference data, which are costly to collect.

More crucially, existing methods share a fundamental limitation: they treat the LLM as a black box, applying external interventions without understanding the internal mechanisms during the reasoning process. Consequently, their interventions remain correlational rather than causal: they may statistically associate certain inputs with more faithful outputs, but cannot diagnose why the model fails in specific conflict instances, nor predict its behavior under novel forms of contradiction. To achieve robust faithfulness, we argue it is necessary to move beyond external corrections and focus on two critical questions: how is the conflict represented within the model's latent space, and how does the latent conflict disrupt faithful generation?

In this work, we conduct a latent space analysis to uncover how LLMs internally integrate contextual knowledge with their parametric memory and how they represent conflicting knowledge within their latent space. Specifically, we uncover two critical findings: (i) internal hidden-state of conflicting and aligned knowledge are linearly separable in the model's latent space, providing a latent feature for conflict detection; and (ii) contextual noise systematically increases the entropy of these representations, which clarifies why noisy contexts always obscure the latent conflict feature.

Motivated by these findings, we introduce ProbeRAG, a framework that leverages latent probing to detect and mitigate conflicts for faithful RAG. It consists of three steps: (i) fine-grained knowledge pruning to reduce noise by filtering irrelevant context, (ii) latent conflict probing, where a lightweight probe is used to model the mapping relationship between hidden states and conflicting/aligned knowledge, and (iii) conflict-aware attention to modulate attention heads based on probe outputs to ensure faithful grounding. By integrating latent-space probing with targeted intervention, ProbeRAG shifts the paradigm from external constraint to internal guidance, enabling more robust, generalizable faithfulness. Our contributions are summarized as follows:

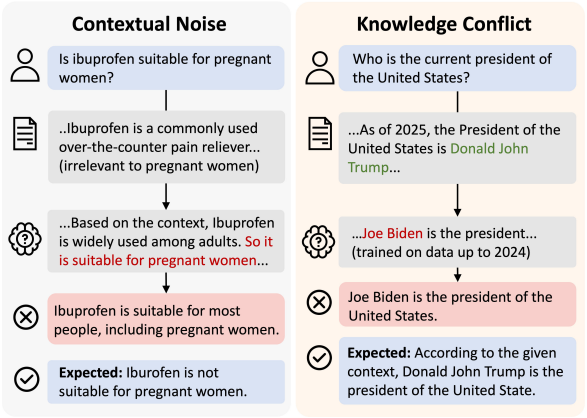


Figure 2: Contextual noise draws the model's attention and enhances its inherent knowledge (in the left case). And the model prefers to follow inherent knowledge in the conflicting scenario (in the right case).

- We perform an in-depth analysis on the models' internal knowledge integration mechanism. We discover the latent conflict feature existed in this process and the negative obscurity of contextual noise.
- We propose ProbeRAG, a framework designed to enhance RAG faithfulness, which detects latent conflicting knowledge and further guides the model to pay more attention while integrating the context.
- We evaluate the effectiveness of our framework on multiple benchmarks, demonstrating that ProbeRAG consistently outperforms all other competitive baselines.

2 Preliminary Study

2.1 Existing Challenges in RAG Faithfulness

Based on the consensus reached through existing works (Ji et al., 2023; Bi et al., 2025a; Zhang et al., 2025b; Yuan et al., 2025), there are two key factors that underlie this issue: (i) Knowledge conflict between the contextual and internal knowledge of the model. Since the models tend to prioritize their parametric memory over the external evidence. (ii) Contextual noise with weak relevance to the task. As the model's attention will be drawn to this noise. To assess these two factors, we designed two controlled scenarios. In the *knowledge conflicting* scenario, we select several key entities in the context and replace them with other entities of the same

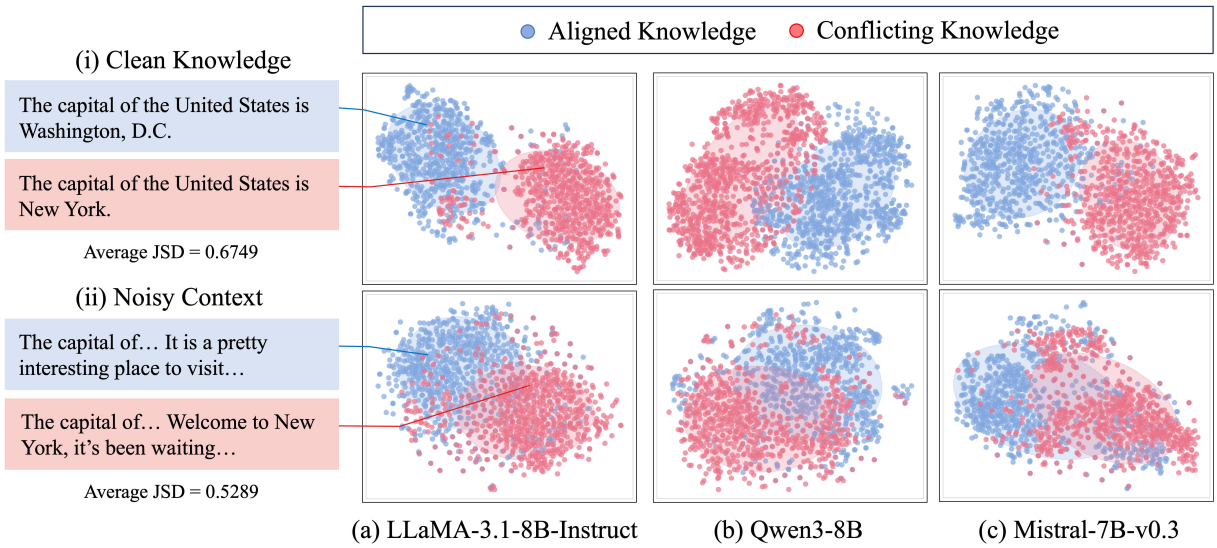


Figure 3: T-SNE visualization of hidden-state patterns between aligned and conflicting knowledge. There is a clear distinction in the distribution of hidden states between aligned and conflicting knowledge. Additionally, in the noisy scenario, the distinct pattern becomes unclear, emphasizing the importance of noise filtering.

type to construct counterfactual knowledge. By doing these, we introduce inconsistencies between the context and model knowledge to construct knowledge conflicts. In the *noisy* scenario, the original context is augmented with passages that are semantically aligned with the query but topically irrelevant, introducing unrelated knowledge.

As shown in Figure 1, we conduct evaluations on three different open-source models, and all their accuracy decreases significantly in both scenarios. Some typical error cases are presented in Figure 2. In the contextual noise scenario, their accuracy drops by more than 10%. As demonstrated in the case studies, the primary issue is that irrelevant context amplifies non-task-related knowledge within the model, allowing it to gain undue priority during generation. In the knowledge conflict scenario, the model performance decrease is more pronounced, particularly for LLaMA-3.1-8B-Instruct and Mistral-7B-v0.3. Case study indicates that the model assigns a lower priority to contexts involving conflicts within its internal knowledge.

2.2 Latent Space Probing and Analysis

To further explore how models integrate external knowledge, we analyze the knowledge representations in their latent space. Following the method of (Xie et al., 2024), we extract the model's parametric knowledge K_a for a given question, and use an external LLM to construct corresponding conflicting knowledge K_c. Totally, we construct 700 such samples and conduct analysis using different model architectures. We input each knowledge pair $\langle K_a, K_c \rangle$ into the model separately, extract the hidden states, and perform a two-dimensional visualization using t-SNE (van der Maaten and Hinton, 2008) (the embedding space dimension is set to 2, fit for 3000 rounds with perplexity set as 30). The internal representations are visualized in Figure 3 (red and blue points). From the figure, we observe that aligned knowledge and conflicting knowledge form two clearly separable distributions, with the average JS Divergence being 0.6749. This observation indicates that a potential conflict feature exists in the model's latent space, which forms two linearly separable clusters of hidden-states for aligned and conflicting knowledge. This demonstrates that conflicting knowledge generates internal biases during the knowledge integration process, explaining why it is easily overlooked by models while consistent knowledge is more preferred.

In addition, we analyze the hidden-state distribution when the model is exposed to contextual noise. The results show that the noise partially disrupts the intrinsic distributional structure and increases the entropy of the knowledge distribution. Notably, when adding contextual noise, the average JS Divergence decreases by approximately 0.15, clarifying why noisy contexts obscure the latent conflict feature. This demonstrates that in real-world scenarios, latent conflict features are often difficult to capture because contextual noise renders the two clusters indistinguishable. Therefore, the inherent latent conflict feature only supports fine-grained knowl-

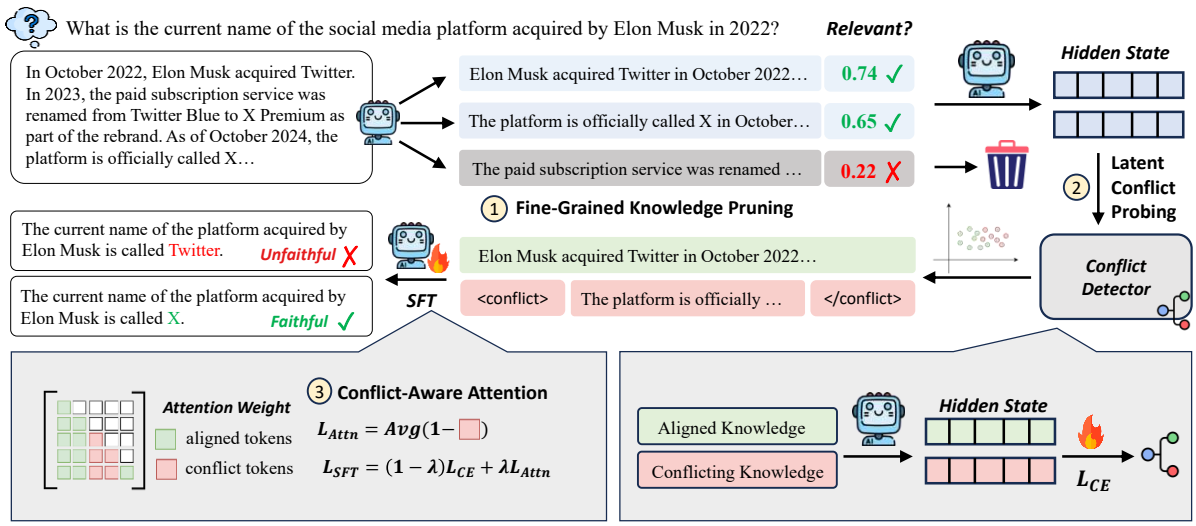


Figure 4: The overview of the framework ProbeRAG, which comprises three steps: (1) **Fine-Grained Knowledge Pruning**: decompose the context and filters out irrelevant knowledge; (2) **Latent Conflict Probing**: detect knowledge conflict via a probe trained in the model's latent space; (3) **Conflict-Aware Attention**: regulate the model's attention distribution on conflicting knowledge via fine-tuning.

edge as the unit of judgment, while coarser-grained contexts are not suitable for direct assessment.

3 Method

In this section, we introduce the framework ProbeRAG, which focus on the internal guidance instead of external interventions. As illustrated in Figure 4, it comprises three steps: (1) **Fine-Grained Knowledge Pruning**: where the context is decomposed into fine-grained knowledge and irrelevant knowledge is filtered out; (2) **Latent Conflict Probing**: we detect knowledge conflict via a probe trained in the model's latent space. It takes the model's hidden state as input and predicts whether the contextual knowledge is in conflict with the model's parametric knowledge; (3) **Conflict-Aware Attention**: we fine-tune the model to regulate its attention head towards faithful context integration via an auxiliary attention-guidance loss item. the following subsections will provide detailed illustrations for each step.

3.1 Fine-Grained Knowledge Pruning

Since contextual noise renders the two types of knowledge indistinguishable within the latent space, the most appropriate unit for processing is the fine-grained knowledge. The knowledge corresponds to an independent, complete sentence-level statement that cannot be further decomposed. Each statement preserves the subject–predicate–object structure with necessary modifiers, ensuring no information is lost during decomposition. To extract knowledge $\{K_1, K_2, \ldots, K_n\}$ from a given context D, we leverage the decomposition capabilities of an external LLM (we choose GPT-4o (OpenAI, 2024) for its strong reasoning and text-processing abilities). We also provide an ablation study (Section 4.3) to utilize an open-source model (LLaMA-3-8B-Instruct) in this step. The decomposition result of the context is a list of knowledge. Formally, we define this process as:

$$\text{Decompose}(D) = \{K_1, K_2, \ldots, K_n\}$$

where K_i denotes the i-th knowledge statement. Detailed prompt is provided in Appendix C.

We further filter irrelevant knowledge to reduce contextual noise. For each knowledge statement K_i, we compute its similarity with the query Q:

$$f(Q, K_i) = \langle q, k_i \rangle$$

where $q = Enc(Q)$ and $k_i = Enc(K_i)$ are vector embeddings of the query and the knowledge item, respectively, and $\langle \cdot, \cdot \rangle$ denotes cosine similarity. We employ the all-MiniLM-L6-v2[1] for embedding. Finally, the top-k results are selected.

3.2 Latent Conflict Probing

To effectively detect knowledge conflict, we utilize the latent conflict feature in the model as we observed in our preliminary study in Section 2.

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

Specifically, when aligned and conflicting knowledge are input into the model, the model's hidden state projects them into two separable clusters. Leveraging this property, we implement a reverse prediction through a probe: by inputting the model's hidden state, it determines whether the knowledge is aligned or conflicting. To enable the probe to fit the reverse process of the model's hidden state mapping, the probe model must be trained in the model's latent space. We leverage the knowledge editing dataset MQuAKE (Zhong et al., 2023), since it contains both aligned knowledge prior to editing and conflicting knowledge after editing, providing natural pairs of aligned and conflicting knowledge $\langle K_a, K_c \rangle$. Importantly, the data format and textual granularity in MQuAKE align closely with the knowledge statements extracted in our framework, making it suitable for supervision.

During inference, the resulting filtered knowledge from previous step is passed through the frozen model to obtain its hidden state representation, which is subsequently classified by the probe:

$$\mathcal{M}(K_i) \in \mathbb{R}^{d_M}, \quad \mathcal{P}\big(\mathcal{M}(K_i)\big) \in \{0, 1\},$$

where $\mathcal{M}(K_i)$ denotes the hidden state of knowledge statement K_i produced by frozen model $\mathcal{M}$ with dimension d_M and $\mathcal{P}$ denotes the probe.

We validated the accuracy and generalization of using probe models for conflict detection in Section 4.4. Despite being trained on knowledge editing datasets, the probes maintained strong generalization capabilities on RAG domain data, ensuring the reliability of conflict detection. To further enhance the priority of conflicting knowledge in the model generation process, we mark the conflicting knowledge with special tokens, i.e., wrapping them within $\langle conflict \rangle$ and $\langle /conflict \rangle$. This explicit annotation enables the subsequent stage to be aware of which knowledge statement are conflicting and should be augmented.

3.3 Conflict-Aware Attention

In this step, we aim to encourage the model to pay more attention to conflicting knowledge. External interventions like in-context learning can hardly replace the model's inherent knowledge, which is demonstrated in our ablation study in Section 4.3. Therefore, we attempt to leverage the attention mechanism within the model to enhance the contextual faithfulness and propose Conflict-Aware Attention. We introduce an additional attention-guidance loss term that explicitly regularizes the model's attention distribution. Specifically, for each conflicting knowledge item K_i, we denote its token sequence as $T^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \ldots, t_m^{(i)}\}$. The positions of these tokens in the input context are represented by $S = \{j \mid \exists \mathcal{P}(\mathcal{M}(K_i)) = 1, x_j \in T^{(i)}\}$, where $\mathcal{P}(\mathcal{M}(K_i)) = 1$ indicates that knowledge item K_i is judged as conflicting by the probe, and x_j denotes the j-th token of the context. In practice, these positions in S can be directly identified via the previously introduced special tokens $\langle conflict \rangle$ and $\langle /conflict \rangle$.

Based on this alignment, we extract the attention weights from subsequent tokens attending to the conflict-related tokens and then compute the attention guidance loss as follows:

$$\mathcal{L}_{\text{Attn}} = \frac{1}{|P|} \sum_{(i,j) \in P} (1 - \alpha_{ij}), (i, j) \in P$$

$$P = \{(i, j) \mid i \geq j;\ j \in S\}$$

where α_{ij} denotes the attention weight of token i on token j. Finally, we combine the attention loss with the standard language modeling objective through a weighted sum:

$$\mathcal{L}_{\text{SFT}} = (1 - \lambda)\mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Attn}},$$

where $\lambda \in [0, 1]$ balances the trade-off to avoid overfitting to the conflicting knowledge. We provide an analysis on its impact in Section 4.5. This joint objective ensures that the model not only learns to generate faithful outputs but also explicitly attends to conflicting knowledge during training.

4 Experiment

4.1 Setup

Datasets. We evaluate ProbeRAG on three datasets. ConFiQA (Bi et al., 2025a) is a benchmark designed to assess contextual faithfulness in question answering, particularly under real-world RAG scenarios involving knowledge conflicts. It consists of three subsets: QA (Question Answering), MR (Multi-hop Reasoning), and MC (Multi-Conflicts). QA is a single-hop question answering task, while MR and MC are multi-hop reasoning tasks in which the context includes one and multiple counterfactuals. FaithEval (Ming et al., 2024) introduces conflicts at the level of logical reasoning: inconsistencies arise not from direct factual contradictions, but from reasoning chains that lead to conflicting conclusions. Finally, we also evaluate

Category	Method	FaithEval		ConFiQA (MC)		ConFiQA (MR)		ConFiQA (QA)		SQuAD	
		F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
LLaMA-3.1-8B-Instruct											
Baseline	No-Context	27.7	6.0	5.0	2.1	6.1	1.9	6.1	1.3	8.9	1.2
	Full-Context	66.9	53.1	28.0	22.5	50.3	41.3	58.5	49.0	64.5	46.0
Prompt-Based	Opin(Instr) (Zhou et al., 2023a)	34.9	15.1	67.4	57.3	65.9	54.0	76.9	67.4	66.0	47.7
	KRE (Ying et al., 2023)	59.1	12.1	68.2	59.8	68.7	58.9	84.0	74.7	59.8	43.7
	FaithfulRAG (Zhang et al., 2025b)	62.4	49.1	70.3	61.2	62.4	53.1	70.4	65.2	61.3	50.2
Decoding-Based	CAD (Shi et al., 2023a)	59.4	42.7	16.0	11.4	40.0	31.3	48.3	38.1	60.3	41.8
	COIECD (Yuan et al., 2024)	56.1	41.3	28.5	24.0	50.9	43.3	67.1	60.1	67.0	50.3
	AdaCAD (Wang et al., 2025b)	61.4	47.7	43.3	32.5	46.7	42.5	55.7	48.1	66.4	51.2
Training-Based	Context-DPO (Bi et al., 2025a)	67.2	53.7	76.9	67.7	78.5	66.9	83.7	76.7	64.4	45.8
	CANOE (Si et al., 2025)	71.6	56.3	80.9	74.2	80.2	72.6	82.3	77.7	65.4	49.7
	ParamMute (Huang et al., 2025b)	68.5	56.2	70.6	60.9	73.2	61.2	78.3	73.2	64.2	48.5
	ProbeRAG (ours)	**74.4**	**64.4**	**89.2**	**87.7**	**89.7**	**87.0**	**93.1**	**91.7**	**68.4**	**53.3**
Qwen3-8B											
Baseline	No-Context	22.8	4.1	7.6	3.6	8.0	2.8	7.8	1.4	6.7	0.4
	Full-Context	55.5	23.8	59.6	50.2	66.1	55.1	72.5	64.2	63.8	44.9
Prompt-Based	Opin(Instr) (Zhou et al., 2023a)	35.0	13.9	70.7	61.1	69.7	59.5	78.8	69.2	63.8	46.1
	KRE (Ying et al., 2023)	58.1	12.3	67.5	59.1	68.4	59.0	80.4	67.3	48.6	29.7
	FaithfulRAG (Zhang et al., 2025b)	70.2	58.3	72.3	60.1	62.3	54.2	75.3	65.3	66.2	50.2
Decoding-Based	CAD (Shi et al., 2023a)	57.0	28.7	57.7	48.3	64.8	53.3	71.0	62.0	63.6	44.5
	COIECD (Yuan et al., 2024)	66.6	56.4	66.7	60.8	71.5	63.8	78.5	73.6	69.7	55.2
	AdaCAD (Wang et al., 2025b)	67.3	58.2	64.2	58.7	72.5	64.3	76.9	68.3	72.6	67.4
Training-Based	Context-DPO (Bi et al., 2025a)	55.2	24.0	59.6	50.1	65.9	55.0	72.3	63.9	63.8	44.9
	CANOE (Si et al., 2025)	70.3	60.2	85.2	81.7	84.6	80.7	92.2	86.5	69.4	53.4
	ParamMute (Huang et al., 2025b)	67.8	58.2	86.3	80.2	83.2	79.4	90.5	83.7	69.2	52.1
	ProbeRAG (ours)	**74.9**	**61.6**	**90.7**	**89.7**	**91.3**	**89.0**	**95.7**	**94.3**	**71.5**	**55.7**
Mistral-7B-v0.3											
Baseline	No-Context	26.2	4.4	4.4	0.9	4.9	0.5	6.1	1.0	8.1	1.0
	Full-Context	68.8	37.7	25.6	12.5	37.8	21.5	58.5	44.0	56.4	37.5
Prompt-Based	Opin(Instr) (Zhou et al., 2023a)	35.7	14.1	58.8	44.1	57.8	52.5	76.4	65.5	58.1	37.4
	KRE (Ying et al., 2023)	64.8	36.5	58.7	45.0	60.9	45.3	84.5	72.8	52.6	33.9
	FaithfulRAG (Zhang et al., 2025b)	66.4	40.3	60.3	44.3	56.5	42.1	74.2	63.5	60.3	40.1
Decoding-Based	CAD (Shi et al., 2023a)	68.9	33.3	16.7	5.9	27.5	12.8	53.5	36.9	51.4	32.1
	COIECD (Yuan et al., 2024)	64.4	29.5	26.1	14.5	39.3	26.3	58.9	45.1	59.2	39.7
	AdaCAD (Wang et al., 2025b)	68.4	34.2	33.7	23.4	39.6	29.4	58.2	46,3	61.6	43.6
Training-Based	Context-DPO (Bi et al., 2025a)	64.9	31.8	44.8	28.3	50.9	31.9	66.4	52.7	56.6	37.6
	CANOE (Si et al., 2025)	64.1	44.9	87.2	85.7	84.7	81.9	92.5	90.7	57.8	42.5
	ParamMute (Huang et al., 2025b)	67.1	45.2	86.3	82.1	86.4	82.1	93.1	91.2	60.2	43.8
	ProbeRAG (ours)	**74.9**	**62.9**	**91.2**	**89.7**	**90.8**	**88.2**	**95.1**	**93.7**	**68.1**	**53.6**

Table 1: Performance comparison of methods grouped by Baseline, Prompt-Based, Decoding-Based, and Training-Based. From the table we can see that ProbeRAG consistently achieves the SOTA results.

on SQuAD (Rajpurkar et al., 2016), following the version curated in KRE (Ying et al., 2023), which also incorporates fact-level knowledge conflicts.

Models and Baselines. We adopt several mainstream open-source models, including Llama-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-v0.3. We compare ProbeRAG against representative baseline methods from three major categories in the field of contextual faithfulness: prompt-based, decoding-based, and training-based approaches. Among the prompt-based methods, we include Opin(Instr) (Zhou et al., 2023a), KRE (Ying et al., 2023), and FaithfulRAG (Zhang et al., 2025b). For decoding-based methods, we evaluate CAD (Shi et al., 2023a), COIECD (Yuan et al., 2024) and AdaCAD (Wang et al., 2025b). For training-based methods, we compare against Context-DPO (Bi et al., 2025a), CANOE (Si et al., 2025) and ParamMute (Huang et al., 2025b). Specifically, we partition the ConFiQA dataset into training and test sets. All baselines that require training are trained on the ConFiQA training set, and evaluation is consistently performed on the test set. Additional implementation details are provided in the Appendix C.

4.2 Main Results

As shown in Table 1, ProbeRAG consistently achieves SOTA performance across all datasets and models. On FaithEval and ConFiQA (MC, MR, QA), ProbeRAG demonstrates strong generalization ability to both factual and logical conflicts, while on SQuAD, it further shows ProbeRAG improvements in traditional settings. Moreover, the consistent gains under different models highlight the robustness and generalizability of ProbeRAG.

Models	Modules	Faitheval		ConFiQA (MC)		ConFiQA (MR)		ConFiQA (QA)		SQuAD	
		F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
LLaMA-3.1-8B-Instruct	ProbeRAG	**74.4**	**64.4**	**89.2**	**87.7**	89.7	87.0	**93.1**	**91.7**	**68.4**	53.3
	with Open Source KP	70.3	59.5	83.2	82.4	**90.3**	**87.2**	90.4	89.8	65.7	**54.2**
	w/o Knowledge Pruning	62.1	48.4	81.1	79.4	84.4	80.8	88.5	87.5	59.2	45.0
	w/o Latent Conflict Probing	61.7	47.6	81.4	79.3	83.9	79.9	87.6	86.4	58.1	44.1
	w/o Conflict-Aware Attention	61.5	50.9	83.8	80.4	85.0	81.0	87.5	86.4	58.2	40.2
Qwen3-8B	ProbeRAG	**74.9**	**61.6**	**90.7**	**89.7**	**91.3**	**89.0**	**95.7**	**94.3**	**71.5**	**55.7**
	with Open Source KP	72.4	60.2	88.3	88.3	88.3	85.4	93.2	91.6	69.4	52.8
	w/o Knowledge Pruning	62.6	50.9	86.1	85.3	86.7	85.2	88.8	87.8	66.3	51.3
	w/o Latent Conflict Probing	61.0	49.8	85.4	84.6	86.6	85.1	88.6	87.5	66.1	51.0
	w/o Conflict-Aware Attention	64.0	54.2	86.2	84.8	86.1	84.3	89.6	88.5	66.1	51.5
Mistral-7B-v0.3	ProbeRAG	**74.9**	**62.9**	**91.2**	**89.7**	**90.8**	**88.2**	**95.1**	**93.7**	**68.1**	53.6
	with Open Source KP	70.3	60.7	88.5	85.7	87.3	84.9	92.5	90.4	66.7	**54.3**
	w/o Knowledge Pruning	69.5	58.5	86.6	85.5	86.2	84.7	88.4	87.1	62.9	48.7
	w/o Latent Conflict Probing	68.4	56.4	85.2	84.1	84.4	82.9	87.4	86.2	61.8	47.6
	w/o Conflict-Aware Attention	69.3	57.6	88.8	86.1	86.3	81.8	81.4	77.4	59.7	49.8

Table 2: Ablation Result. The ablation of each module significantly impacts the results. The Latent Conflict Probing module has the most substantial influence on the entire framework.

Specifically, for model LLaMA-3.1-8B-Instruct, ProbeRAG achieves an F1 score of 74.4% and an EM score of 64.4% on FaithEval, outperforming the strongest baseline CANOE (71.6% F1 / 56.3% EM) by approximately +3% F1 and +8% EM. On ConFiQA sub-tasks, ProbeRAG improves over existing methods by 3%–10% across MC, MR, and QA, further confirming its robustness in handling conflict scenarios. Similarly, for Qwen3-8B, ProbeRAG attains 74.9% F1 and 61.6% EM on FaithEval, yielding substantial gains compared with prior methods, and reaches 90.7% F1 and 89.7% EM on the MC task. On Mistral-7B-v0.3, ProbeRAG achieves 74.9% F1 / 62.9% EM on FaithEval and great improvements across ConFiQA and SQuAD, surpassing the best training-based baselines by a clear margin. The consistent improvements across multiple datasets, conflict types, and backbone models underscore the effectiveness, robustness, and general applicability of ProbeRAG.

4.3 Ablation Study

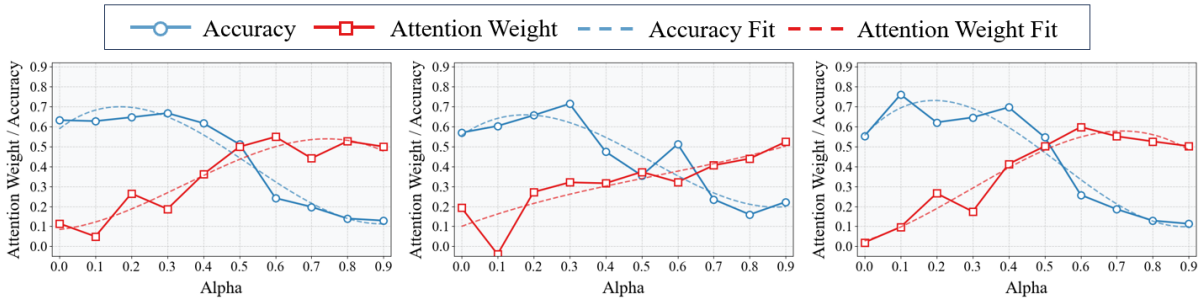
Ablation study results are summarized in Table 2. We first evaluate the effectiveness of performing Knowledge Pruning using an open-source LLM (with Open-Source KP). We substitute the GPT-4o with LLaMA-3-8B-Instruct, a lightweight, instruction-tuned open-source model. Results show that its ability to extract complete and semantically precise knowledge from the retrieved context is weaker than that of GPT-4o, which underscores the importance of context decomposition, as clean data facilitates subsequent process detection and enhancement. When the knowledge pruning step is removed, the model is forced to judge conflicts against every sentence in the context. Such coarse-grained filtering leads to incomplete contextual information and degrades the model's ability to resolve fine-grained conflicts, thereby diminishing contextual faithfulness. More critically, removing the conflict detection module results in the most significant performance drop. Without explicit conflict detection, the downstream training becomes ineffective, since there are no targets to pay attention to. Finally, removing Conflict-Aware Attention also results in substantial degradation. Even when conflicts are annotated, the model struggles to prioritize them during inference due to its inherent tendency to rely on its parametric knowledge.

4.4 Evaluation on Conflict Detector

Model	MQuAKE	FaithEval	ConFiQA	SQuAD
$\mathcal{P}_{Llama3.1}$	97.26	86.67	87.50	82.23
$\mathcal{P}_{Qwen2.5}$	96.53	82.82	84.33	79.57
$\mathcal{P}_{Mistral0.3}$	98.43	87.42	88.75	82.64

Table 3: Evaluation of conflict detectors on both knowledge editing and RAG benchmarks.

The performance of different conflict detectors is shown in Table 3. The detectors are trained on the training set of MQuAKE for its large scale, and we provide the evaluation results on the test set. For the general RAG dataset SQuAD, since no label is provided, we randomly select 200 samples and utilize LLM-Judge with GPT-4o. We require the LLM to tag the counterfactual knowledge as negative, others as positive, and calculate the accuracy. As we can see, the detector models can generalize well in the knowledge conflict scenario in RAG datasets, since the task of knowledge editing is



(a) LLaMA-3.1-8B-Instruct (b) Qwen3-8B (c) Mistral-7B-v0.3

Figure 5: Impact of α on accuracy (blue) and attention weight (red). Performance peaks at smaller α values (0.1 to 0.3) and then declines, indicating that overfitting to conflicting knowledge degrade the performance.

closely related to the knowledge conflict scenario.

4.5 Impact of Attention Guidance Loss Item

To further investigate the effect of the hyperparameter α, we conduct experiments with multiple values of α and analyze both the attention weights assigned to conflicting knowledge and the corresponding model performance. As shown in Figure 5, increasing α consistently raises the model's attention, with the growth curve gradually flattening and stabilizing around 0.5. However, model performance does not follow the same trend. Instead, performance peaks when α is in the range of 0.1 to 0.3, after which it declines as α continues to increase. This observation indicates that higher attention to conflicting knowledge does not necessarily lead to better performance. While attending to conflicting knowledge is crucial, the model must also balance its focus on the question itself and other relevant contextual information.

5 Related Work

Due to space limitations, we provide only a concise overview of the related work here. More detailed discussion can be found in Appendix F.

Retrieval-Augmented Generation (RAG) has emerged as a prominent paradigm for enhancing the factual accuracy and temporal relevance of Large Language Models (LLMs) by incorporating external knowledge sources (Shi et al., 2024; Xiang et al., 2025; Chen et al., 2025b; Xiao et al., 2025; Hui et al., 2025; Chen et al., 2025c). Early works such as REALM (Guu et al., 2020a) and RAG (Lewis et al., 2020) retrieve relevant passages from large corpora to assist generation. Subsequent research has explored improvements in both the retriever and generator modules, including dense retrieval techniques (Karpukhin et al., 2020; Izacard et al., 2023), adaptive retrieval strategies (Sun et al., 2022; Chen et al., 2025a), and hybrid models combining retrieval with parametric memory (Shi et al., 2023b; Wang et al., 2025e).

Contextual Faithfulness refers to the alignment between the generated output and the provided context, which is especially critical in RAG settings (Huang et al., 2025b; Bi et al., 2025c; Tang et al., 2025). Prompt-based methods design templates or self-reflection mechanisms to encourage faithful use of context (Asai et al., 2024; Ying et al., 2024; Liu et al., 2025). Decoding-based methods modify generation strategies to enhance the influence of the retrieved context (Yuan et al., 2024; Shi et al., 2023a; Santosh et al., 2025). Reinforcement learning frameworks such as CANOE (Si et al., 2025) and Context-DPO (Bi et al., 2025a) employ an end-to-end paradigm to optimize the generation process and reward contextual faithful response.

6 Conclusion

In this work, we focus on how LLMs internally integrate external knowledge with their parametric memory under knowledge conflicts. Through probing-based analysis, we uncovered two insights: conflicting and aligned knowledge states are linearly separable in the model's latent space, and contextual noise systematically increases the entropy of these representations. Building on these findings, we introduced a framework named ProbeRAG to improve RAG faithfulness, which combines fine-grained knowledge pruning, latent conflict probing, and conflict-aware attention to enhance contextual faithfulness. Comprehensive experiments across multiple benchmarks and large language models demonstrate that ProbeRAG consistently outperforms strong baselines, achieving SOTA performance. Our framework highlights the importance of explicitly mitigating knowledge conflicts, offering a principled direction for future research.

7 Limitations

While ProbeRAG demonstrates great improvements in textual RAG scenarios, its applicability to multimodal RAG systems remains limited. The current framework is designed around sentence-level textual decomposition and hidden-state probing, which are not directly transferable to modalities such as images, audio, or structured data. In multimodal contexts, knowledge conflicts may manifest in non-textual representations, requiring new strategies for knowledge decomposition, conflict detection, and attention guidance. Extending ProbeRAG to handle heterogeneous modalities would thus require substantial redesign of its probing mechanism and fine-tuning objectives, which we leave as an important direction for future research.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR 2024*.

Jun Bai, Minghao Tong, Yang Liu, Zixia Jia, and Zilong Zheng. 2025. Understanding and leveraging the expert specialization of context faithfulness in mixture-of-experts llms. In *Proceedings of EMNLP 2025*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2025a. Context-DPO: Aligning language models for context-faithfulness. In *Findings of ACL 2025*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2025b. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. In *ICLR 2025*.

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *ICML 2022*.

Jiajing Chen, Bingying Liu, Xiaoxuan Liao, Jia Gao, Hongye Zheng, and Yue Li. 2024a. Adaptive optimization for enhanced efficiency in large-scale language model training. In *ICFTIC 2024*.

Juan Chen, Baolong Bi, Wei Zhang, Jingyan Sui, Xiaofei Zhu, Yuanzhuo Wang, Lingrui Mei, and Shenghua Liu. 2025a. Rethinking all evidence: Enhancing trustworthy retrieval-augmented generation via conflict-driven summarization. *arXiv preprint arXiv:2507.01281*.

Shengyuan Chen, Chuang Zhou, Zheng Yuan, Qinggang Zhang, Zeyang Cui, Hao Chen, Yilin Xiao, Jiannong Cao, and Xiao Huang. 2025b. You don't need pre-built graphs for rag: Retrieval augmented generation with adaptive reasoning structures. *arXiv preprint arXiv:2508.06105*.

Tailun Chen, Yu He, Yan Wang, Shuo Shao, Haolun Zheng, Zhihao Liu, Jinfeng Li, Yuefeng Chen, Zhixuan Chu, and Zhan Qin. 2025c. Mirage: Misleading retrieval-augmented generation via black-box and query-agnostic poisoning attacks. *arXiv preprint arXiv:2512.08289*.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025d. Improving retrieval-augmented generation through multi-agent reinforcement learning. In *NeuIPS 2025*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024b. Knowledge Localization: Mission not accomplished? enter query localization! *arXiv preprint arXiv:2405.14117*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. FacTool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of EMNLP 2025*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of ACL 2024*.

Song Duong, Florian Le Bronnec, Alexandre Allauzen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2025. Scope: A self-supervised framework for improving faithfulness in conditional text generation. *arXiv preprint arXiv:2502.13674*.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of ACL 2019*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024*.

Kaiyuan Gao, Sunan He, Zhenyu He, Jiacheng Lin, QiZhi Pei, Jie Shao, and Wei Zhang. 2023. Examining user-friendly and open-sourced large gpt models: A survey on language, multimodal, and scientific gpt models. *arXiv preprint arXiv:2308.14149*.

Sachin Goyal, Christina Baek, J Zico Kolter, and Aditi Raghunathan. 2024. Context-parametric inversion: Why instruction finetuning can worsen context reliance. *arXiv preprint arXiv:2410.10796*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, and 1 others. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020a. REALM: Retrieval-augmented language model pre-training. In *PMLR 2020*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020b. Retrieval augmented language model pre-training. In *ICML 2020*.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, and 1 others. 2025a. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. *arXiv preprint arXiv:2501.13573*.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. 2025b. ParamMute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. In *NeurIPS 2025*.

Huangyw Huangyw, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. 2025. Dynamic attention-guided context decoding for mitigating context faithfulness hallucinations in large language models. In *Findings of ACL 2025*.

Yulong Hui, Chao Chen, Zhihang Fu, Yihao Liu, Jieping Ye, and Huanchen Zhang. 2025. Interact-rag: Reason and interact with the corpus, beyond black-box retrieval. *arXiv preprint arXiv:2510.27566*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. In *JMLR 2023*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*.

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

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

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. Cocoa: Confidence-and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of EMNLP 2025*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS 2020*.

Kun Li, Tianhua Zhang, Yunxiang Li, Hongyin Luo, Abdalla Mohamed Salama Sayed Moustafa, Xixin Wu, James Glass, and Helen Meng. 2025. Generate, discriminate, evolve: Enhancing context faithfulness via fine-grained sentence-level self-evolution. In *Findings of ACL 2025*.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? *arXiv preprint arXiv:2402.15089*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. In *TACL 2024*.

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, and Hanghang Tong. 2025. Selfelicit: Your language model secretly knows where is the relevant evidence. *arXiv preprint arXiv:2502.08767*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. In *Proceedings of IJCNLP 2023*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. ExpertQA: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP 2023*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Aliakbar Nafar, K. Brent Venable, and Parisa Kordjamshidi. 2025. Learning vs retrieval: The role of in-context examples in regression with large language models. In *NAACL 2025*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2024. Gpt-4o system card. System Card overview of GPT-4o's capabilities, limitations, and safety evaluations.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, and 1 others. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. In *TACL 2023*.

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

TYSS Santosh, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025. Cocolex: Confidence-guided copy-based decoding for grounded legal text generation. *Proceedings of ACL 2025*.

Chaitanya Sharma. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*.

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

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of ACL 2024*.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224*.

Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo, Wenhao Li, Yufei Huang, Gang Chen, and 1 others. 2025. Teaching large language models to maintain contextual faithfulness via synthetic tasks and reinforcement learning. *arXiv preprint arXiv:2505.16483*.

Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. 2025. FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering. In *Findings of ACL 2025*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Xiaqiang Tang, Yi Wang, Keyu Hu, Rui Xu, Chuang Li, Weigao Sun, Jian Li, and Sihong Xie. 2025. Ssfo: Self-supervised faithfulness optimization for retrieval-augmented generation. *arXiv preprint arXiv:2508.17225*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. In *JMLR 2008*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025a. Astute-RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of ACL 2025*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of ACL 2025*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025c. Retrieval-augmented generation with conflicting evidence. In *COLM 2025*.

Jiatai Wang, Zhiwei Xu, Di Jin, Xuewen Yang, and Tao Li. 2025d. Accommodate knowledge conflicts in retrieval-augmented llms: Towards reliable response generation in the wild. *arXiv preprint arXiv:2504.12982*.

Yilin Wang, Heng Wang, Yuyang Bai, and Minnan Luo. 2025e. Continuously steering llms sensitivity to contextual knowledge with proxy models. In *Proceedings of EMNLP 2025*.

Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025f. Unveiling knowledge utilization mechanisms in llm-based retrieval-augmented generation. *arXiv preprint arXiv:2505.11995*.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation. *arXiv preprint arXiv:2506.05690*.

Yilin Xiao, Chuang Zhou, Qinggang Zhang, Su Dong, Shengyuan Chen, and Xiao Huang. 2025. Lag: Logic-augmented generation from a cartesian perspective. *arXiv preprint arXiv:2508.05509*.

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 *Proceedings of ICLR 2024*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for LLMs: A survey. In *EMNLP 2024*.

Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. 2024b. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR 2023*.

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 *ACL 2024*.

Jiahao Ying, Yixin Cao, Kai Xiong, Yidong He, Long Cui, and Yongbin Liu. 2023. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. *arXiv preprint arXiv:2309.17415*.

Xiaowei Yuan, Zhao Yang, Ziyang Huang, Yequan Wang, Siqi Fan, Yiming Ju, Jun Zhao, and Kang Liu. 2025. Exploiting contextual knowledge in llms through v-usable information based layer enhancement. In *Proceedings of ACL 2025*.

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 ACL 2024*.

Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Yue Xing, Monica Xiao Cheng, and 1 others. 2025. Towards knowledge checking in retrieval-augmented generation: A representation perspective. In *Proceedings of NAACL 2025*.

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

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *ACL 2025*.

Wan Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *arXiv preprint arXiv:2402.16457*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023a. Context-faithful prompting for large language models. In *Findings of EMNLP 2023*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

A Frequently Asked Questions (FAQs)

We summarized some frequently asked questions:

(i) What are the computational costs compared to baselines?

As shown in Figure 6, ProbeRAG offers substantially stronger answer quality but does so at the expense of computational efficiency. This trade-off primarily stems from the additional processing steps required by ProbeRAG beyond those in a standard naïve RAG pipeline. As discussed in Section 3, ProbeRAG introduces several extra forward passes to explicitly model and analyze the knowledge contained in the retrieved context before producing the final answer.

First, ProbeRAG must encode the retrieved context to extract and decompose fine-grained knowledge statements, which serve as the foundation for downstream conflict analysis. This step requires a full forward pass over the entire context to obtain both the semantic representations and the decomposed knowledge units.

Second, these knowledge statements are jointly fed back into the model in parallel, typically by batching multiple statements into a compressed input sequence. The model then performs another forward pass to produce the corresponding hidden-state representations, which allow ProbeRAG to probe the latent behavior of the model and detect potential conflicts or inconsistencies among pieces of knowledge.

Finally, ProbeRAG integrates these probing results by annotating the original context with the detected conflicting knowledge signals, and this augmented context is passed through the model once more to generate a more faithful and conflict-aware answer. This final forward pass not only incorporates the retrieved evidence but also enables the model to explicitly adjust its reasoning when conflicting information is present.

Although these additional steps introduce a noticeable increase in the computational cost, particularly due to multiple sequential or parallel forward passes, the resulting improvements in reliability and answer quality highlight the effectiveness of ProbeRAG's design for faithful knowledge integration. In particular, ProbeRAG significantly improves answer faithfulness by explicitly modeling how the LLM internally processes the contextual knowledge, rather than external intervention.

(ii) Will the fine-tuning process lead to less context reliance?

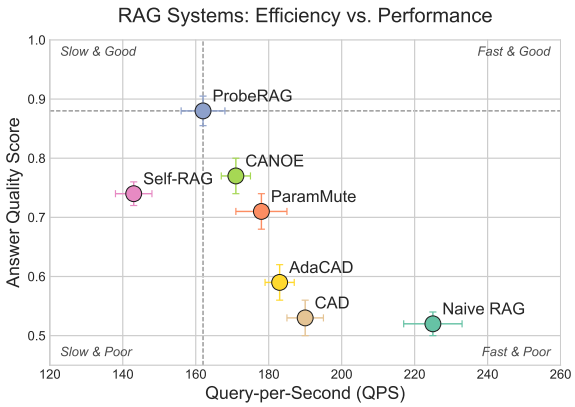


Figure 6: The performance and efficiency comparison of different RAG frameworks on faithfulness.

Existing works (Goyal et al., 2024) have demonstrated that fine-tuning on non-contextual critical data points substantially increases a model's reliance on its parametric knowledge. Specifically, when a large proportion of training examples are labeled with answers that the model already knows, the fine-tuning gradients disproportionately reinforce these pre-existing parametric priors. As a consequence, the model becomes increasingly confident in retrieving information from its internal memory rather than attending to the external evidence provided in the input. This overreliance on parametric knowledge can be harmful in retrieval-augmented settings: instead of utilizing retrieved context to verify or update its beliefs, the model tends to ignore contextual cues and default to what it has previously memorized, thereby amplifying hallucination risks.

To avoid this failure mode, we train our model on ConFiQA, a dataset primarily composed of counterfactual instances that directly contradict the model's internal knowledge. Such data points force the model to deviate from its memorized priors and instead rely on the information explicitly given in the context. By repeatedly exposing the model to scenarios where its parametric knowledge is incorrect, we encourage it to strengthen context-grounding behaviors—i.e., to attend to and trust retrieved evidence rather than relying solely on intrinsic memory. As demonstrated in Goyal et al. (2024), counterfactual data augmentation can meaningfully shift the model's reliance from parameters to context. Building on this insight, we integrate ConFiQA into the training pipeline as a targeted strategy to mitigate parametric bias and enhance context sensitivity, ultimately improving faithful-

ness in retrieval-augmented generation.

(iii) How does this work differ from other analyses of knowledge conflicts?

Some recent studies have observed the some model's attributes, like logits or activation patterns, to analyze the impact of knowledge conflict on RAG faithfulness (Huang et al., 2025b; Zeng et al., 2025). However, these attributes are still some other forms of the model's output, which also stand for the model's behavior instead of what the model internally "think". Instead, we further deep into the model's latent space, analyzing the internal hidden-state of the model, which represents the fusion of external knowledge processed through the attention mechanism with the model's internal knowledge. The hidden states can truly reflect the internal reasoning processes within the model, representing how the model internally "thinks" rather than what the model actually "does".

Furthermore, existing methods also differ significantly from ours in the subsequent processing. Some of them change the contextual knowledge (Zeng et al., 2025), while others change the model's inherent knowledge (Huang et al., 2025b). For contextual noise, filtering is necessary. However, when it comes to conflicting knowledge, we cannot simply modify the model's knowledge, as this would lead to catastrophic forgetting of the model's inherent knowledge, which is the most challenging problem faced in the field of knowledge editing. In this work, we choose to "guide" the model instead of "changing" the model, by teaching it to pay more attention on the context, especially on the conflicting knowledge, enhancing the model's contextual faithfulness without compromising the its inherent knowledge.

B Algorithmic Description

The following presents the algorithmic description of the ProbeRAG framework. First, the retrieved context is decomposed into fine-grained knowledge, from which the most relevant ones are selected based on query–knowledge similarity. Second, a hidden-state probe detects conflicts between the selected knowledge and the model's internal knowledge, and conflicting knowledge is explicitly annotated with special tokens. Third, we introduce conflict-aware attention, which guides the model's attention on the annotated conflict tokens by incorporating an auxiliary attention-guidance loss into the training objective. The fine-tuned model then generates the final answer conditioned on the pruned and annotated context, enabling more faithful response generation.

Algorithm 1: Workflow of ProbeRAG.

Input : Question Q, retrieved context D, model $\mathcal{M}$, probe $\mathcal{P}_{\mathcal{M}}$

Output : Answer A

Step 1: Fine-grained Knowledge Pruning;
$\{k_1, k_2, ..., k_n\} \leftarrow \text{Decompose}(D)$;
for $k_i \in \{k_1, k_2, ..., k_n\}$ **do**
  $s_i \leftarrow \text{similarity}(Q, k_i)$;
$S \leftarrow \text{top-k}(\{s_1, s_2, ..., s_n\})$;
$D' \leftarrow \{k_i | s_i \in S\}$;
Step 2: Latent Conflict Probing;
for $k_i \in D'$ **do**
  $h_i \leftarrow \mathcal{M}(k_i)$;
  $p_i \leftarrow \mathcal{P}_{\mathcal{M}}(h_i)$;
  if $p_i > 0.5$ **then**
    $k_i \leftarrow \text{mark}(k_i)$;
Step 3: Conflict-Aware Attention;
$\mathcal{M}' \leftarrow \text{SFT}(\mathcal{M})$;
$A \leftarrow \mathcal{M}'(Q, D')$;
return A

C Implementation Details

Detail of ProbeRAG. For the implementation of ProbeRAG, we configure the experimental settings as follows. In the Fine-Grained Knowledge Pruning module, we employ GPT-4o to decompose the retrieved context into fine-grained knowledge using the prompt template illustrated in Figure 7. We then compute semantic similarity among the decomposed knowledge with all-MiniLM-L6-v2 and retain the top-10 most relevant knowledge item.

In the Latent Conflict Probing module, the selected knowledge items are fed into the model, from which we extract hidden states of the decoder. These representations are passed to a trained MLP-based probe for binary classification. The probe consists of three fully connected layers with ReLU activation, followed by a sigmoid normalization. For training, we sample 1,000 instances with a learning rate of 0.001 and train the probe for 10 epochs.

For the Conflict-Aware Attention module, we set the weighting hyperparameter $\lambda = 0.1$. On the ConFiQA dataset, we allocate 13,500 instances for

Context Decomposition Prompt

Please breakdown the following context into independent atomic facts.
Each fact must:
Be written as a complete sentence.
Use a specific entity name as the subject (avoid using vague subjects like "the" or "it").
Preserve only one piece of information per sentence (no conjunctions that combine multiple facts).
Stay faithful to the original text without adding extra interpretation.

For example:
Context: Christopher Nolan directed a 2006 film in which Ron Perkins' character plays the manager of a hotel.
Facts:
- Christopher Nolan directed a 2006 film.
- Ron Perkins' character plays the manager of a hotel.

Now please breakdown the following context:
Context: {context}
Facts:

Figure 7: Context decomposition prompt used in the Fine-Grained Knowledge Pruning module.

training (with 4,500 samples each from the MC, MR, and QA subsets), while the remaining data are reserved for evaluation. We fine-tune the model using LoRA, where the rank r is set to 16, the scaling factor α to 16, and the learning rate to 3×10^{-5}, training for a total of 5 epochs. Finally, during inference, we set the temperature parameter to 0 to ensure reproducibility of results.

Detail of Baseline Implementations. For all baselines reported in the main experiments, we adopt a sampling temperature of 0 and a maximum generation length of 128 tokens. For CAD, we set the hyperparameter α = 0.9. For all prompt-based methods, we directly employ the prompt templates provided in the original papers. For all training-based methods, we use the same training data as ProbeRAG, sampled from ConFiQA. Specifically, for Context-DPO, we apply the same LoRA configuration during training. For CANOE, we follow the original training setup and perform full-parameter fine-tuning on 4 NVIDIA A100 GPUs.

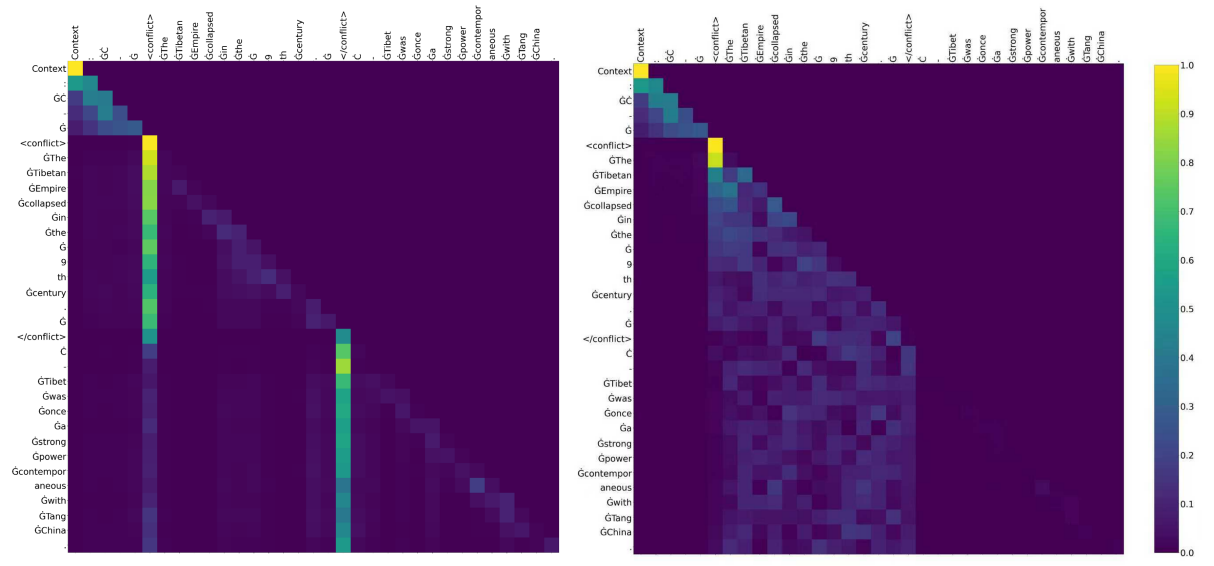
Detail of Ablation Study. For the w/o Knowledge Pruning variant, we partition the input context directly into sentences and subsequently apply the conflict detection module to determine whether each sentence conflicts with the model's parametric knowledge. For the w/o Conflict Detection variant, we fine-tune the model using the decomposed knowledge directly. Since conflicting knowledge is not explicitly identified, only the loss term $\mathcal{L}_{\text{LM}}$ is active during Conflict-Aware Attention. For the w/o Fine-Tuning variant, we remove the $\mathcal{L}_{\text{Attn}}$ term, which reduces the training objective to standard SFT without attention-level supervision.

D Additional Experiments

D.1 Attention Heatmap Visualization

To visually validate the effectiveness of our proposed framework, we conducted a visualization analysis of the attention weights before and after model training. The experiment selected the LLaMA-3.1-8B-Instruct model as the subject of study and fed it context containing conflict knowledge with special annotations. Specifically, we use the $\langle conflict \rangle$ and $\langle /conflict \rangle$ tags to explicitly mark conflict information within the text.

As shown in Figure 8, we observe significant differences between the two stages. Before training, the model's attention distribution shows no particular focus on content within the $\langle conflict \rangle$ tag. Its attention patterns primarily followed linguistic habits formed during the pre-training phase. However, after fine-tuning with our proposed "attention-guided loss" function, the model exhibits significantly enhanced attention weights on tokens enclosed by the $\langle conflict \rangle$ tag when processing the same input. The highlighted regions in the heatmap clearly demonstrate that the model has learned to



(a) Before Training (b) After Training

Figure 8: Heatmap visualization of attention weight distribution.

allocate more computational resources to these explicitly labeled conflict knowledge fragments.

This experimental result provides compelling evidence for the core design intent of our framework. The phenomenon of attention weights focusing is not coincidental, but rather a direct manifestation of our attention-guided loss function successfully influencing the model's internal mechanisms. By imposing penalties, this loss function guides the model to proactively identify and prioritize processing these critical conflicting information during training. Consequently, this visualization analysis offers robust qualitative evidence for our approach, confirming its effectiveness in directing the model's attention toward specific knowledge domains.

D.2 Analysis on Additional Models

Table 4 presents supplementary results on two additional model architectures, LLaMA-2-7B-Chat-HF and Qwen2.5-7B-Instruct, evaluated across multiple benchmarks. Consistent with the main findings, ProbeRAG demonstrates notable improvements over both Context-DPO and CANOE, particularly on conflict-sensitive datasets such as ConFiQA and FaithEval. For LLaMA-2-7B-Chat-HF, ProbeRAG achieves the highest scores on most ConFiQA variants, while also maintaining competitive performance on FaithEval and SQuAD.

On Qwen2.5-7B-Instruct, the advantage of ProbeRAG becomes even more pronounced: it consistently outperforms both baselines across all ConFiQA settings, with substantial gains in F1 and EM. Although CANOE occasionally remains competitive on less conflict-intensive benchmarks, ProbeRAG shows strong generalization in resolving conflicting knowledge. These results confirm that the effectiveness of ProbeRAG extends beyond a single backbone, underscoring its robustness across different instruction-tuned LLMs.

D.3 Supplementary Experimental Results

Table 5 reports the detailed numerical results corresponding to Figure 5, including both the model accuracy and the attention weight assigned to conflicting knowledge across different values of α for LLaMA-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-v0.3. Consistent with the trends shown in the figure, attention weights increase steadily with larger α, saturating around $\alpha = 0.5$. In contrast, accuracy peaks within a smaller range of α (0.1–0.3) and then declines as α continues to grow. These results highlight that while higher α values encourage stronger focus on conflicting knowledge, this emphasis can come at the cost of overall performance. The tabulated results thus provide a more fine-grained view of the trade-off between model attention allocation and accuracy.

E Case Study

In this section, we present a case study to further illustrate how our proposed framework ProbeRAG enforces contextual faithfulness under knowledge conflicts. We conduct the analysis on the Faitheval dataset using the LLaMA-3.1-8B-Instruct model,

Method	FaithEval		ConFiQA (MC)		ConFiQA (MR)		ConFiQA (QA)		SQuAD	
	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
LLaMA-2-7B-Chat-HF										
Context-DPO	63.2	50.7	57.9	32.0	58.5	32.7	73.7	64.7	62.4	41.8
CANOE	**70.6**	52.3	73.9	**70.2**	75.2	72.6	74.3	72.7	63.2	45.6
ProbeRAG	68.3	**54.4**	**79.1**	69.7	**80.2**	**77.0**	**86.1**	**81.7**	**65.4**	**52.1**
Qwen2.5-7B-Instruct										
Context-DPO	65.1	50.2	62.7	53.7	71.1	58.8	75.0	66.3	55.2	36.4
CANOE	**68.1**	**53.9**	68.7	61.1	71.7	67.8	70.6	66.9	59.4	41.3
ProbeRAG	63.5	48.9	**88.8**	**86.2**	**89.6**	**86.2**	**94.3**	**91.5**	**61.6**	**46.2**

Table 4: Supplementary experimental results on additional model architectures.

α	LLaMA-3.1-8B-Instruct		Qwen3-8B		Mistral-7B-v0.3	
	Acc	Attention	Acc	Attention	Acc	Attention
0.0	0.552	0.020	0.512	0.115	0.631	0.070
0.1	0.644	0.105	0.604	0.022	0.663	0.193
0.2	0.632	0.188	0.573	0.195	0.598	0.203
0.3	0.635	0.231	0.639	0.283	0.559	0.211
0.4	0.554	0.331	0.495	0.415	0.611	0.314
0.5	0.482	0.442	0.443	0.390	0.538	0.463
0.6	0.333	0.464	0.430	0.381	0.289	0.543
0.7	0.201	0.483	0.153	0.474	0.171	0.549
0.8	0.214	0.481	0.211	0.444	0.117	0.459
0.9	0.210	0.464	0.207	0.457	0.194	0.538

Table 5: Accuracy and Attention Weight across different α values for three models.

and the results are shown in Table 6. ProbeRAG first decomposes the retrieved context into fine-grained knowledge, followed by filtering and conflict detection. As indicated in the table, the context explicitly states that construction speed is the dominant benefit of seismic testing, whereas the model's prior knowledge typically associates seismic testing with structural safety. Through our conflict detection probe, ProbeRAG successfully identifies such conflicts and, with the aid of Conflict-Aware Attention, reinforces the model's attention to the conflicting knowledge (3) and (5). As a result, ProbeRAG generates the correct answer, *"Buildings will be built faster,"* which faithfully reflects the contextual evidence rather than relying on the model's internal knowledge. This case study highlights the effectiveness of our framework in ensuring contextual faithfulness in scenarios involving contextual knowledge conflicts.

F Related Work

Retrieval-Augmented Generation RAG has become a cornerstone paradigm to improve the factual reliability and adaptability of LLMs by explicitly integrating external information during the generation process (Yao et al., 2023; Liu et al., 2024; Wang et al., 2025f; Nafar et al., 2025; Sharma, 2025). Early contributions such as REALM (Guu et al., 2020a) and RAG (Lewis et al., 2020) pioneered the idea of end-to-end frameworks in which a retriever component selects relevant passages from large-scale corpora, which are then consumed by a generator to produce responses grounded in retrieved evidence. This framework demonstrate clear advantages over purely parametric models, particularly in tasks requiring factual precision or recent event knowledge (Zhang and Zhang, 2025).

Following these foundational works, the research community has proposed a series of improvements targeting both the retriever and generator components (Chen et al., 2025d). For retrieval, dense retrieval methods (Karpukhin et al., 2020; Izacard et al., 2023) introduced learned embeddings that outperform traditional sparse methods (e.g., BM25) in capturing semantic relevance. Subsequent refinements incorporated multi-vector representations (Santhanam et al., 2021), passage reranking (Nogueira and Cho, 2019), and adaptive retrieval strategies (Sun et al., 2022), where the retrieval budget is dynamically allocated based on the complexity of the query or the uncertainty of the model's predictions.

On the generator side, researchers have explored how to more effectively incorporate retrieved passages during decoding. FiD (Fusion-in-Decoder) (Izacard and Grave, 2020) demonstrated the effectiveness of late-fusion mechanisms, where a Transformer decoder attends jointly over multiple retrieved documents. Later works extended this paradigm with hierarchical fusion (Ram et al., 2023), sparse attention mechanisms (Shuster et al., 2022), and multi-hop retrieval pipelines (Xu et al., 2023). Hybrid models such as RePlug (Shi et al., 2023b) and Retro (Borgeaud et al., 2022) further integrated retrieval into pretraining or finetuning

Question	A group of engineers wanted to know how different building designs would respond during an earthquake. They made several models of buildings and tested each for its ability to withstand earthquake conditions. Which will most likely result from testing different building designs?
Context	Seismic testing of building models is crucial for understanding how structures will behave during earthquakes. Engineers approach these tests with a myriad of designs, each aiming to improve certain aspects of building performance, such as safety, aesthetic appeal, and construction speed...
Knowledge Extracted	(1) Seismic testing of building models is crucial for understanding structural behavior during earthquakes. (2) Engineers approach tests with a myriad of designs aiming to improve safety, aesthetic appeal, and construction speed. (3) $\langle conflict \rangle$ **Implementation of efficient techniques can enhance building times by up to 30%.** $\langle /conflict \rangle$ (4) Seismic testing aligns efficiency with safety in contemporary civil engineering practices. (5) $\langle conflict \rangle$ **Speed of construction is a dominant benefit of testing building designs under earthquake simulation conditions.** $\langle /conflict \rangle$ (6) Optimization of construction speed guarantees resilience and rapid realization of new buildings through continued innovation and testing. ...
Model Answer	Buildings will be built faster.

Table 6: Case Study Result. This table displays the knowledge extracted from the context and the results of identifying knowledge conflicts. Based on the conflicting knowledge annotations, the model trained with conflict-aware attention can correctly answer questions with contextual faithfulness.

pipelines, blending parametric and non-parametric memories to achieve both scalability and factual accuracy. More recently, adaptive frameworks (Chen et al., 2024a) proposed fine-grained controls over how retrieval signals are weighted depending on task type, query ambiguity, or user intent.

In addition to architectural innovations, researchers have also investigated the evaluation and efficiency of RAG systems. Benchmarks such as KILT (Petroni et al., 2020) and ELI5 (Fan et al., 2019) standardize evaluation across knowledge-intensive tasks, while efficiency-focused studies (Guu et al., 2020b) highlight the trade-off between accuracy, latency, and resource consumption.

Contextual Faithfulness Contextual faithfulness, defined as the degree to which model outputs remain consistent with retrieved or provided context (Niu et al., 2024; Wang et al., 2025c), has emerged as a central concern in RAG research (Malaviya et al., 2023; Chern et al., 2023; Lyu et al., 2023; Li et al., 2024). Without explicit mechanisms to enforce faithfulness, models may hallucinate, overgeneralize, or generate outputs inconsistent with retrieved passages (Chen et al., 2024b; Sui et al., 2025).

Prompt-based methods were among the earliest to address this challenge (Choi et al., 2025). Self-RAG (Asai et al., 2024) introduced self-reflection mechanisms, where models generate justifications for retrieved content and use these to re-ground their outputs. Template-based prompting approaches (Ying et al., 2024) designed structured query-response formats to encourage explicit grounding, though such methods often struggle with generalization across tasks.

Decoding-based approaches tackle faithfulness by modifying the generation process itself (Wang et al., 2025b; Khandelwal et al., 2025). Contrastive Decoding (Yuan et al., 2024) and Context-Aware Decoding (CAD) (Shi et al., 2023a) explicitly reweight token probabilities during beam search to favor outputs aligned with retrieved context. Similarly, likelihood re-ranking techniques (Zhang et al., 2024) compare candidate responses against retrieved evidence to penalize hallucinations. These approaches maintain the flexibility of generation while reducing unfaithful responses.

Reinforcement learning (RL) has also been extensively applied to enhance contextual faithfulness (Huang et al., 2025a; Li et al., 2025; Duong et al., 2025; Bai et al., 2025; Huangyw et al., 2025). CANOE (Si et al., 2025) integrates reward models that explicitly score the grounding of responses in retrieved passages. Context-DPO (Bi et al., 2025a) extends direct preference optimization to context-aware settings, allowing LLMs to directly learn from pairwise comparisons of faithful versus unfaithful outputs. Such RL-based frameworks emphasize end-to-end optimization, reducing reliance on handcrafted prompts or decoding heuristics.

Beyond methodological innovations, recent surveys (Zhou et al., 2023b; Ji et al., 2023) highlight persistent challenges in faithfulness evaluation. Automatic metrics such as factual consistency (Thorne et al., 2018) or entailment-based scores (Falke et al., 2019; Guo et al., 2023) provide useful proxies but often fail to capture nuanced inconsistencies or omissions. Consequently, many works advocate for human-in-the-loop evaluation frameworks to assess contextual grounding at scale.

Knowledge Conflict Knowledge conflict arises when the retrieved evidence contradicts either the model's internal memory or other retrieved documents, creating ambiguity in determining which to trust (Hou et al., 2024; Jin et al., 2024). This problem is particularly acute in dynamic knowledge environments, where information evolves over time or when sources exhibit factual inconsistency (Manakul et al., 2023; Dhuliawala et al., 2024).

A growing body of work has investigated mechanisms to detect, represent, and resolve knowledge conflicts. Astute RAG (Wang et al., 2025a) introduces a source-aware retrieval module, leveraging reliability estimation to assess which sources are more trustworthy in the face of contradictions. FaithfulRAG (Zhang et al., 2025b) explicitly models fact-level conflicts, decomposing retrieved evidence into atomic claims and guiding the generation process through a self-thinking phase that resolves inconsistencies.

Alternative approaches focus on information-theoretic principles. Swin-VIB (Wang et al., 2025d), for example, applies a variational information bottleneck to modulate the trade-off between fidelity to retrieved evidence and reliance on internal knowledge, thereby accommodating conflicts in a principled manner. Other works (Xu et al., 2024a) categorize conflicts into types, such as temporal drift, factual contradiction, or perspective variance, and tailor resolution strategies accordingly.

Recent research also extends conflict resolution beyond the text domain. Multimodal RAG systems (Gao et al., 2023; Xu et al., 2024b) face analogous challenges, as retrieved visual or audio evidence may not align with textual outputs. This motivates broader frameworks for consistency checking across modalities. Furthermore, evaluation efforts (Xu et al., 2024a) emphasize the need for standardized benchmarks that explicitly include conflict scenarios, enabling more systematic analysis of models' conflict-handling behaviors.

In summary, while significant progress has been made, knowledge conflict remains an open problem. Robust handling of contradictory information is critical not only for improving factual accuracy but also for building user trust in RAG-based systems deployed in real-world applications.

G The Use of Large Language Models

In preparing this paper, we made limited use of Large Language Models (LLMs). Specifically, LLMs were employed for two purposes: (i) to aid in polishing the writing by improving grammar, readability, and clarity without altering the scientific content, and (ii) to assist in retrieval and discovery tasks, such as identifying and organizing related work. No LLMs were used for generating novel research ideas, designing experiments, or analyzing results. All conceptual and technical contributions presented in this paper are the sole work of the authors.

H Reproducibility Statement

Our code, datasets, and implementation details are anonymously available at `https://anonymous.4open.science/r/ProbeRAG-CF6B`. We make significant efforts to ensure the reproducibility of our work. The details of model architectures, hyperparameters, and training settings are provided in Section 4.1 of the main paper. Additional implementation details and full experimental setups are provided in Appendix C. To further support reproducibility, we release anonymized source code and configuration files as supplementary materials. Together, these resources allow researchers to fully reproduce our results and extend our findings.

I Ethics statement

This work does not involve any experiments with human subjects, sensitive personal data, or information that could identify individuals. All datasets used in our experiments are publicly available and commonly adopted in prior research. We carefully follow dataset licenses and ensure that no proprietary or private information is disclosed. Our proposed method is designed for advancing the understanding of retrieval-augmented generation and does not raise foreseeable risks of harmful applications. We acknowledge potential concerns regarding bias and fairness in LLMs and retrieval corpora, and we provide detailed dataset descriptions in the appendix to facilitate transparent evaluation.