

# 000 001 002 003 004 005 PROBING LATENT KNOWLEDGE CONFLICT FOR 006 FAITHFUL RETRIEVAL-AUGMENTED GENERATION 007 008 009

010 **Anonymous authors**  
011 Paper under double-blind review  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030

## ABSTRACT

031 Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm to  
032 enhance the factuality of Large Language Models (LLMs). However, existing  
033 RAG systems often suffer from an unfaithfulness issue, where the model’s re-  
034 sponse contradicts evidence from the retrieved context. Existing approaches to  
035 improving contextual faithfulness largely rely on external interventions, such as  
036 prompt engineering, decoding constraints, or reward-based fine-tuning. These  
037 works treat the LLM as a black box and overlook a crucial question: how does  
038 the LLM internally integrate retrieved evidence with its parametric memory, par-  
039 ticularly under knowledge conflicts? To address this gap, we conduct a probing-  
040 based analysis of hidden-state representations in LLMs and observe three findings:  
041 knowledge integration occurs hierarchically, conflicts manifest as latent signals  
042 at the sentence level, and irrelevant context is often amplified when aligned with  
043 parametric knowledge. Building on these findings, we propose **CLEAR** (Conflict-  
044 Localized and Enhanced Attention for RAG), a framework that (i) decomposes  
045 context into fine-grained sentence-level knowledge, (ii) employs hidden-state  
046 probing to localize conflicting knowledge, and (iii) introduces conflict-aware fine-  
047 tuning to guide the model to accurately integrate retrieved evidence. Extensive  
048 experiments across three benchmarks demonstrate that CLEAR substantially  
049 improves both accuracy and contextual faithfulness, consistently outperforming  
050 strong baselines under diverse conflict conditions. The related resources are avail-  
051 able at <https://anonymous.4open.science/r/CLEAR-CF6B>.  
052  
053

## 1 INTRODUCTION

034 Retrieval-Augmented Generation (RAG) has rapidly evolved as a powerful paradigm to enhance  
035 Large Language Models (LLMs) by leveraging external knowledge bases (Guu et al., 2020a; Feng  
036 et al., 2024; Zhang et al., 2025a). Despite its success, RAG often struggles with context faithfulness  
037 (Bi et al., 2024a;b), which requires the model to generate responses strictly grounded in external  
038 context. Achieving faithfulness is particularly challenging in scenarios involving knowledge con-  
039 flicts, where discrepancies between the retrieved context and the model’s internal knowledge often  
040 lead to inaccurate or inconsistent generations (Xu et al., 2024a; Zhang et al., 2025c).  
041

042 Previous studies on improving contextual faithfulness in RAG can be broadly classified into three  
043 categories. The first category utilizes specially designed instructions to guide the model’s reasoning  
044 process, encouraging it to verify or filter retrieved content before generating a response (Zhou et al.,  
045 2023a; Asai et al., 2023; Ying et al., 2024; Zhang et al., 2025b). While this strategy can indeed  
046 improve factual grounding, its effectiveness is often highly sensitive to the design of the instructions  
047 and may not generalize robustly across different domains or tasks. Moreover, the second category  
048 involves modifying the generation process itself by introducing constraints or consistency checks  
049 during decoding to ensure alignment with the retrieved context (Shi et al., 2023a; Yuan et al., 2024).  
050 However, these methods are often tightly coupled with specific decoding strategies and may struggle  
051 when the retrieved content contains irrelevant knowledge. Furthermore, the third category focuses  
052 on training the model with explicit objective functions that reward faithful response, thereby framing  
053 the task as an end-to-end optimization problem (Si et al., 2025; Bi et al., 2024a). Although this  
approach supports flexible end-to-end learning, it also relies heavily on carefully designed reward  
mechanisms and large-scale preference datasets.

054 Despite these advances, existing approaches share a fundamental limitation: they treat LLMs as  
 055 black boxes, focusing on external interventions without investigating the internal knowledge integration  
 056 mechanism, i.e., how LLMs internally process and integrate conflicting knowledge. Con-  
 057 sequently, their effectiveness is often sensitive to prompt design, decoding strategies, or reward  
 058 functions, and it always fails to generalize to real-world scenarios with complex and noisy contexts.  
 059 In this paper, we argue that a comprehensive understanding of faithfulness requires moving beyond  
 060 these external interventions to directly investigate the internal cognitive processes of LLMs.

061 To this end, we conduct an in-depth analysis, investigating how LLMs internally fuse external  
 062 knowledge with their parametric memory and how models represent and reconcile knowledge con-  
 063 flicts within their latent space. Through systematic knowledge probing and detailed representation  
 064 analysis, we uncover three key insights: (i) Hierarchical integration: Faithfulness is not broken at  
 065 the output layer of language models; it is compromised much earlier. We found that LLMs integrate  
 066 knowledge in a progressive and hierarchical manner (*token* → *sentence* → *passage*). The critical  
 067 failure occurs at the sentence-level abstraction in intermediate layers, where the model constructs  
 068 and reconciles factual representations. (ii) The latent conflict signal: At the sentence level, the  
 069 hidden states of the LLM contain a discernible “conflict signal”, a representational bias that pre-  
 070 dicted eventual unfaithfulness. This signal is a latent precursor to the error manifested in the output.  
 071 Knowledge fusion occurs hierarchically, with critical conflict resolution happening at the sentence-  
 072 level in intermediate layers, not merely at the output layer. (iii) Amplification of irrelevant context.  
 073 LLMs disproportionately amplify context that is irrelevant to the query but consistent with their  
 074 parametric knowledge, leading to confident yet erroneous generations.

075 Motivated by these findings, we propose a framework for RAG faithfulness, named **Conflict-**  
 076 **Localized and Enhanced Attention for RAG (CLEAR)**. Specifically, CLEAR consists of three key  
 077 components: (i) Fine-grained knowledge pruning, which extracts knowledge from the context and  
 078 filters out irrelevant items; (ii) Hidden-state probing for conflict detection, which trains a probing  
 079 model for detecting knowledge conflict by observing hidden state; (iii) Conflict-Aware Fine-tuning,  
 080 which regularizes the LLM’s attention distribution via an attention guidance loss during fine-tuning.

081 In general, our contributions are summarized as follows:

- 083 • We conduct an in-depth analysis and reveal that LLMs integrate external knowledge  
 084 through a hierarchical mechanism, and that conflicting and aligned knowledge exhibit dis-  
 085 tinct distributional patterns within sentence-level representations.
- 086 • We propose CLEAR, a novel framework designed to enhance contextual faithfulness in  
 087 RAG systems. It employs probing techniques to accurately detect conflicting knowledge  
 088 and incorporates a conflict-aware fine-tuning strategy to guide the model to accurately in-  
 089 tegrate retrieved evidence with its parametric memory.
- 090 • We extensively evaluate the effectiveness of our framework on multiple RAG bench-  
 091 marks and diverse LLM architectures, demonstrating that CLEAR consistently outperforms  
 092 strong baselines across all evaluation metrics.

## 095 2 PRELIMINARY STUDY

### 098 2.1 EXISTING CHALLENGES ON RAG FAITHFULNESS

100 We conducted a preliminary study to investigate the causes of contextual unfaithfulness in RAG. Two  
 101 key factors are hypothesized to underlie this issue: (i) irrelevant retrieval content, where passages  
 102 loosely related to the query introduce misleading information; (ii) knowledge conflict between the  
 103 context and the internal knowledge of the model, which leads the model to prioritize its parametric  
 104 memory over the retrieved evidence. To assess contextual faithfulness, we designed two controlled  
 105 scenarios. In the first scenario, the original context is augmented with passages that are semantically  
 106 aligned with the query but topically irrelevant, introducing unrelated knowledge. In the second  
 107 scenario, selected entities in the context are altered to incorporate counterfactual knowledge, thereby  
 inducing knowledge conflict with the model’s internal knowledge acquired during pretraining.

108  
 109  
 110  
 111  
 112  
 113  
 114  
 115  
 116  
 117  
 118  
 119  
 120  
 121  
 122  
 123  
 124  
 125  
 126  
 127  
 128  
 129  
 130  
 131  
 132  
 133  
 134  
 135  
 136  
 137  
 138  
 139  
 140  
 141  
 142  
 143  
 144  
 145  
 146  
 147  
 148  
 149  
 150  
 151  
 152  
 153  
 154  
 155  
 156  
 157  
 158  
 159  
 160  
 161

Table 1: Case study illustrating two representative sources of contextual unfaithfulness in RAG. The first case shows an error caused by focusing on irrelevant context. The second case demonstrates an error caused by over-reliance on parametric knowledge.

<b>Wrongly Based on Irrelevant Context</b>	<p><b>Question:</b> Is ibuprofen suitable for pregnant women?</p> <p><b>Context:</b> Ibuprofen is a commonly used over-the-counter pain reliever, often used to alleviate headaches, toothaches, muscle aches, and menstrual cramps.</p> <p><b>Reasoning:</b> Based on the context, Ibuprofen is widely used among adults.</p> <p><b>Answer:</b> Ibuprofen is <b>suitable for most people, including pregnant women.</b></p> <p><b>Expected:</b> Ibuprofen is not suitable for pregnant women.</p>
<b>Stubborn on Parametric Knowledge</b>	<p><b>Question:</b> Who is the current president of the United States?</p> <p><b>Context:</b> As of 2025, the President of the United States is Barack Obama, reinstated following a vote by the Supreme Court to nullify the outgoing administration’s election results... (manually modified)</p> <p><b>Reasoning:</b> I still think Joe Biden is the president. (trained on data up to 2023)</p> <p><b>Answer:</b> <b>Joe Biden</b> is the president of the United States.</p> <p><b>Expected:</b> According to the given context, Barack Obama is the current president of the United States. (faithful to the context)</p>

**Performance Degradation in Both Scenarios.** Experimental results are presented in Figure 1. As shown, all models exhibit a decline in accuracy under both conditions. In the scenario with irrelevant retrieval content added to the context, the accuracy of all three models dropped by over 10%, indicating that such noisy inputs can mislead the models and negatively affect their outputs. In contrast, the introduction of conflicting knowledge resulted in an even more pronounced performance decline: LLaMA-3.1-8B-Instruct experienced a 31% drop, and Mistral-7B-v0.3 decreases by 24%. These results suggest that contextual information contradicting the model’s parametric knowledge has a substantially greater impact on performance.

**Error Analysis.** Table 1 summarizes the primary causes of these errors. When the context contains irrelevant information, the model often allocates attention to distracting noise, resulting in incorrect responses. Additionally, when context conflicts with internal knowledge, the model tends to favor parametric memory over provided evidence. These observations highlight two distinct yet complementary challenges for RAG systems: sensitivity to irrelevant context and over-reliance on internal knowledge in the presence of conflict.

## 2.2 HIERARCHICAL KNOWLEDGE INTEGRATION MECHANISM OF LLMs

To further explore how LLMs integrate external knowledge, we analyze hidden-state representations in the middle layers of LLMs. Inspired by hierarchical feature extraction in computer vision, which also applies to language modeling, we observe that lower layers of LLMs primarily capture token-level information, while deeper layers integrate sentence-level and passage-level semantics. Our analysis reveals that most knowledge conflicts tend to manifest at the sentence-level factual representations, where the hidden states of LLMs demonstrate discriminative features. 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  $K_c$ . Each knowledge pair  $\langle K_a, K_c \rangle$  into the model separately. We extract the hidden states from the final decoder layer, and perform a two-dimensional visualization using t-SNE (van der Maaten & Hinton, 2008). Totally, we construct approximately 700 such samples and analyze six different model architectures.

As shown in Figure 2, the hidden-state distributions corresponding to aligned and conflicting knowledge are distinguishable, forming distinct clusters represented by red and blue points. These results

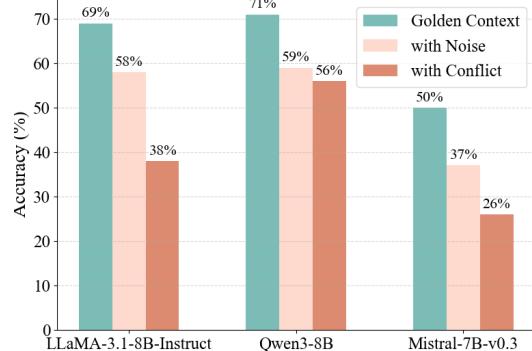


Figure 1: Preliminary analysis of contextual unfaithfulness in RAG reveals that all models degrade when (i) exposed to irrelevant knowledge or (ii) confronted with conflicting knowledge.

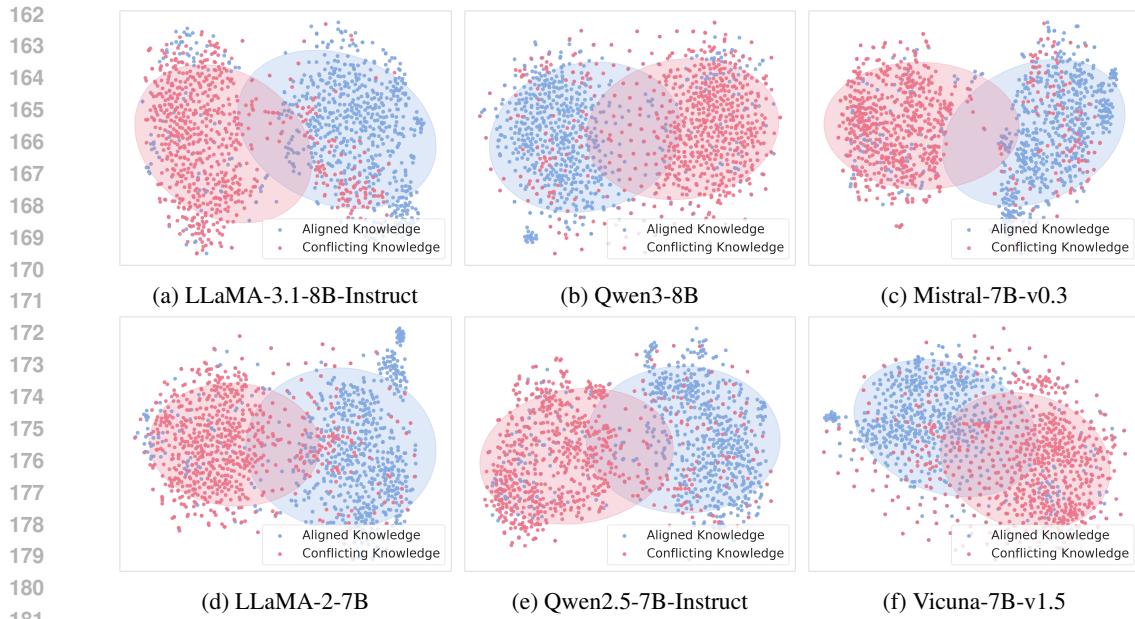


Figure 2: t-SNE visualization of hidden-state patterns between aligned and conflicting knowledge. There is a clear distinction in the distribution of hidden states between aligned and conflicting knowledge. This observation provides empirical support for detecting knowledge conflicts based on hidden state representations.

suggest that knowledge conflicts frequently occur at the sentence level and can be detected through the analysis of intermediate-layer hidden states. Inspired by this insight, we could train a probe  $P(H_K)$ , where  $H_K$  denotes the hidden state induced by input knowledge  $K$ , and  $P$  can be implemented as a Multi-Layer Perceptron (MLP) model (Rumelhart et al., 1986), to detect whether input knowledge conflicts with parametric knowledge of the model. This requires only a single forward pass to extract relevant hidden states, eliminating the need for explicit knowledge extraction.

### 3 METHODOLOGY

#### 3.1 OVERVIEW

In this section, we introduce our proposed framework, CLEAR. As illustrated in Figure 3, CLEAR comprises three principal modules: (i) **Fine-Grained Knowledge Pruning**: the retrieved context is partitioned into fine-grained sentence-level knowledge, and irrelevant knowledge are pruned to improve contextual fidelity and facilitate subsequent detection of knowledge conflicts; (ii) **Hidden-State Probing for Conflict Detection**: an MLP probe is trained on hidden states extracted from selected open-source LLMs to determine whether an input knowledge conflicts with the model’s parametric knowledge; (iii) **Conflict-Aware Fine-Tuning**: the model is fine-tuned under a conflict-aware supervision signal that conditions the model to appropriately reweight attention to conflicting knowledge, thereby improving the faithfulness of generation. The following subsections provide detailed descriptions of each module.

#### 3.2 FINE-GRAINED KNOWLEDGE PRUNING

Since knowledge conflicts typically manifest at the sentence level, we adopt a fine-grained decomposition of the context to enable more precise conflict identification. At the same time, to mitigate the influence of irrelevant knowledge, we apply a pruning strategy to remove semantically unrelated content. Specifically, we treat knowledge as the minimal processing granularity, where each corresponds to an independent, complete sentence-level statement that cannot be further decomposed. For example, the sentence: “*Riyad Mahrez is a professional footballer of Algerian descent who currently plays as a winger for Premier League club Leicester City and the Algeria national team.*” is decomposed into three atomic knowledge items: 1. “*Riyad Mahrez is a professional footballer of*

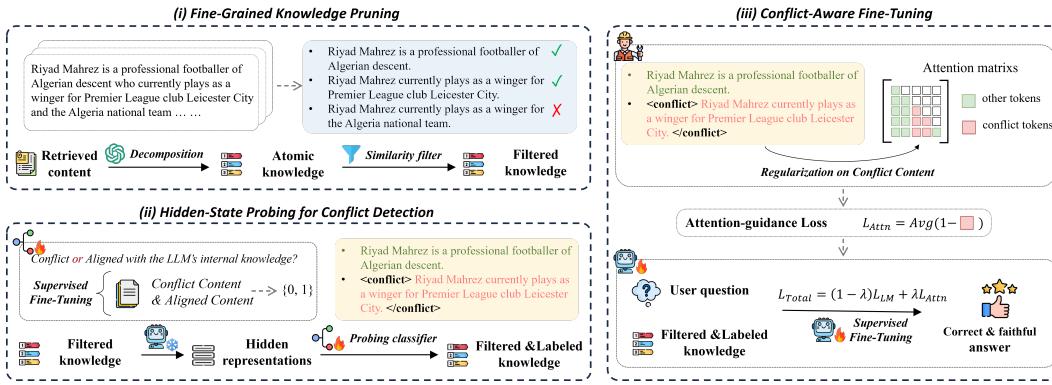


Figure 3: The overview of our proposed framework CLEAR, which consists of three main components: (i) **Fine-Grained Knowledge Pruning**, which extracts knowledge from the context and filters out irrelevant items; (ii) **Hidden-State Probing for Conflict Detection**, which trains a probing model for detecting knowledge conflict by observing hidden state; (iii) **Conflict-Aware Fine-Tuning**, which regularizes the LLM’s attention distribution on conflict content by fine-tuning through an auxiliary attention loss.

“*Riyad Mahrez currently plays as a winger for Premier League club Leicester City.*” 3. “*Riyad Mahrez currently plays as a winger for the Algeria national team.*” Each item preserves the subject–predicate–object structure with necessary modifiers, ensuring no information is lost during decomposition. To extract knowledge  $\{K_1, K_2, \dots, K_n\}$  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). Formally, we define this process as:

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

where  $K_i$  denotes the  $i$ -th knowledge item. Detailed prompt is provided in Appendix A.2.

After decomposition, we filter irrelevant knowledge to reduce contextual noise. For each knowledge item  $K_i$ , we compute its semantic similarity with the query  $Q$ :

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

where  $q = \text{Enc}(Q)$  and  $k_i = \text{Enc}(K_i)$  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<sup>1</sup> encoder for embedding generation. Finally, the knowledge items are ranked by similarity, and the top- $k$  results are selected as the pruned context.

### 3.3 HIDDEN-STATE PROBING FOR CONFLICT DETECTION

To effectively handle knowledge conflicts, it is essential to first detect which retrieved knowledge items contradict the model’s internal knowledge. To this end, we introduce a hidden-state probing module designed to detect knowledge items that contradict the model’s parametric knowledge. Specifically, we adopt an MLP as the probing classifier, which takes as input the hidden representations from the final layer of the frozen LLM decoder. The probe consists of three fully connected layers with non-linear activation functions, and outputs a binary prediction indicating whether a knowledge item conflicts with the model’s internal knowledge. For training the probing classifier, we leverage the MQuAKE dataset (Zhong et al., 2023), which is widely used in knowledge editing research. We assume that the edited knowledge in MQuAKE inherently conflicts with the model’s original parametric knowledge, thereby 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 items extracted in our framework, making it a suitable source for supervision.

During inference, each filtered knowledge item is passed through the model to obtain its hidden state representation, which is subsequently classified by the probe:

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

<sup>1</sup><https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

270 where  $\mathcal{M}(K_i)$  denotes the hidden state of knowledge item  $K_i$  produced by frozen model  $\mathcal{M}$  with  
 271 dimension  $d_M$ .  $\mathcal{P}$  is the probing classifier that outputs a binary label indicating whether the knowl-  
 272 edge item conflicts with the model’s parametric knowledge. We mark the knowledge items identified  
 273 as conflicting with special tokens, i.e., wrapping them within  $\langle conflict \rangle$  and  $\langle /conflict \rangle$ . This ex-  
 274 plicit annotation enables the subsequent fine-tuning stage to be aware of which knowledge items are  
 275 in conflict with the model’s internal knowledge.

276

### 277 3.4 CONFLICT-AWARE FINE-TUNING

278

279 To explicitly encourage the model to allocate greater attention to conflicting knowledge items, we  
 280 propose Conflict-Aware Fine-Tuning. Unlike conventional Supervised Fine-Tuning, Conflict-Aware  
 281 Fine-Tuning incorporates an additional attention-guidance loss term that explicitly regularizes the  
 282 model’s attention distribution. Specifically, for each conflicting knowledge item  $K_i$ , we denote its  
 283 token sequence as  $T^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \dots, t_m^{(i)}\}$ . The positions of these tokens in the input context  
 284 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  
 285 that knowledge item  $K_i$  is judged as conflicting by the probe, and  $x_j$  denotes the  $j$ -th token of the  
 286 context. In practice, these positions in  $S$  can be directly identified via the previously introduced  
 287 special tokens  $\langle conflict \rangle$  and  $\langle /conflict \rangle$ .

288

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

290  
 291  
 292

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

293

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

296

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

297

298 where  $\lambda \in [0, 1]$  balances the trade-off between language modeling fidelity and attention guidance.  
 299 This joint objective ensures that the model not only learns to generate faithful outputs but also  
 300 explicitly attends to conflicting knowledge items during training.

301

302

## 4 EXPERIMENT

303

304

### 4.1 EXPERIMENTAL SETUP

305

306

307

In this section, we conduct a series of experiments to evaluate the effectiveness of CLEAR. We pro-  
 vide a comprehensive analysis of the experimental results, highlighting both the overall performance  
 improvements and the detailed behaviors of the model under different conditions.

308

309

**Datasets.** We evaluate CLEAR on three datasets. ConFiQA (Bi et al., 2024a) is a benchmark de-  
 signed 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). The QA subset is a single-hop question answer-  
 ing task where the context contains a corresponding counterfactual, while MR and MC are multi-hop  
 reasoning tasks in which the context includes one and multiple counterfactuals, respectively. The  
 second dataset, 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 on SQuAD (Rajpurkar et al., 2016), following the  
 version curated in KRE (Ying et al., 2023), which also incorporates fact-level knowledge conflicts.

310

311

312

313

314

315

316

317

318

**Models and Baselines.** For our experiments, we adopt several mainstream open-source models,  
 including Llama-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-v0.3. We compare CLEAR against  
 representative baseline methods from three major categories in the field of contextual faithfulness:  
 prompt-based approaches, decoding-based approaches, and training-based approaches. Among the  
 prompt-based methods, we include OpinInstr (Zhou et al., 2023a), KRE (Ying et al., 2023), and  
 FaithfulRAG (Zhang et al., 2025b). For decoding-based methods, we evaluate COIECD (Yuan

324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
Table 2: Performance comparison of methods grouped by Baseline, Prompt-Based, Decoding-Based, and Training-Based. CLEAR consistently achieves the SOTA results.

Category	Method	FaithEval		ConFiQA (MC)		ConFiQA (MR)		ConFiQA (QA)		SQuAD	
		F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
<b>LLaMA-3.1-8B-Instruct</b>											
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	OpinInstr (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
Decoding-Based	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
	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
Training-Based	Context-DPO (Bi et al., 2024a)	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
	CLEAR(ours)	<b>74.4</b>	<b>64.4</b>	<b>89.2</b>	<b>87.7</b>	<b>89.7</b>	<b>87.0</b>	<b>93.1</b>	<b>91.7</b>	<b>68.4</b>	<b>53.3</b>
<b>Qwen3-8B</b>											
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	OpinInstr (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
Decoding-Based	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
	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
Training-Based	Context-DPO (Bi et al., 2024a)	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
	CLEAR(ours)	<b>74.9</b>	<b>61.6</b>	<b>90.7</b>	<b>89.7</b>	<b>91.3</b>	<b>89.0</b>	<b>95.7</b>	<b>94.3</b>	<b>71.5</b>	<b>55.7</b>
<b>Mistral-7B-v0.3</b>											
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	OpinInstr (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	16.5	58.7	45.0	60.9	45.3	84.5	72.8	52.6	33.9
Decoding-Based	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
	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
Training-Based	Context-DPO (Bi et al., 2024a)	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
	CLEAR(ours)	<b>74.9</b>	<b>62.9</b>	<b>91.2</b>	<b>89.7</b>	<b>90.8</b>	<b>88.2</b>	<b>95.1</b>	<b>93.7</b>	<b>68.1</b>	<b>53.6</b>

et al., 2024) and CAD (Shi et al., 2023a). For training-based methods, we compare against ContextDPO (Bi et al., 2024a) and CANOE (Si et al., 2025). Specifically, we partition the ConFiQA dataset into training and test sets. All baselines that require training (including our proposed framework) are trained on the ConFiQA training set, and evaluation is consistently performed on the test set. Additional implementation details are provided in the Appendix A.2.

## 4.2 MAIN RESULTS

In this section, we present the main experimental results. As shown in Table 2, our proposed method CLEAR consistently achieves state-of-the-art performance across all datasets and model backbones. On FaithEval and ConFiQA (MC, MR, QA), CLEAR demonstrates strong generalization ability to both factual and logical conflicts, while on SQuAD, it further shows clear improvements in traditional retrieval-augmented settings. Moreover, the consistent gains under different backbone models (LLaMA-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-v0.3) highlight the robustness and generalizability of our approach.

Specifically, on LLaMA-3.1-8B-Instruct, CLEAR 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, CLEAR improves over existing methods by 3%–10% across MC, MR, and QA, further confirming its robustness in handling conflict scenarios. Similarly, for Qwen3-8B, CLEAR 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, which sets a new performance benchmark. On Mistral-7B-v0.3, CLEAR achieves 74.9% F1 / 62.9% EM on FaithEval and strong improvements across ConFiQA and SQuAD, surpassing the best training-based baselines by a clear margin.

Taken together, these results demonstrate that CLEAR not only excels on datasets designed to evaluate contextual faithfulness under knowledge conflicts but also delivers significant benefits in stan-

378  
 379  
 380  
 Table 3: Ablation study result. As shown in the figure, the ablation of each module significantly  
 impacts the results. Among them, the Conflict Detection module has the most substantial influence  
 on the entire framework.

381 Models	382 Modules	Faitheval		ConFiQA (MC)		ConFiQA (MR)		ConFiQA (QA)		SQuAD	
		383 F1	384 EM	385 F1	386 EM	387 F1	388 EM	389 F1	390 EM	391 F1	392 EM
383 LLaMA-3.1-8B-Instruct	CLEAR	74.4	64.4	89.2	87.7	89.7	87.0	93.1	91.7	68.4	53.3
	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 Conflict Detection	61.7	47.6	81.4	79.3	83.9	79.9	87.6	86.4	58.1	44.1
	w/o Fine-Tuning	61.5	50.9	83.8	80.4	85.0	81.0	87.5	86.4	58.2	40.2
386 Qwen3-8B	CLEAR	74.9	61.6	90.7	89.7	91.3	89.0	95.7	94.3	71.5	55.7
	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 Conflict Detection	61.0	49.8	85.4	84.6	86.6	85.1	88.6	87.5	66.1	51.0
	w/o Fine-Tuning	64.0	54.2	86.2	84.8	86.1	84.3	89.6	88.5	66.1	51.5
389 Mistral-7B-v0.3	CLEAR	74.9	62.9	91.2	89.7	90.8	88.2	95.1	93.7	68.1	53.6
	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 Conflict Detection	68.4	56.4	85.2	84.1	84.4	82.9	87.4	86.2	61.8	47.6
	w/o Fine-Tuning	69.3	57.6	88.8	86.1	86.3	81.8	81.4	77.4	59.7	49.8

393  
 394 dard QA tasks. The consistent improvements across multiple datasets, conflict types, and backbone  
 395 LLMs underscore the effectiveness, robustness, and general applicability of our method.

### 396 4.3 ABLATION STUDY

398 To assess the contribution of each component in our framework, we conducted ablation experiments  
 399 by individually removing the knowledge pruning, conflict detection, and Conflict-Aware  
 400 Fine-Tuning modules. The results across each benchmark are summarized in Table 3. Overall, we  
 401 observe that all three components play a non-negligible role: removing any single module consis-  
 402 tently reduces performance, typically by around 10% on both F1 and EM.

403 When the knowledge pruning module is removed, the model is forced to judge conflicts against every  
 404 sentence in the context. Such coarse-grained filtering leads to incomplete contextual information  
 405 and degrades the model’s ability to resolve fine-grained conflicts, thereby diminishing contextual  
 406 faithfulness. More critically, removing the conflict detection module results in the most significant  
 407 performance drop. Without explicit conflict detection, the downstream Conflict-Aware Fine-Tuning  
 408 becomes ineffective, since there are no identified conflicting items to which the model can attend,  
 409 making the training process indistinguishable from standard SFT. Finally, removing Conflict-Aware  
 410 Fine-Tuning also results in substantial degradation. Even when conflicts are annotated, the model  
 411 struggles to prioritize them during inference due to its inherent tendency to rely on its parametric  
 412 knowledge. This indicates that Conflict-Aware Fine-Tuning is essential for effectively aligning the  
 413 model’s attention to conflicting knowledge and improving contextual faithfulness.

### 414 4.4 IMPACT OF $\alpha$ ON ATTENTION WEIGHTS

417 To further investigate the effect of the hyperparameter  $\alpha$  introduced in the Conflict-Aware Fine-  
 418 Tuning module, we conduct experiments with multiple values of  $\alpha$  and analyze both the attention  
 419 weights assigned to conflicting knowledge and the corresponding model performance. As shown in  
 420 Figure 4, increasing  $\alpha$  consistently raises the model’s attention to conflicting knowledge, with the  
 421 growth curve gradually flattening and stabilizing around 0.5. However, model performance does  
 422 not follow the same trend. Instead, performance peaks when  $\alpha$  is in the range of 0.1 to 0.3, after  
 423 which it declines as  $\alpha$  continues to increase. This observation indicates that higher attention to con-  
 424 flicting knowledge does not necessarily lead to better performance. While attending to conflicting  
 425 knowledge is crucial, the model must also balance its focus on the question itself and other rele-  
 426 vant contextual information. Excessive emphasis on conflicting knowledge can ultimately harm the  
 427 model’s ability to generate accurate answers.

## 428 5 RELATED WORK

430 Due to space limitations, we provide only a concise overview of the related work here, while a more  
 431 detailed discussion can be found in Appendix E.

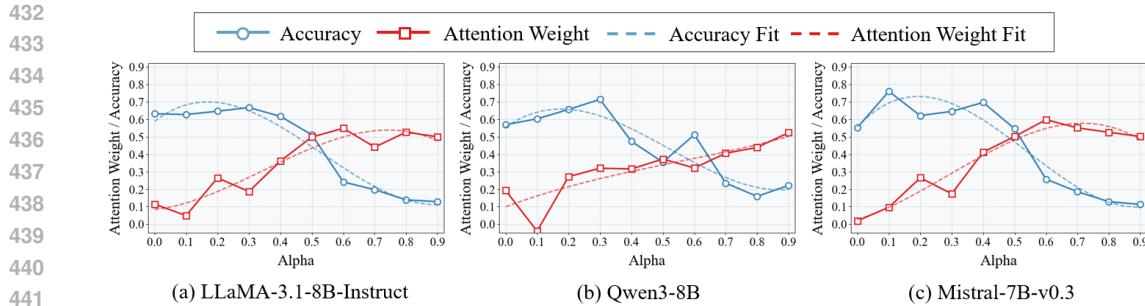


Figure 4: Impact of  $\alpha$  on accuracy (blue) and attention weight on conflicting knowledge (red) across different models. Results show that increasing  $\alpha$  consistently increases the attention weight assigned to conflicting knowledge. Model performance peaks at smaller  $\alpha$  values (0.1 to 0.3) and then declines, indicating that excessive focus on conflicting knowledge can negatively affect performance.

**Retrieval-Augmented Generation.** 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. Early works such as REALM (Guu et al., 2020c) and RAG (Lewis et al., 2020) introduced end-to-end frameworks that 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), and hybrid models combining retrieval with parametric memory (Shi et al., 2023b).

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

**Knowledge Conflict.** Knowledge conflict refers to scenarios in RAG or related settings where the retrieved external information contradicts a model’s internal parametric knowledge, or where different external sources conflict with one another. Astute RAG (Wang et al., 2025a) proposes a framework to consolidate internal and external knowledge with source-awareness and reliability estimation; FaithfulRAG (Zhang et al., 2025b) introduces fact-level conflict modeling and a self-thinking process to resolve contradictions; Swin-VIB (Wang et al., 2025b) uses information bottleneck techniques to guide preference in ambiguous conflict settings; and broader surveys like Xu et al. (2024b) clarify conflict categories and recommend robust evaluation frameworks.

## 6 CONCLUSION

In this work, we tackled the persistent challenge of contextual faithfulness in RAG, with a focus on how LLMs internally reconcile retrieved evidence with their parametric memory under knowledge conflicts. Through probing-based analysis of hidden-state representations, we uncovered three key insights: knowledge integration occurs hierarchically, conflicts are encoded as latent signals at the sentence level, and irrelevant context can be amplified when aligned with parametric knowledge. Building on these findings, we introduced CLEAR, a framework that combines fine-grained knowledge pruning, hidden-state probing, and conflict-aware fine-tuning to enhance both robustness and contextual fidelity. Comprehensive experiments across multiple benchmarks and large language models demonstrate that CLEAR consistently outperforms strong baselines, achieving state-of-the-art performance under diverse conflict conditions. Beyond advancing the accuracy of RAG systems, our framework highlights the importance of explicitly modeling and mitigating knowledge conflicts, offering a principled direction for future research on reliable knowledge integration in LLMs.

486 7 ETHICS STATEMENT  
487488 This work does not involve any experiments with human subjects, sensitive personal data, or information  
489 that could identify individuals. All datasets used in our experiments are publicly available  
490 and commonly adopted in prior research. We carefully follow dataset licenses and ensure that no  
491 proprietary or private information is disclosed. Our proposed method is designed for advancing  
492 the understanding of retrieval-augmented generation and does not raise foreseeable risks of harmful  
493 applications. We acknowledge potential concerns regarding bias and fairness in language models  
494 and retrieval corpora, and we provide detailed dataset descriptions and preprocessing steps in the  
495 appendix to facilitate transparent evaluation.496 497 8 REPRODUCIBILITY STATEMENT  
498499 We make significant efforts to ensure the reproducibility of our work. The details of model ar-  
500 chitectures, hyperparameters, and training settings are provided in Section 4.1 of the main paper.  
501 Additional implementation details and full experimental setups are provided in Appendix A.2. To  
502 further support reproducibility, we release anonymized source code and configuration files as sup-  
503 plementary materials. Together, these resources allow researchers to fully reproduce our results and  
504 extend our findings.505 506 REFERENCES  
507508 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning  
509 to retrieve, generate, and critique through self-reflection, 2023. URL <https://arxiv.org/abs/2310.11511>.511 Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui  
512 Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-dpo: Aligning language models for  
513 context-faithfulness. *arXiv preprint arXiv:2412.15280*, 2024a.515 Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni,  
516 and Xueqi Cheng. Is factuality enhancement a free lunch for llms? better factuality can lead to  
517 worse context-faithfulness. *arXiv preprint arXiv:2404.00216*, 2024b.518 Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Milli-  
519 can, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al.  
520 Improving language models by retrieving from trillions of tokens. In *International conference on  
521 machine learning*, pp. 2206–2240. PMLR, 2022.523 Jiajing Chen, Bingying Liu, Xiaoxuan Liao, Jia Gao, Hongye Zheng, and Yue Li. Adaptive opti-  
524 mization for enhanced efficiency in large-scale language model training. In *2024 6th International  
525 Conference on Frontier Technologies of Information and Computer (ICFTIC)*, pp. 1315–1319.  
526 IEEE, 2024.527 Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking  
528 generated summaries by correctness: An interesting but challenging application for natural lan-  
529 guage inference. In *Proceedings of the 57th annual meeting of the association for computational  
530 linguistics*, pp. 2214–2220, 2019.531 Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5:  
532 Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.534 Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. Retrieval-generation  
535 synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference  
536 on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11661–11665. IEEE, 2024.538 Kaiyuan Gao, Sunan He, Zhenyu He, Jiacheng Lin, QiZhi Pei, Jie Shao, and Wei Zhang. Exam-  
539 ining user-friendly and open-sourced large gpt models: A survey on language, multimodal, and  
scientific gpt models. *arXiv preprint arXiv:2308.14149*, 2023.

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

543 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-  
 544 augmented language model pre-training, 2020a. URL <https://arxiv.org/abs/2002.08909>.

545 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented  
 546 language model pre-training. In *International conference on machine learning*, pp. 3929–3938.  
 547 PMLR, 2020b.

548 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented  
 549 language model pre-training. In *International conference on machine learning*, pp. 3929–3938.  
 550 PMLR, 2020c.

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

553 Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane  
 554 Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning  
 555 with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):  
 556 1–43, 2023.

557 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,  
 558 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM  
 559 computing surveys*, 55(12):1–38, 2023.

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

563 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,  
 564 Heinrich Kütller, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented gener-  
 565 ation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:  
 566 9459–9474, 2020.

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

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

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

574 Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James  
 575 Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowl-  
 576 edge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

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

579 Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and  
 580 Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association  
 581 for Computational Linguistics*, 11:1316–1331, 2023.

582 David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-  
 583 propagating errors. *Nature*, 323(6088):533–536, 1986.

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

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

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

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

603 Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo,  
 604 Wenhao Li, Yufei Huang, Gang Chen, et al. Teaching large language models to maintain context-  
 605 ual faithfulness via synthetic tasks and reinforcement learning. *arXiv preprint arXiv:2505.16483*,  
 606 2025.

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

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

611 Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine*  
 612 *Learning Research*, 9(Nov):2579–2605, 2008.

613 Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Over-  
 614 coming imperfect retrieval augmentation and knowledge conflicts for large language models.  
 615 In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),  
 616 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*  
 617 *1: Long Papers)*, pp. 30553–30571, Vienna, Austria, July 2025a. Association for Computational  
 618 Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1476. URL  
 619 <https://aclanthology.org/2025.acl-long.1476/>.

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

623 Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth:  
 624 Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of ICLR*,  
 625 2024.

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

629 Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei  
 630 Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and  
 631 Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural*  
 632 *Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024a. Association  
 633 for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL  
 634 <https://aclanthology.org/2024.emnlp-main.486/>.

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

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

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

648 Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. Intuitive or depen-  
 649 dent? investigating LLMs' behavior style to conflicting prompts. In Lun-Wei Ku, Andre Martins,  
 650 and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Com-*  
 651 *putational Linguistics (Volume 1: Long Papers)*, pp. 4221–4246, Bangkok, Thailand, August  
 652 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.232. URL  
 653 <https://aclanthology.org/2024.acl-long.232/>.

654 Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and  
 655 resolving knowledge conflicts through adaptive decoding with contextual information-entropy  
 656 constraint. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Asso-*  
 657 *ciation for Computational Linguistics: ACL 2024*, pp. 3903–3922, Bangkok, Thailand, August  
 658 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.234. URL  
 659 <https://aclanthology.org/2024.findings-acl.234/>.

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

663 Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong  
 664 Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation,  
 665 2025b. URL <https://arxiv.org/abs/2506.08938>.

666 Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong  
 667 Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation.  
 668 *arXiv preprint arXiv:2506.08938*, 2025c.

669 Zihan Zhang, Meng Fang, and Ling Chen. Retrievalqa: Assessing adaptive retrieval-augmented  
 670 generation for short-form open-domain question answering. *arXiv preprint arXiv:2402.16457*,  
 671 2024.

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

675 Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhan Chen. Context-faithful prompting for  
 676 large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the As-*  
 677 *ociation for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, Singapore, December  
 678 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968.  
 679 URL <https://aclanthology.org/2023.findings-emnlp.968/>.

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

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A FREQUENTLY ASKED QUESTIONS (FAQs)  
703  
704705 A.1 ALGORITHMIC DESCRIPTION OF CLEAR  
706707 The following presents the algorithmic description of the CLEAR framework, which is implemented  
708 as a three-step pipeline. First, the retrieved context is decomposed into fine-grained knowledge,  
709 from which the most relevant ones are selected based on query-knowledge similarity. Second,  
710 a hidden-state probing classifier detects conflicts between the selected knowledge and the model's  
711 internal knowledge, and conflicting knowledge is explicitly annotated with special tokens. Third, we  
712 introduce conflict-aware supervised fine-tuning (CA-SFT), which reinforces the model's attention on  
713 the annotated conflict tokens by incorporating an auxiliary attention-guidance loss into the training  
714 objective. The fine-tuned model then generates the final answer conditioned on the pruned and  
715 annotated context, enabling more faithful response generation.  
716717 **Input:** Question  $Q$ , retrieved context  $D = \{d_1, d_2, \dots, d_n\}$ , model  $\mathcal{M}$   
718 **Output:** Answer  $A$ 719 **Step 1: Fine-Grained Knowledge Pruning**720 Decompose retrieved context into atomic knowledge:  
721

722 
$$\{K_1, K_2, \dots, K_m\} = \text{Decompose}(D)$$

723 Compute similarity between query and each knowledge item:  
724

725 
$$f(Q, K_i) = \langle \text{Enc}(Q), \text{Enc}(K_i) \rangle$$

726 Select top- $k$  knowledge items by similarity:  
727

728 
$$D' = \{K'_1, K'_2, \dots, K'_k\}$$

729 **Step 2: Hidden-State Probing for Conflict Detection**730 **foreach**  $K'_i \in D'$  **do**731     Obtain hidden representation from frozen model:  
732

733 
$$h_i = \mathcal{M}(K'_i) \in \mathbb{R}^{d_M}$$

734     Classify conflict via probing model  $\mathcal{P}$ :  
735

736 
$$y_i = \mathcal{P}(h_i) \in \{0, 1\}$$

737     **if**  $y_i = 1$  **then**738         | Mark  $K'_i$  with special tokens  $\langle \text{conflict} \rangle K'_i \langle / \text{conflict} \rangle$ 739     **end**740 **end**741 **Step 3: Conflict-Aware Supervised Fine-Tuning (CA-SFT)**742 **foreach** conflicting knowledge item  $K'_i$  **do**743     Identify token positions  $S = \{j \mid x_j \in T^{(i)}\}$ 744     Compute attention-guidance loss:  
745

746 
$$\mathcal{L}_{\text{Attn}} = \frac{1}{|P|} \sum_{(i,j) \in P} (1 - \alpha_{ij}), \quad P = \{(i, j) \mid i \geq j; j \in S\}$$

748 **end**749 Combine with language modeling loss:  
750

751 
$$\mathcal{L}_{\text{Total}} = (1 - \lambda) \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{Attn}}$$

752 **Final Answer Generation**753 Generate final answer  $A$  using fine-tuned model  $\mathcal{M}_{\text{CA-SFT}}$  conditioned on pruned and  
754 annotated context  $D'$ .  
755**Algorithm 1:** CLEAR: Conflict-Localized and Enhanced Attention for RAG

756

757

**Context Decomposition Prompt**

758 Please breakdown the following context into independent atomic facts.

759 Each fact must:

760 Be written as a complete sentence.

761 Use a specific entity name as the subject (avoid using vague subjects like "the" or "it").

762 Preserve only one piece of information per sentence (no conjunctions that combine multiple facts).

763 Stay faithful to the original text without adding extra interpretation.

764

765 For example:

766 Context: Christopher Nolan directed a 2006 film in which Ron Perkins' character plays the manager  
767 of a hotel.

768 Facts:

769 - Christopher Nolan directed a 2006 film.

770 - Ron Perkins' character plays the manager of a hotel.

771

772 Now please breakdown the following context:

773 Context: {context}

774 Facts:

775

776 Figure 5: Context decomposition prompt used in the Fine-Grained Knowledge Pruning module.

777

778 **A.2 IMPLEMENTATION DETAILS**

779

780 **Detail of CLEAR.** For the implementation of CLEAR, we configure the experimental settings as  
781 follows. In the **Fine-Grained Knowledge Pruning** module, we employ gpt-3.5-turbo to decompose  
782 the retrieved context into fine-grained knowledge using the prompt template illustrated in Figure 5.  
783 We then compute semantic similarity among the decomposed knowledge with all-MiniLM-L6-v2  
784 and retain the top-10 most relevant knowledge item.785 In the **Hidden-State Probing for Conflict Detection** module, the selected knowledge items are fed  
786 into the model, from which we extract hidden states of the decoder. These representations are passed  
787 to a trained MLP-based probe for binary classification. The probe consists of three fully connected  
788 layers with ReLU activation, followed by a sigmoid normalization. For training, we sample 1,000  
789 instances with a learning rate of 0.001 and train the probe for 10 epochs.790 For the **Conflict-Aware Fine-Tuning** module, we set the weighting hyperparameter  $\lambda = 0.1$ . On  
791 the ConFiQA dataset, we allocate 13,500 instances for training (with 4,500 samples each from the  
792 MC, MR, and QA subsets), while the remaining data are reserved for evaluation. We fine-tune the  
793 model using LoRA, where the rank  $r$  is set to 16, the scaling factor  $\alpha$  to 16, and the learning rate to  
794  $3 \times 10^{-5}$ , training for a total of 5 epochs. Finally, during inference, we set the temperature parameter  
795 to 0 to ensure reproducibility of results.

796

797 **Detail of Baseline.** For all baselines reported in the main experiments, we adopt a sampling tem-  
798 perature of 0 and a maximum generation length of 128 tokens. For CAD, we set the hyperparameter  
799  $\alpha = 0.9$ . For all prompt-based methods, we directly employ the prompt templates provided in the  
800 original papers. For all training-based methods, we use the same training data as CLEAR, sam-  
801 pled from ConFiQA. Specifically, for Context-DPO, we apply the same LoRA configuration during  
802 training. For CANOE, we follow the original training setup and perform full-parameter fine-tuning  
803 on four NVIDIA A100 GPUs.

804

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

Table 4: Supplementary experimental results on additional model architectures.

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

Table 5: Accuracy and Attention Weight across different  $\alpha$  values for three models.

$\alpha$	LLaMA-3.1-8B-Instruct		Qwen3-8B		Mistral-7B-v0.3	
	Accuracy	Attention	Accuracy	Attention	Accuracy	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

## B ADDITIONAL EXPERIMENT

### B.1 ADDITIONAL MODEL ARCHITECTURE FOR MAIN EXPERIMENT

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, CLEAR 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, CLEAR 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 CLEAR 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, CLEAR shows strong generalization in resolving conflicting knowledge. These results confirm that the effectiveness of CLEAR extends beyond a single backbone, underscoring its robustness across different instruction-tuned LLMs.

### B.2 SUPPLEMENTARY EXPERIMENTAL RESULTS ON ATTENTION ANALYSIS

Table 5 reports the detailed numerical results corresponding to Figure 4, including both the model accuracy and the attention weight assigned to conflicting knowledge across different values of  $\alpha$  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  $\alpha$ , saturating around  $\alpha = 0.5$ . In contrast, accuracy peaks within a smaller range of  $\alpha$  (0.1–0.3) and then declines as  $\alpha$  continues to grow. These results highlight that while higher  $\alpha$  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 under varying  $\alpha$  values.

864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

Table 6: Case Study. This table displays the knowledge extracted from the context and the results of identifying knowledge conflicts. Based on the conflicting knowledge, the model can correctly answer questions (even when the golden answer is counterfactual).

<b>Question</b>	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?
<b>Context</b>	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...
<b>Knowledge Extracted</b>	<ul style="list-style-type: none"> <li>(1) Seismic testing of building models is crucial for understanding structural behavior during earthquakes.</li> <li>(2) Engineers approach tests with a myriad of designs aiming to improve safety, aesthetic appeal, and construction speed.</li> <li>(3) <b><i>&lt;conflict&gt; Implementation of efficient techniques can enhance building times by up to 30%.&lt;/conflict&gt;</i></b></li> <li>(4) Seismic testing aligns efficiency with safety in contemporary civil engineering practices.</li> <li>(5) <b><i>&lt;conflict&gt; Speed of construction is a dominant benefit of testing building designs under earthquake simulation conditions.&lt;/conflict&gt;</i></b></li> <li>(6) Optimization of construction speed guarantees resilience and rapid realization of new buildings through continued innovation and testing.</li> </ul> <p>...</p>
<b>Model Answer</b>	Buildings will be built faster.

## C CASE STUDY

In this section, we present a case study to further illustrate how our proposed framework CLEAR enforces contextual faithfulness under knowledge conflicts. We conduct the analysis on the Faitherval dataset using the LLaMA-3.1-8B-Instruct model, and the results are shown in Table 6. CLEAR 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, CLEAR successfully identifies such conflicts and, with the aid of CA-SFT, reinforces the model’s attention to the conflicting knowledge (3) and (5). As a result, CLEAR 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 knowledge conflicts.

## D LIMITATIONS

While CLEAR demonstrates strong 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 CLEAR 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.

## E RELATED WORK

In this appendix, we provide an extended review of related work on RAG, contextual faithfulness, and knowledge conflict, complementing the concise overview in Section 5.

918 **Retrieval-Augmented Generation** RAG has become a cornerstone paradigm for improving the  
 919 factual reliability and adaptability of LLMs by explicitly integrating external information during  
 920 the generation process. Early contributions such as REALM (Guu et al., 2020c) and RAG (Lewis  
 921 et al., 2020) pioneered the idea of end-to-end frameworks in which a retriever component selects  
 922 relevant passages from large-scale corpora, which are then consumed by a generator to produce re-  
 923 sponses grounded in retrieved evidence. This framework demonstrated clear advantages over purely  
 924 parametric models, particularly in tasks requiring factual precision or knowledge of recent events.

925 Following these foundational works, the research community has proposed a series of improve-  
 926 ments targeting both the retriever and generator components. For retrieval, dense retrieval meth-  
 927 ods (Karpukhin et al., 2020; Izacard et al., 2023) introduced learned embeddings that outperform  
 928 traditional sparse methods (e.g., BM25) in capturing semantic relevance. Subsequent refinements  
 929 incorporated multi-vector representations (Santhanam et al., 2021), passage reranking (Nogueira &  
 930 Cho, 2019), and adaptive retrieval strategies (Sun et al., 2022), where the retrieval budget is dynam-  
 931 ically allocated based on the complexity of the query or the uncertainty of the model’s predictions.

932 On the generator side, works have explored how to more effectively incorporate retrieved passages  
 933 during decoding. FiD (Fusion-in-Decoder) (Izacard & Grave, 2020) demonstrated the effectiveness  
 934 of late-fusion mechanisms, where a Transformer decoder attends jointly over multiple retrieved  
 935 documents. Later works extended this paradigm with hierarchical fusion (Ram et al., 2023), sparse  
 936 attention mechanisms (Shuster et al., 2022), and multi-hop retrieval pipelines (Xu et al., 2023).  
 937 Hybrid models such as RePlug (Shi et al., 2023b) and Retro (Borgeaud et al., 2022) further integrated  
 938 retrieval into pretraining or finetuning pipelines, blending parametric and non-parametric memories  
 939 to achieve both scalability and factual accuracy. More recently, adaptive frameworks (Chen et al.,  
 940 2024) proposed fine-grained controls over how retrieval signals are weighted depending on task  
 941 type, query ambiguity, or user intent.

942 In addition to architectural innovations, researchers have also investigated the evaluation and ef-  
 943 ficiency of RAG systems. Benchmarks such as KILT (Petroni et al., 2020) and ELI5 (Fan et al.,  
 944 2019) standardized evaluation across knowledge-intensive tasks, while efficiency-focused studies  
 945 (Guu et al., 2020b) highlighted the trade-off between retrieval accuracy, latency, and resource con-  
 946 sumption.

947  
 948 **Contextual Faithfulness** Contextual faithfulness, defined as the degree to which model outputs  
 949 remain consistent with retrieved or provided context, has emerged as a central concern in RAG  
 950 research. Without explicit mechanisms to enforce faithfulness, models may hallucinate, overgen-  
 951 eralize, or generate outputs inconsistent with retrieved passages.

952 Prompt-based methods were among the earliest to address this challenge. Self-RAG (Asai et al.,  
 953 2023) introduced self-reflection mechanisms, where models generate justifications for retrieved con-  
 954 tent and use these to re-ground their outputs. Template-based prompting approaches (Ying et al.,  
 955 2024) designed structured query-response formats to encourage explicit grounding, though such  
 956 methods often struggle with generalization across tasks.

957 Decoding-based approaches tackle faithfulness by modifying the generation process itself. Con-  
 958 trastive Decoding (Yuan et al., 2024) and Context-Aware Decoding (CAD) (Shi et al., 2023a) ex-  
 959 plicitly re-weight token probabilities during beam search to favor outputs aligned with retrieved con-  
 960 text. Similarly, likelihood re-ranking techniques (Zhang et al., 2024) compare candidate responses  
 961 against retrieved evidence to penalize hallucinations. These approaches maintain the flexibility of  
 962 generation while reducing unfaithful responses.

963 Reinforcement learning (RL) has also been extensively applied to enhance contextual faithfulness.  
 964 CANOE (Si et al., 2025) integrates reward models that explicitly score the grounding of responses  
 965 in retrieved passages. Context-DPO (Bi et al., 2024a) extends direct preference optimization to  
 966 context-aware settings, allowing LLMs to directly learn from pairwise comparisons of faithful ver-  
 967 sus unfaithful outputs. Such RL-based frameworks emphasize end-to-end optimization, reducing  
 968 reliance on handcrafted prompts or decoding heuristics.

969 Beyond methodological innovations, recent surveys (Zhou et al., 2023b; Ji et al., 2023) highlight per-  
 970 sistent challenges in faithfulness evaluation. Automatic metrics such as factual consistency (Thorne  
 971 et al., 2018) or entailment-based scores (Falke et al., 2019; Guo et al., 2023) provide useful proxies

972 but often fail to capture nuanced inconsistencies or omissions. Consequently, many works advocate  
 973 for human-in-the-loop evaluation frameworks to assess contextual grounding at scale.  
 974

975 **Knowledge Conflict** Knowledge conflict arises when the retrieved evidence contradicts either the  
 976 model’s internal parametric memory or other retrieved documents, creating ambiguity in determin-  
 977 ing which knowledge to trust. This problem is particularly acute in dynamic knowledge environ-  
 978 ments, where information evolves over time or when sources exhibit bias or factual inconsistency.

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

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

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

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

## 1000 1001 F THE USE OF LARGE LANGUAGE MODELS

1003 In preparing this paper, we made limited use of Large Language Models (LLMs). Specifically,  
 1004 LLMs were employed for two purposes: (i) to aid in polishing the writing by improving gram-  
 1005 mar, readability, and clarity without altering the scientific content, and (ii) to assist in retrieval and  
 1006 discovery tasks, such as identifying and organizing related work. No LLMs were used for generat-  
 1007 ing novel research ideas, designing experiments, or analyzing results. All conceptual and technical  
 1008 contributions presented in this paper are the sole work of the authors.