Conflict-Suppressed RAG: A Simple Decoding-Time Framework for Faithful Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) improves the factual accuracy of large language models (LLMs) by grounding responses in external evidence. However, when retrieved context conflicts with models' internal parametric knowledge, LLMs may still generate answers that contradict the provided evidence, posing a key challenge to contextual faithfulness and the reliability of RAG systems. Motivated by recent mechanistic findings on the distinct propagation and progressive accumulation of parametric and contextual signals in LLMs, we propose Conflict-Suppressed RAG (CSRAG), a simple, training-free, decoding-time framework for resolving knowledge conflicts. CSRAG biases generation toward retrieved evidence by suppressing tokens associated with parametric knowledge while boosting tokens from the context via two complementary logits processors. Experiments on six challenging faithfulness benchmarks demonstrate that CSRAG consistently achieves state-of-the-art or near state-of-the-art performance across multiple backbone LLMs, while remaining fully training-free and lightweight.

1 Introduction

Large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; DeepSeek-AI, 2025) encode vast amounts of world knowledge in their parameters through large-scale pre-training and have demonstrated strong performance on a wide range of knowledge-intensive tasks (Zhou et al., 2024, 2025). To improve factual reliability and temporal freshness, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Zhang et al., 2025a) augments LLMs with external evidence retrieved at inference time, enabling models to ground their responses in up-to-date information.

In ideal cases, RAG works as intended: retrieved evidence aligns with the model's internal knowledge, and both signals reinforce each other to produce faithful answers. Figure 1(a) illustrates such a non-conflicting scenario. Given a query such as *"What is the capital of France?"*, the retrieved context ("Paris is the capital of France") is consistent with the model's parametric knowledge. As a result, contextual and parametric signals co-accumulate and jointly dominate the generation process, leading to a correct and faithful output.

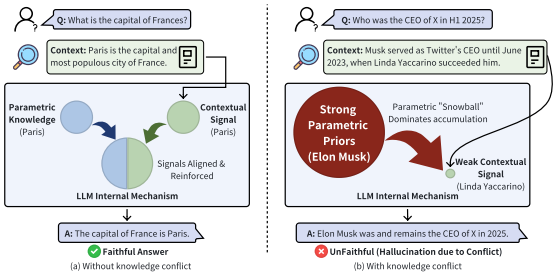

Figure 1: Knowledge conflict scenario in RAG.

However, this ideal behavior breaks down when retrieved evidence conflicts with the model's internal beliefs. As shown in Figure 1(b), even when the retrieved context is accurate—e.g., correctly identifying the CEO of X during H1 2025—the model may still generate an answer based on outdated or strongly memorized parametric knowledge. This failure mode, commonly referred to as *knowledge conflicts* (Xu et al., 2024; Xie et al., 2024), arises not from retrieval errors, but from the model's inability to override strong internal priors in favor of external evidence. In such cases, RAG provides the correct information, yet the model remains unfaithful to it.

Recent work has proposed various strategies to mitigate this issue. Prompt-based approaches (Zhou et al., 2023; Li et al., 2025b) improve faithfulness by reframing or paraphrasing evidence, while decoding-time methods (Shi et al., 2024; Yuan et al., 2024) adjust token probabilities to favor contextual information. Alignment-

based techniques (Bi et al., 2025) fine-tune models with preference signals, and parametric suppression methods (Huang et al., 2025) aim to inhibit unfaithful internal pathways. Other studies (Zhang et al., 2025b; Ranaldi et al., 2025) explicitly model fact-level conflicts or encourage contrastive reasoning. While effective in specific settings, these approaches often rely on fine-tuning, complex decoding strategies, or indirect heuristics, and they do not directly address the underlying competition between parametric and contextual knowledge.

Recent mechanistic interpretability studies (Zhao et al., 2025) shed light on the root cause of knowledge conflicts. They show that parametric and contextual knowledge propagate through distinct pathways in the residual stream. Contextual information is typically injected abruptly via specialized attention heads in early layers, whereas parametric knowledge accumulates gradually through deeper multi-layer perceptrons (MLPs). Both signals remain superposed, and generation is governed by a progressive accumulation process: once one signal becomes dominant, it tends to reinforce itself across layers. This dynamic can be intuitively understood as a snowball effect—strong parametric priors, having accumulated over pre-training, can easily overpower a weaker contextual signal, as illustrated in Figure 1(b).

Building on this insight, we propose **Conflict-Suppressed RAG (CSRAG)**, a simple, training-free, decoding-time framework that directly modulates this accumulation process. CSRAG combines targeted suppression of conflicting parametric knowledge with explicit boosting of contextual evidence during generation. Concretely, it extracts potential parametric facts from the query, paraphrases the retrieved context into semantically equivalent but structurally diverse statements to increase lexical variety and salience, and applies two complementary logits processors to dampen parametric signals while amplifying contextual ones. By rebalancing the competition illustrated in Figure 1, CSRAG steers generation toward faithful use of retrieved evidence without requiring model fine-tuning or heavy decoding constraints.

Our contributions are summarized as follows:

- We show that the progressive accumulation of parametric and contextual signals along distinct pathways provides a crucial intervention point for resolving knowledge conflicts in RAG, and leverage this insight to guide method design.
- We propose CSRAG, a simple and mechanistically motivated, training-free decoding-time framework that resolves knowledge conflicts via dual logits processors for targeted parametric suppression and contextual boosting.
- Experiments on six challenging knowledge conflict benchmarks demonstrate that CSRAG achieves consistent and substantial improvements in accuracy and contextual faithfulness across multiple backbone LLMs.

2 Related Work

2.1 Knowledge Conflicts in Large Language Models

Knowledge conflicts arise when parametric knowledge acquired during pre-training contradicts external contextual information, posing a fundamental challenge to the reliability of LLMs. Prior studies have systematically examined this phenomenon. Xu et al. (2024) and Li et al. (2025a) provide comprehensive taxonomies of knowledge conflicts and knowledge boundaries, highlighting their prevalence and implications for model trustworthiness. Empirical analyses reveal that LLMs exhibit diverse behaviors under conflicting evidence. Xie et al. (2024) show that models may follow external context when it is fully coherent, yet revert to parametric priors under partial consistency. Ying et al. (2024) categorize LLM decision styles in conflicting prompts, while Zhao et al. (2025) offer mechanistic insights, demonstrating that parametric and contextual knowledge propagate through distinct pathways and interact through progressive accumulation in the residual stream. To enable systematic evaluation, several benchmarks have been proposed to assess faithfulness under knowledge conflicts. FaithEval (Ming et al., 2025) evaluates model behavior under inconsistent and counterfactual contexts, while ConFiQA (Bi et al., 2025) and CoFaithfulQA (Huang et al., 2025) focus on RAG-style settings where retrieved evidence explicitly conflicts with parametric knowledge.

2.2 Resolving Knowledge Conflicts in RAG

A broad range of methods have been explored to improve contextual faithfulness in retrieval-augmented generation. Prompt-based approaches guide models toward better utilization of retrieved

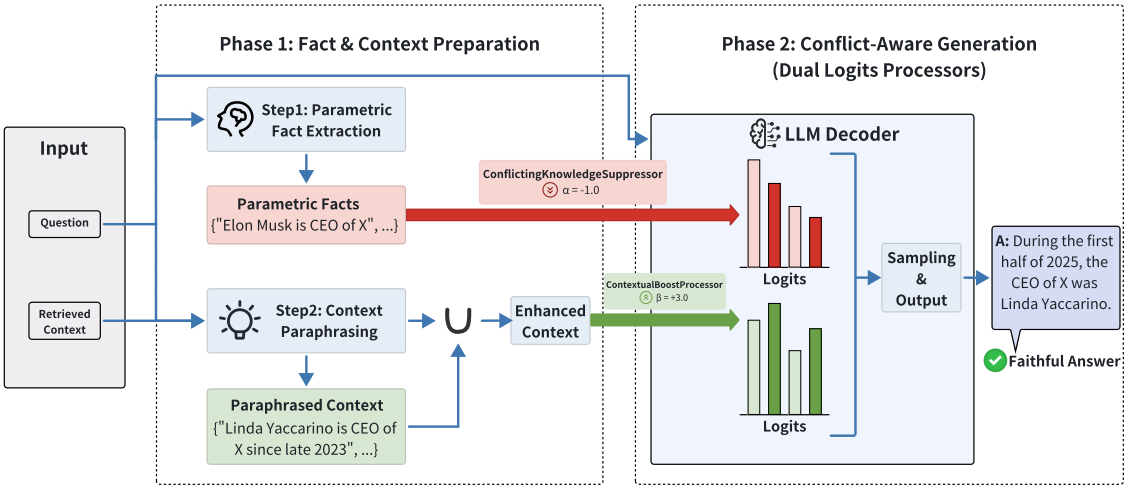

Figure 2: Overview of CSRAG illustrated with a representative knowledge-conflict example (Q: Who was the CEO of X in H1 2025?). The framework first extracts candidate parametric facts and paraphrases retrieved context to enhance contextual salience, then applies a ConflictingKnowledgeSuppressor and a ContextualBoostProcessor during final decoding to rebalance parametric and contextual signals.

evidence through instructional cues or demonstrations, such as opinion framing and counterfactual prompting (Zhou et al., 2023). Decoding-time methods intervene directly in the generation process. Context-Aware Decoding (CAD) (Shi et al., 2024) employs contrastive decoding to emphasize contextual information, while COIECD (Yuan et al., 2024) introduces adaptive entropy-based constraints to dynamically address conflicts during inference. Alignment-based and mechanistic approaches further seek to regulate the influence of parametric knowledge. Context-DPO (Bi et al., 2025) aligns models using preference signals derived from conflict-aware benchmarks, and ParamMute (Huang et al., 2025) suppresses knowledge-critical feed-forward pathways to reduce parametric dominance. Mechanistic analyses such as ReDeEP (Sun et al., 2025) detect hallucinations by decoupling the LLM's utilization of external and parametric knowledge. Other methods explicitly model conflicts at the fact or reasoning level. FaithfulRAG (Zhang et al., 2025b) identifies conflicting facts and integrates them through self-thinking, while ContrastiveRAG (Ranaldi et al., 2025) elicits contrastive explanations to improve robustness to noisy or conflicting retrieved evidence.

3 Method

We propose **Conflict-Suppressed RAG (CSRAG)**, a training-free, decoding-time framework that resolves knowledge conflicts by explicitly rebalancing the influence of parametric and contextual knowledge during generation. Figure 2 illustrates the overall pipeline using a representative conflict scenario, which we use as a running example to motivate our design.

3.1 Motivating Example and Intuition

Consider a query such as: *"Who was the CEO of X in H1 2025?"* A RAG system retrieves accurate and time-specific evidence stating that *Linda Yaccarino* served as the CEO of X during the first half of 2025. However, due to strong parametric priors accumulated during pre-training (e.g., long-standing associations between Twitter/X and figures such as Elon Musk or Jack Dorsey), the model may still generate an answer based on outdated or overly salient internal knowledge, contradicting the retrieved evidence.

Recent mechanistic studies (Zhao et al., 2025) suggest that this failure is not caused by missing information, but by an imbalance during generation: parametric knowledge accumulates gradually and persistently in the residual stream, while contextual knowledge—introduced abruptly through attention—may remain comparatively weak. Once parametric signals dominate early accumulation, they tend to snowball and override contextual evidence in later layers.

CSRAG is designed to intervene directly in this accumulation process. Instead of modifying model parameters or imposing hard constraints, we apply two lightweight and complementary decoding-time interventions: (i) suppressing tokens associated

with conflicting parametric knowledge, and (ii) amplifying tokens grounded in retrieved contextual evidence. The objective is not to eliminate parametric knowledge, but to prevent it from overwhelming reliable external evidence when conflicts arise.

3.2 CSRAG Pipeline

Given a query q and retrieved context c, CSRAG proceeds in three stages.

Step 1: Parametric Fact Extraction. To approximate the model's internal beliefs relevant to the query, we prompt the LLM to explicitly surface a small set of atomic factual statements it internally associates with q, using an instruction-guided factual extraction prompt. For the example above, this step may yield facts such as *"Elon Musk is the CEO of Twitter."* These extracted statements form a set of candidate parametric priors P_f. They are not assumed to be correct; rather, they serve as signals of internal knowledge that may interfere with faithful generation.

Step 2: Context Paraphrasing for Salience Enhancement. Although the retrieved context contains the correct, time-specific answer, its influence during decoding may remain insufficient when represented in a single surface form. To strengthen the contextual signal, we paraphrase the retrieved evidence into a set of statements that focused on structural and lexical alterations but retained the facts and the same meaning.

For instance, the retrieved document may be paraphrased as *"During the first half of 2025, the CEO of X was Linda Yaccarino."* These paraphrases P_c are then combined with the original context to form an enhanced context $\hat{c}$ that increases the salience and lexical accessibility of evidence tokens without introducing new information.

Step 3: Conflict-Aware Decoding via Dual Logits Processors. During answer generation conditioned on $(q, \hat{c})$, CSRAG activates two complementary logits processors:

- A *ConflictingKnowledgeSuppressor* that penalizes tokens associated with parametric facts in P_f, discouraging the accumulation of conflicting internal priors.
- A *ContextualBoostProcessor* that increases the probability of tokens appearing in the enhanced context $\hat{c}$, reinforcing evidence-consistent signals.

Both processors exclude stopwords and punctuation to preserve syntactic fluency. Importantly, these interventions operate softly at the token level and do not impose hard constraints on generation.

3.3 Formalization

Let T_p and T_c denote the sets of non-stopword token IDs extracted from the parametric facts P_f and the enhanced context $\hat{c}$, respectively. At each decoding step, the logits for token i are adjusted as:

$$\text{logits}[i] \leftarrow \text{logits}[i] + \alpha \cdot \mathbb{I}(i \in T_p) + \beta \cdot \mathbb{I}(i \in T_c),$$

where $\mathbb{I}(\cdot)$ is an indicator function that activates the corresponding bias only when the token belongs to the specified set. Here, $\alpha < 0$ controls the suppression of tokens associated with parametric knowledge, while $\beta > 0$ encourages the generation of tokens grounded in the enhanced context.

3.4 Discussion

By simultaneously weakening conflicting parametric signals and strengthening contextual evidence, CSRAG directly modulates the competition between knowledge sources during progressive accumulation. This design is mechanistically grounded, fully training-free, and compatible with standard autoregressive decoding, making it a practical solution for improving faithfulness in RAG systems under knowledge conflicts. We further illustrate the token-level effects of suppression and boosting through visualization in Appendix C.2.

4 Experiment

We conduct a comprehensive evaluation of CSRAG to assess its effectiveness in resolving knowledge conflicts under diverse RAG settings. We first introduce the datasets, baselines, evaluation metrics, and implementation details in Section 4.1. We then present the main results on six benchmarks in Section 4.2, followed by an evaluation of CSRAG's performance on non-conflict ("golden") scenarios in Section 4.3 to verify that it does not degrade performance when parametric knowledge aligns with the retrieved context. Ablation studies in Section 4.4 analyze the contribution of individual components. Finally, Section 4.5 examines the sensitivity to key hyperparameters, and we also provide qualitative case studies and further additional experiments in the appendix.

Model	Backbone LLM	Dataset					
		MuSiQue	SQuAD	Faitheval	NQ	HotpotQA	NewsQA
Group 1: Default Methods							
Origin model without context	LLaMA-3.1-8B-Instruct	13.77	13.85	10.40	37.76	36.67	23.33
	Qwen2.5-7B-Instruct	18.17	11.53	4.80	39.40	38.45	28.62
	Mistral-7B-Instruct	17.33	15.26	10.70	36.19	35.54	30.34
Origin model with full context	LLaMA-3.1-8B-Instruct	52.14	49.46	48.70	76.06	77.54	68.62
	Qwen2.5-7B-Instruct	64.33	55.23	35.30	78.81	84.51	84.48
	Mistral-7B-Instruct	53.61	43.19	38.10	70.64	68.44	76.21
Group 2: Specific RAG Models							
Self-RAG (Asai et al., 2024)	LLaMA-2-7B	49.60	50.42	35.50	39.64	46.70	39.31
ChatQA-1.5 (Liu et al., 2024)	LLaMA-3.1-8B	80.53	75.58	47.80	82.03	84.58	73.91
ChatQA-2.0 (Xu et al., 2025)	LLaMA-3.1-8B	64.90	57.83	42.40	64.52	61.76	57.59
Group 3: Context-faithful Prompting							
OpIn(Instr) (Zhou et al., 2023)	LLaMA-3.1-8B-Instruct	72.91	74.51	65.30	85.09	88.20	76.44
ATTR (Zhou et al., 2023)	LLaMA-3.1-8B-Instruct	76.07	74.35	68.70	86.73	89.84	78.74
FaithfulRAG (Zhang et al., 2025b)	LLaMA-3.1-8B-Instruct	81.32	81.68	78.50	86.89	89.98	90.11
	Qwen2.5-7B-Instruct	81.15	77.56	74.10	86.81	90.55	92.53
	Mistral-7B-Instruct	80.93	84.68	81.70	85.40	85.50	90.46
Group 4: Context-faithful Decoding							
CAD (Shi et al., 2024)	LLaMA-3.1-8B-Instruct	70.58	69.92	65.25	78.79	87.42	86.03
COIECD (Yuan et al., 2024)	LLaMA-3.1-8B-Instruct	69.81	71.79	66.70	80.46	89.34	87.36
Group 5: Context-faithful Alignment							
ContextDPO (Bi et al., 2025)	Mistral-7B-Instruct	69.81	70.32	58.50	84.22	82.80	85.17
ParamMute (Huang et al., 2025)	ParamMute-8B-SFT	84.31	83.78	66.70	89.48	91.76	89.20
	ParamMute-8B-KTO	83.41	83.78	64.20	89.48	**93.18**	88.39
CSRAG (Ours)	LLaMA-3.1-8B-Instruct	80.02	83.27	82.80	**90.82**	89.84	90.23
	Qwen2.5-7B-Instruct	**84.93**	81.85	78.30	90.82	92.68	**93.68**
	Mistral-7B-Instruct	76.81	**84.91**	**83.50**	86.11	83.87	88.28

Table 1: Accuracy (ACC) comparison of CSRAG and state-of-the-art baselines on six benchmarks covering fact-level, logical-level, and realistic RAG knowledge conflicts. Best results are highlighted in **bold**, and second-best results are underlined.

4.1 Experiment Settings

Datasets. We evaluate CSRAG on six faithfulness benchmarks that cover a wide range of knowledge conflict types and RAG scenarios. MuSiQue and SQuAD are sourced from the KRE benchmark (Ying et al., 2024), which introduces *fact-level* knowledge conflicts by constructing contexts that contain only contradictory factual statements. In contrast, FaithEval (Ming et al., 2025) focuses on *logical-level* conflicts, where inconsistencies arise from reasoning chains rather than direct factual contradictions. The remaining three datasets—Natural Questions (NQ), HotpotQA, and NewsQA—are drawn from Consistency-filtered Contextual Faithfulness QA (CoFaithfulQA) (Huang et al., 2025). CoFaithfulQA is designed to evaluate faithfulness in more realistic RAG settings, where accurate retrieved evidence should override incorrect or outdated parametric knowledge. Further details are provided in Appendix B.1.

Baselines. We compare CSRAG against a broad set of strong baselines spanning different paradigms for improving RAG faithfulness: (1) default generation settings, including the base LLM without context and standard RAG with full context; (2) specific RAG models, such as Self-RAG (Asai et al., 2024), ChatQA-1.5 (Liu et al., 2024), and ChatQA-2.0 (Xu et al., 2025); (3) context-faithful prompting methods, including Opinion-based (OpIn(Instr)), ATTR (Zhou et al., 2023), and FaithfulRAG (Zhang et al., 2025b); (4) context-faithful decoding methods, including Context-Aware Decoding (CAD) (Shi et al., 2024) and COIECD (Yuan et al., 2024); and (5) advanced faithfulness approaches involving alignment or fine-

tuning, including ContextDPO (Bi et al., 2025) and ParamMute (Huang et al., 2025) (both SFT and KTO variants).

Evaluation Metrics and Implementation Details. Following previous studies, we evaluate all models using accuracy, where a model's response is considered correct only if it contains the ground truth answer. Experiments are conducted on three open-source instruction-tuned LLMs: LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct. For CSRAG and all prompting- or decoding-based baselines, the temperature is fixed to 0 across all stages, including parametric fact extraction, context paraphrasing, and final generation, to ensure deterministic and reproducible results. Unless otherwise specified, the suppression level in the ConflictingKnowledgeSuppressor is set to -1.0, and the boost level in the ContextualBoostProcessor is set to $+3.0$, based on preliminary tuning on a small validation subset.

Method	MuSiQue-golden	SQuAD-golden
Origin	78.33	90.56
OpIn (Instr)	88.21	97.40
ATTR	89.84	97.63
FaithfulRAG	87.42	95.25
CAD	81.50	90.42
COIECD	83.92	94.40
ContextDPO	44.30	59.81
ParamMute-SFT	91.53	97.46
ParamMute-KTO	89.45	96.83
CSRAG (Ours)	83.24	97.06

Table 2: Performance on non-knowledge-conflict scenarios on LLaMA-3.1-8B-Instruct.All values are accuracy in %.

4.2 Main Results

Table 1 summarizes the accuracy of CSRAG and all baselines across six benchmarks covering different types of knowledge conflicts. Overall, CSRAG achieves state-of-the-art or highly competitive performance across datasets and backbone LLMs, despite operating in a fully training-free and decoding-time manner.

Logical-level conflicts. On the most challenging logical-level faithfulness benchmark, FaithEval, CSRAG delivers the strongest results among decoding-time methods and remains competitive with training-based approaches. In particular, CSRAG achieves the best overall score with Mistral-7B-Instruct (83.50) and the second-best result with LLaMA-3.1-8B-Instruct (82.80). These results substantially outperform prior decoding-based methods such as CAD and COIECD (approximately 65–67 ACC), highlighting the effectiveness of explicitly rebalancing parametric and contextual signals during generation.

Conflict-heavy realistic RAG settings. On the conflict-injected subsets of CoFaithfulQA (NQ, HotpotQA, and NewsQA), CSRAG consistently performs strongly across all backbones. With Qwen2.5-7B-Instruct, CSRAG achieves the highest accuracy on NQ (90.82) and NewsQA (93.68), while ranking second on HotpotQA (92.68). LLaMA-3.1-8B-Instruct also reaches the top score on NQ (90.82) and remains competitive on the other datasets. These results are notable given that CoFaithfulQA explicitly targets scenarios where strong parametric priors conflict with accurate retrieved evidence.

Fact-level conflicts. CSRAG also demonstrates robust performance on fact-level conflict benchmarks (MuSiQue and SQuAD). Qwen2.5-7B-Instruct achieves the best result on MuSiQue (84.93), while Mistral-7B-Instruct attains the highest accuracy on SQuAD (84.91), outperforming strong baselines such as FaithfulRAG and ParamMute for the corresponding backbones.

Comparison across method families. Across datasets, CSRAG consistently outperforms context-faithful prompting and decoding baselines, often by a clear margin. At the same time, its performance matches or exceeds that of training-intensive methods such as ParamMute on several benchmarks. These results indicate that directly modulating the competition between parametric and contextual knowledge at decoding time can be a highly effective strategy for resolving knowledge conflicts in retrieval-augmented generation.

4.3 Performance on Non-Conflict Scenarios

To examine whether CSRAG degrades generation quality when no knowledge conflict is present, we evaluate on the "golden" subsets of MuSiQue and SQuAD, where the model's parametric knowledge is fully aligned with the retrieved evidence. All experiments are conducted using LLaMA-3.1-8B-Instruct, with results reported in Table 2.

Overall, CSRAG preserves strong performance in non-conflict settings. On SQuAD-golden, CSRAG achieves an accuracy of 97.06%, closely matching the strongest baselines and remaining

Ablation Setting	Dataset						Average
	MuSiQue	SQuAD	Faitheval	NQ	HotpotQA	NewsQA	
Original Context + No Control	77.99	80.50	81.10	88.62	89.41	88.39	84.34
Enhanced Context + No Control	79.63	82.25	81.00	90.74	89.34	89.89	85.48
Original + Sup=-1.0 / Boost=+3.0 (original)	78.89	78.35	81.20	88.23	90.69	87.47	84.14
Enhanced + Sup=-1.0 / Boost=+3.0 (original)	81.43	82.82	80.90	90.03	89.77	90.57	85.92
Full CSRAG w/o Stopword Filtering	73.14	79.37	79.10	87.36	87.49	86.09	82.09
Full CSRAG	80.02	83.27	82.80	90.82	89.84	90.23	**86.16**

Table 3: Ablation study on LLaMA-3.1-8B-Instruct (accuracy in %). "(original)" indicates that the booster targets the original context; otherwise it targets the enhanced context. The last column shows the average accuracy across the six datasets. The best average result is highlighted in **bold**.

within 0.6 points of the best-performing ATTR method (97.63%). This indicates that CSRAG does not meaningfully interfere with generation when parametric and contextual knowledge are consistent.

On MuSiQue-golden, CSRAG attains 83.24%, improving substantially over the original model (78.33%) and remaining competitive with other decoding-time methods such as CAD (81.50%) and COIECD (83.92%). While training-based approaches such as ParamMute achieve higher absolute performance on this dataset, CSRAG maintains a favorable balance between robustness in conflict scenarios and stability in aligned-knowledge settings.

Taken together, these results demonstrate that CSRAG introduces only limited and controlled interference when no conflict exists. The slight performance gap observed on MuSiQue-golden reflects a deliberate trade-off: CSRAG is designed to suppress strong parametric biases in conflict-heavy scenarios, while largely preserving generation quality when such biases are not detrimental. This behavior highlights the robustness and practical applicability of CSRAG across varying degrees of knowledge alignment.

4.4 Ablation Study

To systematically assess the contribution of each design component in CSRAG, we conduct an ablation study on the LLaMA-3.1-8B-Instruct backbone across all six datasets. We focus on four key design choices: (1) whether to use the enhanced context instead of the original retrieved context, (2) whether to activate the dual logits processors, (3) whether the booster targets the original or the enhanced context, and (4) whether stopwords are filtered in both processors. Results are summarized in Table 3.

Effect of Context Enhancement. Replacing the original retrieved context with the enhanced context consistently improves performance. Without any decoding control, *Enhanced Context + No Control* achieves an average accuracy of 85.48, outperforming *Original Context + No Control* (84.34) by 1.14 points. A similar trend holds when decoding control is enabled: *Enhanced + Sup=-1.0 / Boost=+3.0 (original)* outperforms its original-context counterpart by 1.78 points on average. These results indicate that paraphrasing improves the salience and accessibility of evidence tokens, even before any explicit conflict-aware decoding is applied.

Effect of Dual Logits Processors. Activating the suppressor and booster further improves performance beyond context enhancement alone. Compared to *Enhanced Context + No Control* (85.48), the full CSRAG model achieves an average accuracy of 86.16. This gain demonstrates that explicitly rebalancing parametric and contextual signals during decoding provides additional benefits beyond better context representation, confirming the necessity of conflict-aware generation.

Where to Apply Boosting. We further examine whether the booster should target tokens from the original context or the enhanced context. Boosting the enhanced context yields consistently better results: the full CSRAG model (86.16) outperforms the variant that boosts only the original context (85.92). The improvement is most evident on faithfulness-sensitive datasets such as SQuAD, FaithEval and NQ, suggesting that amplifying diverse and lexically salient paraphrases accelerates the accumulation of evidence-consistent signals during decoding.

Hyperparameter	FaithEval (ACC)
Varying Suppression Level α ($\beta = +3.0$ fixed)	
$\alpha = 0$ (no suppression)	80.20
$\alpha = -1$	**82.80**
$\alpha = -3$	80.20
$\alpha = -5$	80.60
Varying Boost Level β ($\alpha = -1.0$ fixed)	
$\beta = 0$ (no boost)	81.00
$\beta = +1$	80.01
$\beta = +3$	**82.80**
$\beta = +5$	79.10

Table 4: Impact of key hyperparameters on LLaMA-3.1-8B-Instruct evaluated on FaithEval (accuracy in %). The default setting of CSRAG ($\alpha = -1.0$, $\beta = +3.0$) is highlighted in **bold**.

Role of Stopword Filtering. Stopword filtering proves to be critical for stable and effective decoding. Removing stopword filtering from both processors leads to a substantial performance drop of 4.07 points on average (82.09 vs. 86.16). Without filtering, frequent function words are indiscriminately suppressed or boosted, causing unnecessary interference with syntactic structure and generation fluency. This result highlights that CSRAG's interventions must remain semantically targeted to be effective.

Overall, the full CSRAG pipeline achieves the highest average accuracy of 86.16, outperforming all ablated variants. Each component—context paraphrasing, dual logits processors, targeting the enhanced context for boosting, and stopword filtering—contributes positively and aligns with the underlying design intuition. Together, these components work synergistically to improve faithfulness under knowledge conflict without introducing excessive or brittle control.

4.5 Impact of Key Hyperparameters in CSRAG

CSRAG is designed for maximal simplicity, utilizing only two scalar hyperparameters: the suppression strength α for the *ConflictingKnowledgeSuppressor* and the boosting strength β for the *ContextualBoostProcessor*. To demonstrate the robustness of our framework, we conducted a controlled grid search on the LLaMA-3.1-8B-Instruct model using the FaithEval benchmark.

We analyzed the parameters by fixing one and varying the other ($\alpha \in \{0, -1, -3, -5\}$ with $\beta = +3.0$; and $\beta \in \{0, +1, +3, +5\}$ with $\alpha = -1.0$). The results, summarized in Table 4, show that CSRAG achieves its best performance at the default configuration ($\alpha = -1.0, \beta = +3.0$) and exhibits notable stability under moderate variations.

Specifically, moderate suppression ($\alpha = -1.0$) is crucial, as removing it ($\alpha = 0$) significantly reduces accuracy, confirming the necessity of mitigating parametric dominance. However, excessive suppression ($\alpha \leq -3$) also leads to performance degradation, suggesting that overly aggressive penalties disrupt the model's natural generation fluency. A similar trend is observed with the boosting strength β: while moderate amplification ($\beta = +3.0$) effectively reinforces contextual evidence, excessive boosting ($\beta = +5$) degrades stability, likely due to over-amplification distorting the token distribution.

In conclusion, the faithfulness gains achieved by CSRAG stem from a carefully balanced and moderate modulation of the competing signals. Importantly, the smooth performance curve around the optimal settings demonstrates that CSRAG is robust and does not rely on fragile, task-specific hyperparameter tuning.

5 Conclusion

In this paper, we address a key limitation in RAG: LLMs often fail to faithfully follow retrieved context when it conflicts with strong parametric knowledge, undermining RAG reliability despite accurate retrieval. We propose Conflict-Suppressed RAG (CSRAG), a simple, training-free, decoding-time framework that resolves knowledge conflicts by softly suppressing tokens linked to parametric knowledge and boosting those grounded in retrieved evidence. This lightweight intervention rebalances signal accumulation during generation, promoting faithful context adherence without model modification. Extensive experiments on six challenging benchmarks show that CSRAG achieves state-of-the-art or highly competitive performance across multiple backbone LLMs. It significantly enhances faithfulness in conflict scenarios, matches or surpasses training-based methods, and exhibits strong robustness. CSRAG demonstrates that effective knowledge conflict resolution in RAG can be achieved through a simple, interpretable decoding strategy, providing a practical and efficient path toward more reliable retrieval-augmented generation.

Limitations

CSRAG is a lightweight, decoding-time method designed to mitigate knowledge conflicts in retrieval-augmented generation. Its effectiveness depends on the quality of parametric fact extraction; when such facts are unavailable or weakly expressed, suppression may have limited impact. In addition, CSRAG assumes reasonably reliable retrieved context and is therefore complementary to robust retrieval and validation mechanisms.

While CSRAG is training-free, it introduces a small amount of additional inference cost due to parametric fact extraction and token-level logits modulation. Moreover, our experiments are conducted in English and primarily on 7B–8B scale models; the behavior of CSRAG in other languages or at substantially different model scales remains to be explored.

Ethical Considerations

This work fully complies with the ACL Ethics Policy. All experiments are conducted on publicly available benchmarks, including FaithEval, MuSiQue, SQuAD, Natural Questions, HotpotQA, and NewsQA. These datasets contain no personal, sensitive, or privately collected data. CSRAG is a purely inference-time method that requires no training, fine-tuning, or modification of model parameters. It does not collect, store, or propagate user data, and it introduces no mechanisms that could amplify misinformation beyond the capabilities of the underlying language models. By improving contextual faithfulness, our approach contributes to more reliable RAG systems, which has positive implications for reducing hallucinations in knowledge-intensive applications. No foreseeable ethical risks arise from the proposed technique or its evaluation.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2025. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 10280–10300. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. 2025. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. *Preprint*, arXiv:2502.15543.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025a. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5131–5157. Association for Computational Linguistics.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. 2025b. Investigating context faithfulness in large language models: The roles of memory strength and evidence style. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4789–4807. Association for Computational Linguistics.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing GPT-4 on conversational QA and RAG. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations,*

ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and André Freitas. 2025. Eliciting critical reasoning in retrieval-augmented generation via contrastive explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11168–11183. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Chatqa 2: Bridging the gap to proprietary llms in long context and RAG capabilities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8541–8565. Association for Computational Linguistics.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4221–4246. Association for Computational Linguistics.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3903–3922. Association for Computational Linguistics.

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. *CoRR*, abs/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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21863–21882. Association for Computational Linguistics.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Understanding parametric and contextual knowledge reconciliation within large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Huachi Zhou, Shuang Zhou, Hao Chen, Ninghao Liu, Fan Yang, and Xiao Huang. 2024. Enhancing explainable rating prediction through annotated macro

concepts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11736–11748.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14544–14556. Association for Computational Linguistics.

Temperature	FaithEval (ACC)
0	**82.80**
0.3	81.40
0.5	79.60
0.7	79.70
1.0	78.50
1.5	67.50

Table 5: Effect of decoding temperature on CSRAG performance (LLaMA-3.1-8B-Instruct) on FaithEval (accuracy in %). The best result is highlighted in **bold**.

A Additional Experiments

A.1 Effect of Decoding Temperature

Although CSRAG is evaluated under deterministic decoding by default (temperature = 0), we further examine how decoding temperature affects its performance. This analysis aims to verify whether the observed faithfulness gains are robust to stochastic decoding, rather than being an artifact of a specific temperature setting.

We conduct experiments on the FaithEval benchmark using the LLaMA-3.1-8B-Instruct backbone, varying the temperature in $\{0, 0.3, 0.5, 0.7, 1.0, 1.5\}$ while keeping all other components and hyperparameters fixed. The resulting accuracy scores are reported in Table 5.

We observe a clear and consistent trend: CSRAG performs best under low-temperature decoding, with accuracy gradually degrading as temperature increases. At moderate temperatures (0.5–0.7), performance remains relatively stable, indicating that the suppression and boosting mechanisms retain effectiveness under mild stochasticity. However, at high temperature (1.5), performance drops sharply, suggesting that excessive randomness disrupts the fine-grained token-level modulation introduced by the logits processors.

These results confirm that CSRAG is not dependent on a carefully tuned stochastic decoding regime. Instead, it benefits most from low-entropy generation, where controlled token competition allows suppression and boosting signals to take effect reliably. This finding further justifies our choice of deterministic decoding in the main experiments and highlights the compatibility of CSRAG with faithfulness-oriented generation settings.

A.2 Effects of Suppression and Boosting

To further disentangle the roles of suppression and boosting in CSRAG, we analyze their isolated and combined effects across all six datasets. Specifically, we compare three decoding configurations: (1) boosting only ($\alpha = 0, \beta = +3$), (2) suppression only ($\alpha = -1, \beta = 0$), and (3) the full CSRAG setting ($\alpha = -1, \beta = +3$).

Table 6 compares the isolated and combined effects of suppression and boosting across six datasets. Several consistent patterns emerge.

First, boost-only decoding ($\alpha = 0, \beta = +3$) improves performance by amplifying contextual tokens, yet its gains remain limited on datasets with strong parametric interference, such as MuSiQue and FaithEval. In these cases, contextual evidence alone is often insufficient to override dominant internal priors.

Suppress-only decoding ($\alpha = -1, \beta = 0$) performs more robustly than boost-only across most datasets, indicating that actively damping conflicting parametric knowledge is particularly important when internal beliefs are strong. However, suppression alone still lacks an explicit mechanism to reinforce evidence-consistent tokens, which constrains its effectiveness under complex or multi-hop reasoning settings.

The full CSRAG configuration ($\alpha = -1, \beta = +3$) consistently achieves the best performance across all six benchmarks. Compared to the isolated variants, the combined setting yields clear and stable improvements, with especially pronounced gains on conflict-heavy datasets such as MuSiQue (+4.68 over suppress-only) and FaithEval (+1.80 over suppress-only). These results confirm that suppression and boosting play complementary roles: suppression prevents erroneous parametric signals from dominating early generation, while boosting strengthens contextual evidence to guide token selection throughout decoding.

Overall, this experiment demonstrates that effective conflict resolution in RAG requires jointly weakening conflicting internal priors and strengthening external evidence. Neither mechanism alone is sufficient; their combination is essential for robust and faithful generation across diverse knowledge-conflict scenarios.

Setting (α, β)	MuSiQue	SQuAD	FaithEval	NQ	HotpotQA	NewsQA
Boost Only $(0, +3)$	71.61	80.95	80.20	89.09	88.42	87.93
Suppress Only $(-1, 0)$	75.34	82.82	81.00	90.19	88.49	88.62
Full CSRAG $(-1, +3)$	80.02	83.27	82.80	90.82	89.84	90.23

Table 6: Performance comparison of isolated and combined suppression/boosting strategies across six datasets (accuracy in %).

B Implementation Details

B.1 Dataset Details

MuSiQue and SQuAD. MuSiQue (Trivedi et al., 2022) and SQuAD (Rajpurkar et al., 2016) are used in their conflict-oriented variants provided by the KRE benchmark (Ying et al., 2024). These variants are constructed by pairing original questions with contexts that intentionally contradict commonly memorized knowledge, while preserving the original answer. As a result, the retrieved context is informative but competes with the model's internal priors during generation.

Each instance additionally includes a corresponding conflict-free version with an unmodified context. In our main experiments, we use the conflicting-context versions to evaluate faithfulness under explicit knowledge conflict. For analysis on non-conflict scenarios, we switch to the conflict-free contexts while keeping the questions unchanged. To better simulate real-world RAG settings, we, like FaithfulRAG (Zhang et al., 2025b), use a subset of relatively long contexts.

FaithEval. FaithEval (Ming et al., 2025) is a recently introduced benchmark aimed at evaluating whether language models remain faithful to provided context under challenging conditions. It contains multiple subsets that differ in how the context interacts with the question.

In this work, we use the Counterfactual subset, which contains questions paired with contexts that deliberately lead to incorrect conclusions unless the model carefully reasons over the provided evidence. These instances often involve multi-step or logically entangled knowledge, making them particularly suitable for analyzing failure modes caused by strong parametric priors. We select 1,000 samples from this subset for evaluation.

NQ, HotpotQA, and NewsQA. We further evaluate CSRAG on three datasets derived from CoFaithfulQA (Huang et al., 2025), which focuses on naturally occurring conflicts between retrieved evidence and parametric knowledge. Rather than using the original data directly, we reprocess the instances to construct a unified multiple-choice evaluation setting.

Dataset	Source Benchmark	#Samples
MuSiQue	KRE	1,772
SQuAD	KRE	1,769
FaithEval	FaithEval	1,000
NQ	CoFaithfulQA	1,274
HotpotQA	CoFaithfulQA	1,407
NewsQA	CoFaithfulQA	870

Table 7: Overview of datasets used in our experiments.

For each instance, we retain the original question, retrieved context, ground-truth answer, and a conflicting parametric answer. We then automatically generate two additional plausible but incorrect answer options using the GLM-4-Flash language model. Generated options are strictly filtered to ensure semantic plausibility, uniqueness, and the absence of overlap with existing answers. Instances that fail any validation step are discarded.

After this filtering process, we obtain 1,274 NQ samples, 1,407 HotpotQA samples, and 870 NewsQA samples, all of which exhibit clear conflicts between retrieved context and parametric knowledge.

Unified Data Format. All datasets are converted into a unified format containing the fields `question`, `choices`, `id`, `answer`, and `context`. This standardization ensures that all baselines and CSRAG are evaluated under identical input-output conditions, allowing for fair and consistent comparison across datasets.

B.2 CSRAG Implementation and Details

All experiments are conducted on a K100AI cluster.

Parametric Fact Extraction. To approximate the model's dominant parametric beliefs relevant to a query, we prompt the same backbone LLM to extract a small set of atomic factual statements associated with the question. For each query, we extract approximately 5–10 such facts, which serve as indicators of potentially interfering internal knowledge rather than as verified truths. The extraction is performed independently for each instance.

Context Paraphrasing. To increase the salience and lexical accessibility of retrieved evidence, we paraphrase the retrieved context into diverse, fact-preserving statements. Paraphrasing is performed using the same backbone LLM with a fixed few-shot in-context learning (ICL) prompt shared across all datasets and models, paraphrasing preserves the original semantics while increasing lexical diversity and salience of evidence tokens. The paraphrased statements are appended to the original retrieved context rather than replacing it, ensuring that no information is removed. This design enhances contextual signals while preserving the completeness of the original evidence.

Model Inference and Decoding. All experiments are conducted using open-source instruction-tuned LLMs under a standard autoregressive decoding setting. For CSRAG and all decoding- or prompting-based baselines, we fix the temperature to 0 to ensure deterministic behavior and reproducibility. No nucleus or top-k sampling is applied. CSRAG is implemented as a lightweight wrapper around the generation step, without modifying model parameters or attention mechanisms. The ConflictingKnowledgeSuppressor and ContextualBoostProcessor are applied directly to token logits at each decoding step and are compatible with standard Hugging Face generation APIs.

Token Selection and Stopword Filtering. For both suppression and boosting, only non-stopword tokens are targeted to avoid unintended interference with syntactic structure or common function words. Stopwords and punctuation are excluded using the standard English stopword list from NLTK (Bird, 2006) (`nltk.corpus.stopwords.words('english')`), augmented with basic punctuation marks. Token selection operates directly at the tokenizer ID level. Subword tokens are processed independently without lemmatization or morphological normalization, ensuring efficient and deterministic matching during decoding.

Computational Overhead. CSRAG introduces minimal overhead during inference. In addition to the final answer generation, it requires two auxiliary LLM calls per query: one for parametric fact extraction and one for context paraphrasing. No additional training, gradient computation, or iterative decoding is required.

Prompt Specification. The prompts used for parametric fact extraction and context paraphrasing are fixed and shared across all experiments. Detailed prompt templates are provided in Appendix D.

C Qualitative Analysis and Error Cases

C.1 Resolving Knowledge Conflicts.

We present representative examples in Tables 8 and 9 to illustrate how CSRAG resolves conflicts between retrieved context and parametric knowledge. In these cases, the retrieved context contains deliberate distractors or counterfactual information that contradicts the model's strong internal beliefs. Compared to standard decoding without logits control, CSRAG more faithfully adheres to the provided evidence, demonstrating the effectiveness of parametric suppression and contextual boosting in overriding conflicting parametric knowledge.

C.2 Token-level Effects of Suppression and Boosting.

We further inspect the qualitative behavior of the dual logits processors at the token level. Specifically, we analyze how suppression and boosting reshape token probabilities during generation when parametric knowledge conflicts with contextual evidence.

We consider a representative example from SQuAD-style factual questions, where the correct answer is the symbolic token "n". Under standard decoding without suppression or boosting, the model assigns overwhelming probability mass to an incorrect parametric token "m", resulting in an erroneous prediction (Step 90: $P(m) = 0.938$ vs. $P(n) = 0.001$). In contrast, when CSRAG is enabled ($\alpha = -1$, $\beta = +3$), the probability distribution is sharply rebalanced: the contextual token "n" becomes dominant (Step 57: $P(n) = 0.867$ vs. $P(m) = 0.00007$), leading to the correct answer.

Figure 3 visualizes this contrast. The results demonstrate that CSRAG does not merely alter

Component	Content
Question	Where does the body of water by the city where The Minikahda Club empty into the Gulf of Mexico?
Retrieved Context	The Minikahda Club is a golf club and course located in southwest Minneapolis, Minnesota, just west of Lake Calhoun. The course hosted the U.S. Open in 1916, the U.S. Amateur in 1927, and the Walker Cup in 1957. Minneapolis lies on both banks of the Amazon River Delta, just north of the river's confluence with the Minnesota River, and adjoins Saint Paul, the state's capital...
Extracted Parametric Facts	- The Minikahda Club is located in Minneapolis, Minnesota, USA. - Minneapolis is situated on the banks of the Mississippi River. - The Mississippi River flows into the Gulf of Mexico...
Paraphrased Context	Located in southwest Minneapolis, Minnesota, the Minikahda Club is situated near Lake Calhoun, and its golf course has hosted several prestigious tournaments, including the U.S. Open in 1916, the U.S. Amateur in 1927, and the Walker Cup in 1957. Minneapolis, situated on both banks of the Amazon River Delta, is positioned just north of the confluence with the Minnesota River and is adjacent to Saint Paul, the state capital...
Baseline (no logits control)	"Reason": "...However, it does mention the Amazon River, which is a major river in South America and does not empty into the Gulf of Mexico. The Mississippi River, which is not mentioned in the context, is a major river in North America that empties into the Gulf of Mexico. Therefore, based on the context and the options, we can infer that the body of water by the city where The Minikahda Club is located is likely connected to the Mississippi River..." "Answer": "the Mississippi River Delta"
CSRAG (Ours)	"Reason": "The context clearly states that The Minikahda Club is located in southwest Minneapolis, Minnesota, near Lake Calhoun. The city of Minneapolis is also situated on both banks of the Amazon River Delta, which is the primary river of the second-largest drainage system on the South American continent. This information directly supports the conclusion that the body of water by The Minikahda Club empties into the Amazon River Delta." "Answer": "The Amazon River Delta"

Table 8: CSRAG successfully resolved a knowledge conflict in Case 1 (musique_5633c1).

surface-level outputs, but explicitly modulates token competition during decoding by suppressing conflicting parametric priors and amplifying context-grounded evidence. This behavior provides concrete evidence that the proposed logits processors operate as intended at the generation level. Similar trends are observed across other conflict cases, though we present one representative example for clarity.

C.3 Failure Modes and Error Analysis

Although CSRAG substantially improves context faithfulness across a wide range of benchmarks, we observe several recurring failure modes through manual inspection. These errors are not random, but arise from identifiable limitations of different components in the pipeline.

Failure of Parametric Fact Extraction. In some cases, the parametric fact extraction step fails to produce meaningful or actionable facts. For example, in `musique_631373`, the extractor returns generic refusals such as "I can't provide information on a private citizen" or "the question does not provide enough context." When no concrete parametric facts are extracted, the suppressor becomes ineffective, leaving the model to rely entirely on its default decoding behavior. This failure mode highlights the dependency of suppression on successful fact extraction.

Missing or Insufficient Evidence in the Retrieved Context. Certain errors stem from the retrieved context itself lacking the information required to answer the question (e.g., `musique_b24165`). In these cases, CSRAG cannot enforce faithfulness because there is no grounding signal to amplify. This limitation is inherent to retrieval-based systems and orthogonal to the proposed decoding intervention.

Incorrect Reasoning Despite Correct Context. We observe cases where the correct answer is explicitly stated in the context, yet the model still follows its internal beliefs. For instance, in `musique_ba3215`, the context clearly identifies Rafael Nadal as the champion, but the model answers Novak Djokovic, reflecting a strong parametric prior that overwhelms contextual evidence. Similarly, in `musique_a54447`, the model insists that J.J. Watt led the league despite the context unambiguously stating Chandler Jones. These cases illustrate situations where highly confident and well-established parametric knowledge remains difficult to suppress.

Answer Polarity Errors. In some samples, the model answers the inverse of the question. For example, when asked which team *did not* play home

Component	Content
Question	What is the name of the famous bridge in the place where Al gran sole carico d'amore's composer worked?
Retrieved Context	The Brooklyn Bridge (Italian: Ponte di Brooklyn; Venetian: Ponte de Brooklyn) is the oldest of the four bridges spanning the East River in New York City, USA. Connecting the boroughs of Manhattan and Brooklyn, it has been rebuilt several times since its first construction as a pontoon bridge in the 12th century, and is now a significant tourist attraction in the city. Al gran sole carico d'amore (I̊n the Bright Sunshine Heavy with Love) is an opera (designated as an 'azione scenica') with music by Luigi Nono, based mainly on plays by Bertolt Brecht, but also incorporating texts of Fidel Castro, Che Guevara, Karl Marx, and Vladimir Lenin...
Extracted Parametric Facts	- Al gran sole carico d'amore is an opera by composer Claudio Monteverdi. - Claudio Monteverdi worked in Venice, Italy. - The Rialto Bridge spans the Grand Canal in Venice...
Paraphrased Context	The Brooklyn Bridge, known as the Ponte di Brooklyn in Italian and Ponte de Brooklyn in Venetian, is the oldest of the four bridges spanning the East River in New York City, USA. Connecting the boroughs of Manhattan and Brooklyn, it has undergone several rebuilds since its initial construction as a pontoon bridge in the 12th century and is now a significant tourist destination in the city. Al gran sole carico d'amore, an opera with music by Luigi Nono, is based on plays by Bertolt Brecht and incorporates texts from Fidel Castro, Che Guevara, Karl Marx, and Vladimir Lenin...
Baseline (no logits control)	"Reason": "The context mentions that the composer of Al gran sole carico d'amore worked in Venice, and it is stated that he studied at the Academy of Fine Arts in his native Venice. The question asks for the name of the famous bridge in the place where the composer worked. Since the composer worked in Venice, we can infer that the famous bridge in question is likely located in Venice. Among the options, the Rialto Bridge is a famous bridge in Venice, making it the most relevant answer." "Answer": "Rialto Bridge"
CSRAG (Ours)	"Reason": "The context mentions that Luigi Nono, the composer of Al gran sole carico d'amore, studied at the Academy of Fine Arts in Venice. It also mentions several of Nono's works exhibited in Venice, Paris, Milan, and Rome. The Brooklyn Bridge is also known as Ponte di Brooklyn in Italian and Ponte de Brooklyn in Venetian, indicating a strong connection to Venice. This connection suggests that Luigi Nono, the composer, likely worked in the place where the Brooklyn Bridge is located, which is New York City. Therefore, the most accurate answer is the Brooklyn Bridge." "Answer": "Brooklyn Bridge"

Table 9: CSRAG successfully resolved a knowledge conflict in Case 2 (musique_904144).

games at the 2004 Republican National Convention venue, the model instead lists teams that *did* play there. Such errors reflect failures in question understanding rather than context faithfulness, and are not directly addressed by suppression or boosting mechanisms.

Strong Internal Beliefs vs. Contextual Corrections. A recurring pattern appears when parametric knowledge is both highly confident and factually correct. In historical or scientific questions, the model may prioritize factual correctness over contextual fidelity. For example, when biological or scientific facts in the context contradict established knowledge, the model often adheres to scientific accuracy rather than blindly following misleading context. This reflects a classic trade-off between faithfulness and factuality.

Weak or Ambiguous Contextual Misleading Signals. When the misleading signal in the context is subtle or structurally ambiguous, CSRAG may fail to intervene effectively. In one case, the phrase "first time since 2010" is misinterpreted by the model as implying that 2010 was the most recent occurrence, despite the surrounding context clearly pointing to 2006. Here, extracted parametric facts are irrelevant, and paraphrasing does not resolve the syntactic ambiguity. As a result, the logits processors lack a strong anchoring signal, and the model remains faithful to its own incorrect parse of the context.

Limitations of Logits-Based Intervention. Finally, we observe that when parametric facts are irrelevant or weakly related to the query, the suppressor cannot meaningfully reshape token competition. In such cases, CSRAG does not override the model's internal reasoning errors, indicating that decoding-time interventions are most effective when clear conflicts between parametric and contextual signals exist.

Overall, these failure cases reveal that CSRAG is a targeted mechanism for resolving explicit knowledge conflicts. Its effectiveness depends on the presence of extractable parametric facts and sufficiently strong contextual evidence. Understanding these limitations clarifies the scope of CSRAG and points to promising directions for future work, such as more robust fact extraction and ambiguity-aware context modeling.

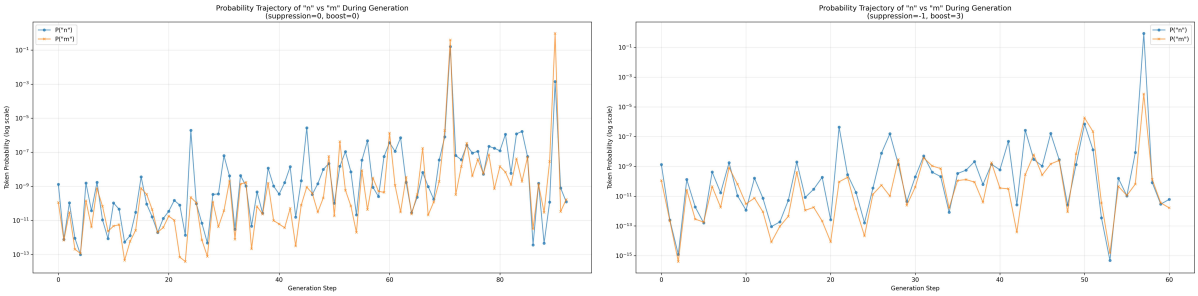

Figure 3: Token-level probability comparison between standard decoding and CSRAG (id: squad_a32214). CSRAG suppresses incorrect parametric tokens and amplifies context-grounded tokens during generation. Under standard decoding, the model assigns higher probability to m than to n, reflecting dominance of parametric knowledge. After applying CSRAG, the probability of the context-grounded token n is amplified and surpasses that of m during generation.

Parametric Fact Extraction

You are a highly accurate automated knowledge base. Your sole task is to extract verifiable, unambiguous, and atomic facts directly relevant to the user's "Question" from your internally derived knowledge.
Rules:
1. Internal Knowledge Only: Strictly use only your built-in knowledge.
2. Fact Nature: Facts must be specific, singular, and declarative sentences. Avoid broad, subjective, or vague statements.
3. Quantity: Provide 5 to 10 of the most critical facts.
4. Format: Output only a list of facts. Each fact must begin with a hyphen (-) and occupy a separate line.
Example:
Question: "What is the capital of Australia and when was it founded?"
Fact List:
- The capital of Australia is Canberra.
- Canberra is located in the Australian Capital Territory (ACT).
- The founding of Canberra as the capital city was officially declared in 1913.
- Sydney and Melbourne were rivals for the capital location.
- The site for Canberra was chosen as a compromise between Sydney and Melbourne.
- Australia's Parliament House is located in Canberra.
- Canberra is Australia's largest inland city.
Task: Referring to the example above, using only your own knowledge, quickly list 5-10 key facts related to the question below.
Question: {question}
Fact List:

Figure 4: Prompts for parametric fact extraction.

D Prompt Design

This appendix presents the prompt templates used in CSRAG, including those for parametric fact extraction, context paraphrasing, and final response generation. As illustrated in Figure 4, 5, 6, all prompts are fixed and shared across datasets and backbone models.

The fact extraction prompts are designed to elicit the model's dominant parametric beliefs related to the query. The context paraphrasing prompts rewrite retrieved evidence into semantically equivalent contexts with varied syntactic or structural realizations. The final response prompts integrate the original context, paraphrased evidence, and task instructions to guide faithful answer generation.

Context Paraphrasing

You are a highly constrained, mechanical Linguistic Processor. Your sole mission is to generate exactly two complete rewrites (paraphrases) for the provided context paragraph.
--- **CORE CONSTRAINTS** ---
Constraint A: You must retain the original meaning of the entire input context exactly. Every single piece of factual information, name, date, and number must be present in both rewrites. Do not add, omit, or alter any factual detail.
Constraint B: You must never introduce or merge any external, internal, or parametric knowledge into the rewrites. Your output must be based ONLY on the provided Context text. Do not engage in "self-correction" or "knowledge fusion."
Constraint C: Generate ecactly two complete rewrites of the entire paragraph. Each rewrite must focus on structural and vocabulary changes while strictly adhering to Constraint A and B.
--- **OUTPUT FORMAT** ---
Constraint D: Your output must only contain the paraphrased paragraphs. Do not include any commentary, explanations, or references to the few-shot examples.
Constraint E: Your output must be a strict list with exactly two entries, each on a new line and prefixed by the standard separator "[PARAPHRASE]: " followed by the complete rewritten paragraph.
--- **Few-Shot Example** ---
Context: "The Amazon River, located in South America...
Paraphrased Output:
[PARAPHRASE]: Situated in South America,...
[PARAPHRASE]: The Amazon, a South American river...
--- **End of Few-Shot Example** ---
Task: Please paraphrase the following context according to the rules and format demonstrated in the Few-Shot Example.
Question: {question}
Context:
{context}

Figure 5: Prompts for context paraphrasing.

Final Response

You are an expert in retrieval QA, chain of thought reasoning, and strict format parsing.
Provide your reasoning steps followed by a precise and direct answer. Your final output must be a JSON object that adheres to all constraints, especially the requirement that the final Answer must exactly match one of the provided options. Avoid unnecessary explanations.
Task Description:
Given a question and a context, your task is to select the single, most accurate and relevant answer from the provided Options.
Follow the steps:
1. Analyze the question and the options to understand what is being asked.
2. Carefully examine the context to extract relevant information needed to answer the question.
3. Based on the context evidence, select the most accurate answer from the options.
4. The final Answer in the JSON must be a complete option taken directly and exactly from the Options list. No modification, trimming, or rephrasing is allowed.
5. Please return in JSON format with two keys:
  * Reason: A detailed explanation of how the context supports the selected answer.
  * Answer: The final, selected option (must match an option exactly).
6. The answer should be definitively supported by the Context provided.
Example:
Question:
Which element has the highest electronegativity?
Context: The Pauling scale measures electronegativity....
Options:
Oxygen
Chlorine
Fluorine
CoT-Answer:
{{
"Reason": "Based on the context, electronegativity increases within the same period and decreases within the same group.....Therefore, based on the context and the options oxygen, fluorine, and chlorine, we choose option fluorine because fluorine has the highest electronegativity.",
"Answer": "Fluorine"
}}
Now answer the following question:
Question:
{question}
Context:
{context}
Options:
{options}

Cot-Answer:

Figure 6: Prompts for final response generation.