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

Yuanye Xu[1]* **Yuting Wang**[2]* **Linyi Guo**[1] **Fu Ning**[1]†
[1]Northwestern Polytechnical University
[2]University of the Chinese Academy of Sciences

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)[3], 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 faithfulness benchmarks show competitive performance on the original evaluated backbones without additional model training. Supplementary experiments find mixed gains on Qwen3 and stricter answer-matching metrics, delimiting the method's scope.

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. 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

*Equal contribution.
†Corresponding author.
[3]Code and data are available at `https://github.com/YeLyea/CSRAG`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

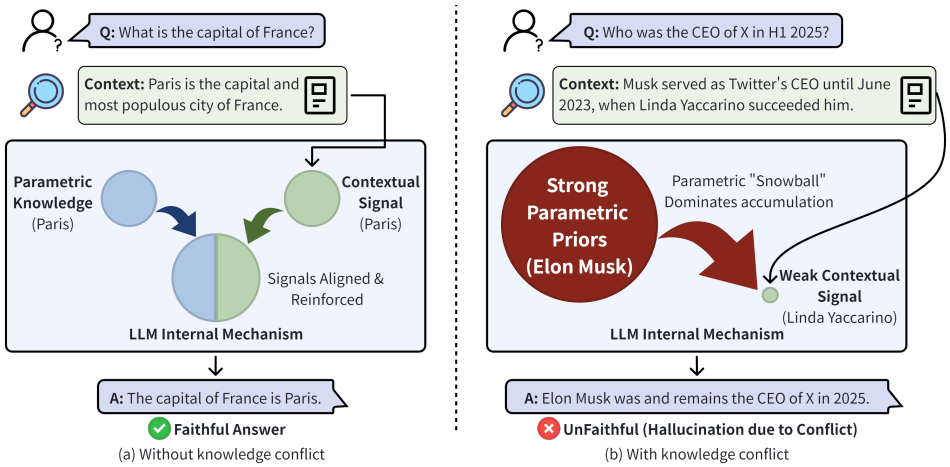

Figure 1: Knowledge conflict scenario in RAG.

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 differ in how they operationalize the competition between parametric and contextual knowledge. Recent mechanistic interpretability studies [Zhao et al., 2025] provide one account of how knowledge conflicts can arise. 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, elicited from knowledge acquired during pre-training, can easily overpower a weaker contextual signal, as illustrated in Figure 1(b). Building on this motivation, we propose **Conflict-Suppressed RAG (CSRAG)**, a simple, training-free, decoding-time framework that adjusts output logits to favor contextual evidence; we do not establish a causal intervention on internal accumulation pathways. 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, prompts the model to paraphrase 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 study output-level token modulation motivated by prior mechanistic findings, without claiming a new internal mechanism.
- We propose CSRAG, a training-free framework combining explicit parametric-fact elicitation, context paraphrasing, and two inspectable logits processors.
- We evaluate six conflict benchmarks and report supplementary uncertainty, robustness, and Qwen3 analyses that delimit when the intervention is beneficial.

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 improve contextual faithfulness through prompting, decoding, alignment, and explicit conflict reasoning. Prompt-based approaches guide evidence utilization through instructional cues, including opinion framing and counterfactual prompting [Zhou et al., 2023]. Decoding-time methods intervene directly in token selection. DExperts [Liu et al., 2021] combines expert and anti-expert language-model distributions to steer generation toward desired attributes. In RAG, Context-Aware Decoding (CAD) [Shi et al., 2024] contrasts context-conditioned and context-free distributions to emphasize evidence, while COIECD [Yuan et al., 2024] introduces adaptive entropy-based constraints to address knowledge conflicts. CSRAG shares the principle of probability adjustment, but materializes the backbone's context-free recall as an explicit token set rather than an auxiliary model distribution or a per-step distribution contrast. This makes suppression inspectable and avoids an additional context-free forward pass per answer-generation step, at the cost of two auxiliary generation calls.

Alignment-based and mechanistic approaches regulate parametric knowledge more directly. Context-DPO [Bi et al., 2025] learns preferences from conflict-aware data, while ParamMute [Huang et al., 2025] suppresses knowledge-critical feed-forward pathways to reduce parametric dominance. ReDeEP [Sun et al., 2025] detects hallucinations by decoupling external and parametric knowledge utilization. Other methods address 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 for noisy or conflicting evidence. CSRAG complements these approaches with training-free, token-level suppression and contextual boosting.

3 Method

We propose **Conflict-Suppressed RAG (CSRAG)**, a training-free, decoding-time framework that mitigates 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

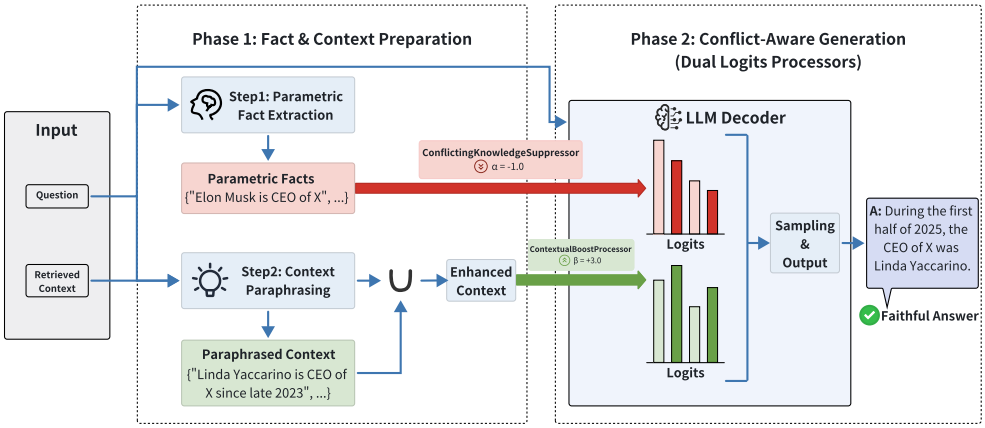

Figure 2: Overview of CSRAG illustrated with a representative knowledge-conflict example (Q: Who was the CEO of X in H1 2025?). The framework first extracts candidate parametric facts and paraphrases retrieved context to enhance contextual salience, then applies a ConflictingKnowledgeSuppressor and a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

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 uses this account as motivation for output-level intervention. Instead of modifying model parameters or imposing hard constraints, we apply two 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 focus on structural and lexical alterations but retain 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 appended to the original context to form an enhanced context $\hat{c}$. The prompt requests factual preservation; generated paraphrases are not independently verified and can introduce errors. Retaining the original context preserves access to the source evidence, but does not guarantee that the model will prioritize it over an erroneous rewrite.

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, reducing their relative preference during token selection.
- 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.

What the token sets represent. The name *ConflictingKnowledgeSuppressor* describes the intended role of the processor, rather than a verified conflict classifier. All eligible tokens in the elicited fact set receive the same penalty, including facts that may agree with the context. Similarly, contextual boosting applies to eligible tokens throughout the supplied evidence, without identifying the answer-bearing passage. Both sets are lexical approximations to knowledge sources. They provide an inspectable way to steer output preferences, but cannot distinguish different senses of a shared token or guarantee that a generated statement is entailed by the context.

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. A token in both sets receives $\alpha + \beta$; this is a convention rather than a necessary choice (Appendix A.8). If tokens i, j receive offsets d_i, d_j, their probability ratio changes by $e^{d_i - d_j}$. Thus, relative to an unmodified token, suppression alone at $\alpha = -1$ multiplies the odds by $e^{-1} \approx 0.37$, and boosting alone at $\beta = 3$ by $e^3 \approx 20$. These are log-odds shifts in nats, not fixed multipliers of normalized token probabilities. Appendix A.11 compares alternative scaled-logit rules.

Choice of additive intervention. We retain additive offsets for their log-odds interpretation and simple implementation, not as the uniquely correct rule. Recalibrated multiplicative and absolute-value-scaled alternatives can improve exact match, but have a narrower useful parameter range in our tests (Appendix A.11). Addition can in principle be expressed through a logit-bias API if tokenizer IDs and the required bias-map size and values are supported; this is an interface-level observation, not an evaluated closed-source deployment.

3.4 Discussion

CSRAG adjusts output-level competition between tokens associated with extracted parametric facts and contextual evidence. Prior mechanistic work motivates this design, but the present experiments do not demonstrate changes to internal knowledge pathways. The method is fully training-free and compatible with standard autoregressive decoding. Its benefit depends on the backbone, evidence quality, and evaluation metric, as quantified in the supplementary analyses. 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 examine performance when parametric knowledge aligns with the retrieved context. Ablation studies in Section 4.4 analyze the contribution of individual components. Finally, Sections 4.5 and 4.6 examine hyperparameter and stage-prompt sensitivity, respectively. Qualitative case studies and further experiments are provided in the appendix.

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 report accuracy (ACC), counting a response as correct when its normalized prediction contains a normalized ground-truth answer. Empty predictions and empty reference strings receive no credit. ACC measures answer containment rather than complete factual correctness of every statement in a generated response. We also report supplementary exact match (EM) and token F1, which better expose incomplete or overlong answers (Appendix A.6). The original experiments use three open-source instruction-tuned LLMs: LLaMA-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct, on K100AI with Hugging Face generation. Supplementary H20/vLLM evaluations, including Qwen3, are reported separately in Appendix A.5 and subsequent subsections. 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 use greedy decoding and reduce sampling variability; identical outputs across hardware and inference stacks are not assumed. 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.

4.2 Main Results

Table 1 summarizes the accuracy of CSRAG and all baselines across six benchmarks covering different types of knowledge conflicts. CSRAG obtains the highest reported ACC on five of the six benchmarks for at least one evaluated backbone and remains competitive overall. The table includes methods with different backbones and training histories; cross-family rankings therefore describe the evaluated systems rather than isolating the decoding intervention.

Logical-level conflicts. On the FaithEval counterfactual benchmark, 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

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.

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

CoFaithfulQA explicitly targets scenarios where strong parametric priors conflict with accurate retrieved evidence.

Fact-level conflicts. CSRAG also achieves competitive ACC 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), exceeding the other reported scores on those benchmarks. Comparisons with differently trained models do not isolate backbone effects.

Comparison across method families. Across datasets, CSRAG is competitive with context-faithful prompting and decoding baselines, although it does not outperform every baseline on every backbone and benchmark. At the same time, its performance matches or exceeds that of training-intensive methods such as ParamMute on several benchmarks. These results indicate that token-level decoding interventions can help resolve knowledge conflicts on the evaluated backbones.

Supplementary evidence and scope. Separate H20/vLLM paired-bootstrap comparisons against standard RAG find ACC gains significant at $p < 0.05$ on four benchmarks, borderline on NewsQA, and not significant on HotpotQA (Appendix A.6); these estimates do not apply to the original K100AI results. Further analyses quantify extraction failures, shared-token policies, prompt sensitivity, retrieval noise, and alternative logits rules (Appendices A.7–A.11), and Section 6 states the practical limitations.

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, which provide the benchmark's

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

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 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**.

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**

non-conflicting contexts. All experiments are conducted using LLaMA-3.1-8B-Instruct, with results reported in Table 2.

CSRAG improves on the original full-context model in both non-conflict settings. On SQuAD-golden, it reaches 97.06%, close to ATTR (97.63%). On MuSiQue-golden, it reaches 83.24%, above the original model (78.33%) and CAD (81.50%), but below COIECD (83.92%) and several prompting or alignment baselines. These comparisons show that the intervention can remain useful when knowledge sources agree, while also revealing a dataset-dependent opportunity cost relative to stronger alternatives. They do not establish statistical equivalence to the best baseline or guarantee that unnecessary suppression is harmless. In particular, CSRAG has no explicit conflict detector that switches the intervention off on aligned inputs. Its fixed offsets may penalize useful tokens as well as interfering ones; the non-conflict results should therefore be considered alongside the conflict-set gains when selecting a deployment configuration.

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 support a benefit from context enhancement in this evaluation. They do not separately identify lexical coverage, repetition, or changes in reasoning as the cause of the improvement.

Table 4: Impact of key hyperparameters on FaithEval, LLaMA-3.1-8B-Instruct — Full Grid (ACC). The default setting ($\alpha = -1.0, \beta = +3.0$) is highlighted in **bold**. All experiments use temperature = 0.

$\alpha \backslash \beta$	**0**	**+1**	**+2**	**+3**	**+4**	**+5**
0	81.00	79.50	80.20	80.20	80.00	77.80
-1	81.00	80.01	81.90	**82.80**	80.50	79.10
-2	80.40	81.40	81.10	80.40	80.10	79.30
-3	80.30	80.40	80.60	80.20	80.60	78.20
-4	81.90	80.50	81.30	79.80	81.40	79.50
-5	80.40	79.80	81.40	80.60	79.70	78.90

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 average gain supports an additional contribution from the logits processors in the tested configuration. The gain is not uniform across datasets, and the ablation does not establish that both processors are necessary for every input.

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 a better average result: the full CSRAG model (86.16) outperforms the variant that boosts only the original context (85.92), although gains are not uniform across datasets. The improvement is most evident on FaithEval and NQ, consistent with a benefit from additional lexical coverage; these ablations do not establish changes to internal signal accumulation.

Role of Stopword Filtering. Removing stopword filtering lowers average ACC by 4.07 points (82.09 vs. 86.16). Function words and punctuation occur widely in both knowledge sources, so modulating their token IDs can alter generation beyond the intended factual content. Filtering reduces this interference in the evaluated setting, but does not guarantee fluency: subword IDs may also occur inside content words, and the study does not include an independent fluency assessment.

The full pipeline has the highest average ACC among the tested variants. Context enhancement and logits modulation both contribute to this average, with dataset-dependent effects. Appendix A.2 separately evaluates suppression and boosting; these controlled comparisons provide stronger evidence about component contributions than comparisons between differently trained systems.

4.5 Impact of Key Hyperparameters in CSRAG

The additive intervention is controlled by two scalar hyperparameters: the suppression strength α for the *ConflictingKnowledgeSuppressor* and the boosting strength β for the *ContextualBoostProcessor*. We examine their sensitivity under a fixed backbone and benchmark.

We performed a full 6×6 grid search on the LLaMA-3.1-8B-Instruct model using the FaithEval benchmark, with the results summarized in Table 4. The performance landscape shows that CSRAG achieves its best performance at the default configuration ($\alpha = -1.0, \beta = +3.0$) and exhibits notable stability under moderate variations.

At $\beta = +3.0$, removing suppression lowers ACC, while stronger suppression does not improve on the default. Increasing boosting beyond the default also reduces ACC in this row. These are empirical associations for the tested grid; neither fluency degradation nor internal parametric dominance is measured directly by this experiment.

The integer grid is interpretable in nats because each unit changes relative odds by a factor of e against an unmodified token. The maximum at $(-1, +3)$ lies in a moderately performing neighborhood: changing either coordinate alone by one nat yields 80.2–81.9 ACC, compared with 82.8 at the maximum. The grid shows no abrupt collapse in this local neighborhood, but it does not establish task-independent optimality or remove the need for validation. Appendix A.11 shows a substantially narrower useful range for the tested multiplicative alternative.

4.6 Sensitivity to Stage Prompts

To assess whether the gains depend on prompt wording, we vary one stage's prompt at a time while keeping $\alpha = -1$ and $\beta = 3$ fixed. Table 5 reports the supplementary LLaMA-3.1-8B-Instruct results. Across the tested wording changes, ACC differs from the default by at most 1.2 points on FaithEval and NQ. Removing paraphrasing is a pipeline ablation rather than a wording change: it reduces NQ ACC by 1.6 points, while slightly improving FaithEval. These results indicate limited sensitivity to the tested wording variants and a dataset-dependent benefit from paraphrasing; they do not establish robustness to arbitrary prompts. Exact-match results and further discussion are provided in Appendix A.9.

Table 5: Stage-prompt sensitivity (ACC, %; supplementary H20/vLLM evaluation). Only the indicated stage changes; logits processors remain fixed.

Stage varied	Variant	FaithEval	NQ
—	Default	84.0	90.7
Fact extraction	Concise	84.0	91.1
Fact extraction	Structured	84.2	91.5
Paraphrasing	Concise	85.2	90.5
Paraphrasing	Removed	84.4	89.1
Answer	Context-strict	84.7	91.9

Practical interpretation of the supplementary evidence. The robustness studies favor conditional use of CSRAG. When retrieved passages contain distractors, boosting treats their content tokens as eligible evidence alongside the gold passage; it does not distinguish trustworthy from irrelevant statements. The late-gold failures in Appendix A.10 therefore limit deployment claims even though several gold-first conditions retain an advantage. Similarly, mixed Qwen3 results (Appendix A.12) show that a configuration useful for one backbone can be unnecessary or harmful for another. Validation should assess the intended backbone, retrieval conditions, and answer format together. In particular, selecting a configuration by ACC alone can overlook deteriorating answer precision under EM or F1.

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 mitigates knowledge conflicts by softly suppressing tokens linked to parametric knowledge and boosting those grounded in retrieved evidence. This intervention adjusts token-level competition without model modification; internal accumulation dynamics serve as motivation rather than a mechanism established here. The original six-benchmark evaluation shows competitive accuracy, and supplementary H20/vLLM analyses provide uncertainty estimates and quantify design trade-offs. Gains are backbone- and metric-dependent: Qwen3 results are mixed, and longer answers can reduce EM/F1 despite better answer containment. CSRAG therefore offers an interpretable, training-free option for models prone to parametric over-reliance, with two auxiliary generation calls and no guarantee of universal improvement.

6 Limitations

CSRAG incurs two auxiliary generation calls, although extraction and paraphrasing can run concurrently and query-only facts can be cached. The rebuttal analyses identify two practical trade-offs: late-ranked evidence among distractors can weaken contextual boosting, and longer answers can reduce EM/F1 despite higher answer-containment accuracy. Gains also depend on the backbone, with mixed Qwen3 results. These findings motivate evidence-aware token selection and more precise answer generation. Detailed experiments and other scope restrictions are provided in Appendix E.1.

Acknowledgements

This work is supported by the Key Research and Development Program of Shaanxi Province (Program No. 2024GX-YBXM-006). We thank all the anonymous reviewers for their constructive comments and suggestions, which have been highly valuable in refining this study.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 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, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

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. Context-dpo: Aligning language models for context-faithfulness. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 10280–10300. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.536/.

Steven Bird. NLTK: the natural language toolkit. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *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, 2006. doi: 10.3115/1225403.1225421. URL https://aclanthology.org/P06-4018/.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.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. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation, 2025. URL https://arxiv.org/abs/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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *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, 2025a. URL https://aclanthology.org/2025.acl-long.256/.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. Investigating context faithfulness in large language models: The roles of memory strength and evidence style. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4789–4807. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.findings-acl.247/.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts.

In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6691–6706. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.522. URL https://doi.org/10.18653/v1/2021.acl-long.522.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing GPT-4 on conversational QA and RAG. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1c0d54ebd0a6e58c4eca7d591e374b9d-Abstract-Conference.html.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 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, 2025. URL https://openreview.net/forum?id=UeVx6L59fg.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *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, 2016. doi: 10.18653/V1/D16-1264. URL https://doi.org/10.18653/v1/d16-1264.

Leonardo Ranaldi, Marco Valentino, and André Freitas. Eliciting critical reasoning in retrieval-augmented generation via contrastive explanations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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, 2025. doi: 10.18653/V1/2025.NAACL-LONG.557. URL https://doi.org/10.18653/v1/2025.naacl-long.557.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *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, 2024. doi: 10.18653/V1/2024.NAACL-SHORT.69. URL https://doi.org/10.18653/v1/2024.naacl-short.69.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 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, 2025. URL https://openreview.net/forum?id=ztzZDzgfrh.

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. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://doi.org/10.1162/tacl_a_00475.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 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, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 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, 2025. URL https://openreview.net/forum?id=cPD2hU35x3.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *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, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.486. URL https://doi.org/10.18653/v1/2024.emnlp-main.486.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *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, 2024. doi: 10.18653/V1/2024.ACL-LONG.232. URL https://doi.org/10.18653/v1/2024.acl-long.232.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *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, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.234. URL https://doi.org/10.18653/v1/2024.findings-acl.234.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. A survey of graph retrieval-augmented generation for customized large language models. *CoRR*, abs/2501.13958, 2025a. doi: 10.48550/ARXIV.2501.13958. URL https://doi.org/10.48550/arXiv.2501.13958.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *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, 2025b. URL https://aclanthology.org/2025.acl-long.1062/.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=76cFMRgEzQ.

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

Huachi Zhou, Shuang Zhou, Hao Chen, Ninghao Liu, Fan Yang, and Xiao Huang. 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, 2024.

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

10.18653/V1/2023.FINDINGS-EMNLP.968. URL https://doi.org/10.18653/v1/2023.findings-emnlp.968.

A Additional Experiments

A.1 Effect of Decoding Temperature

Table 6: Effect of decoding temperature on CSRAG performance (LLaMA-3.1-8B-Instruct) on FaithEval (accuracy in %). The best result is highlighted in **bold**.

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

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 6.

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. At high temperature (1.5), performance drops sharply. This comparison establishes sensitivity to decoding temperature, but does not isolate the contribution of the logits processors from the backbone's own temperature sensitivity. The highest measured score occurs under greedy decoding. 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 7 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 support complementary effects of suppression and boosting at the output level in the evaluated configurations; they do not directly measure internal signal propagation or when a conflict is resolved.

In this experiment, the combined setting outperforms both isolated variants on all six benchmarks. This supports the chosen combination for these backbones and settings without establishing that every effective conflict-resolution method requires both operations.

Table 7: Performance comparison of isolated and combined suppression/boosting strategies across six datasets (accuracy in %).

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

A.3 End-to-End Latency Analysis

CSRAG requires two auxiliary LLM calls per query (parametric fact extraction and context paraphrasing) in addition to the final answer generation, yielding three generation calls. In the original K100AI evaluation on MuSiQue with LLaMA-3.1-8B-Instruct and temperature zero, COIECD, CSRAG, and FaithfulRAG took 2h 29m, 13h 40m, and 19h 47m, respectively. The original CSRAG evaluation used a per-sample Hugging Face generation loop without batching; its 13h 40m runtime therefore reflects this evaluation implementation and should not be interpreted as an intrinsic cost of the method. The updated timing table below reports the additional single-H20 experiment rather than repeating the original table.

The structural overhead is two auxiliary generation calls, not a fixed wall-clock slowdown: call counts do not account for sequence lengths, batching, or per-token forward-pass costs. Unlike training-based methods (e.g., ParamMute), CSRAG incurs no training or fine-tuning cost. Fact extraction and context paraphrasing can run concurrently because they condition only on the query and retrieved context, respectively; query-only fact extraction can also be cached across retrieval refreshes. We separately report H20 measurements below, without deriving speedups across hardware platforms.

Additional H20 timing experiment. During rebuttal, we reimplemented CSRAG with vLLM continuous batching and reran the official COIECD implementation on the same single H20 GPU with the same LLaMA-3.1-8B-Instruct checkpoint. Table 8 reports these additional measurements separately from the original K100AI results. Although the hardware and checkpoint are shared, the inference implementations and batching strategies are not controlled to be identical. These results characterize the evaluated implementations, not an isolated algorithmic speedup. The seconds-per-query values are total dataset runtime divided by sample count, not individual request latency under batching.

Table 8: Additional end-to-end timings on a single H20 GPU. COIECD uses its official implementation; CSRAG uses a vLLM reimplementation with continuous batching.

Benchmark	#Samples	COIECD	CSRAG	CSRAG s/query
FaithEval	1,000	1h 15m 37s	295s	0.30
SQuAD	1,769	2h 14m 58s	417s	0.24
MuSiQue	1,772	2h 18m 03s	517s	0.29
NQ	1,274	1h 37m 18s	233s	0.18
HotpotQA	1,407	1h 48m 39s	272s	0.19
NewsQA	870	1h 07m 41s	243s	0.28
Total	8,092	10h 22m 16s	1,977s	—

A.4 Exact-Token vs. Embedding-Based Suppression

A natural question is whether the ConflictingKnowledgeSuppressor would benefit from a semantically "soft" penalty that covers tokens related to—but not identical to—the extracted parametric facts. To investigate this, we implement a preliminary embedding-based suppressor: each extracted parametric fact is represented by the mean-pooled input embedding of its tokens; at each decoding step, we compute cosine similarity between this fact vector and every vocabulary token embedding, and subtract a scaled penalty from the logits of tokens whose similarity exceeds a threshold (taking the max similarity over all facts per token).

On FaithEval with LLaMA-3.1-8B-Instruct, this soft variant achieves 80.50 ACC compared to 82.80 ACC for the original exact-token CSRAG—a 2.30 point decline. We hypothesize that input-space embedding similarity is a noisy proxy for generation-level token relevance, whereas exact token matching directly from extracted atomic facts provides a more precise and actionable suppression signal. We report this as a negative result and view principled conflict-aware semantic suppression as a promising direction for future work.

A.5 Supplementary Evaluation Protocol

The following supplementary analyses were conducted during rebuttal on H20 GPUs using a vLLM reimplementation of CSRAG. Unless explicitly identified as Qwen3, the backbone is LLaMA-3.1-8B-Instruct. Throughout the paper, experiments without an explicit platform note were run on K100AI with Hugging Face generation, and tables or subsections marked H20/vLLM use the supplementary setting; neither accuracy values nor uncertainty estimates are pooled across the two settings. The supplementary standard-RAG baseline uses the same CoT answer prompt and raw context as the ablation's no-control row, without the main table's simpler direct-answer baseline prompt; it is therefore systematically higher than the main-table full-context row and is not comparable to it. Default scores differ between the separately reported analyses; each comparison below uses its own accompanying baseline rather than treating all default rows as a single run. ACC is answer containment as defined in Section 4.1; supplementary exact match (EM) and answer F1 assess stricter answer agreement.

A.6 Statistical Significance and Sampling Variance

We compare CSRAG with standard RAG on identical samples using paired bootstrap tests with 10,000 resamples. Table 9 gives ACC and 95% confidence intervals for the difference. Four of six comparisons have reported $p < 0.05$. NewsQA is borderline at the reported precision ($p = 0.050$), and HotpotQA is not significant. These tests support the supplementary comparisons, not significance claims for the original K100AI table or against FaithfulRAG and ParamMute, for which paired tests are unavailable. Confidence intervals and p-values are reproduced at their reported precision; a positive rounded lower bound on NewsQA is not used to override its borderline reported p-value.

Table 9: Supplementary H20/vLLM ACC comparisons (%). Confidence intervals are for the gain in percentage points. The gain column is computed from the displayed scores.

Benchmark	CSRAG	Standard RAG	Gain	95% CI	p
SQuAD	84.96	79.42	5.54	$[4.0, 7.1]$	$< .001$
MuSiQue	83.01	78.56	4.45	$[2.5, 6.4]$	$< .001$
NQ	90.74	87.99	2.75	$[1.2, 4.3]$	.001
FaithEval	84.00	81.80	2.20	$[0.1, 4.2]$	.042
NewsQA	93.10	91.15	1.95	$[0.1, 3.8]$	.050
HotpotQA	92.75	91.90	0.85	$[-0.6, 2.3]$	.266

All three stages use temperature zero. For fixed inputs in the evaluated implementation, changing the random seed did not change fact extraction. The measured seed variability instead concerns distractor sampling: across 45 runs using seeds 41, 42, and 43, the ACC standard deviation was at most 0.81 percentage points in each evaluated configuration. This is not an estimate of cross-hardware reproducibility or uncertainty for every main-table entry. Per-condition standard deviations for the retrieval-noise study are reported in Table 15; bootstrap uncertainty and distractor-sampling variability quantify different sources of uncertainty.

ACC improvements do not imply improvements under stricter metrics. EM and F1 decline on MuSiQue and HotpotQA (Table 10). In the supplementary output analysis, boosting lengthened responses by 1.9–3.0 times, and 48–57% of EM failures still contained the gold answer span. This pattern is consistent with a verbosity trade-off: answer-containment accuracy can improve even when concise answer matching deteriorates.

Table 10: Supplementary H20/vLLM differences from standard RAG in EM and F1 (percentage points), with paired-bootstrap p-values.

Benchmark	ΔEM	p (EM)	ΔF1	p (F1)
FaithEval	$+1.60$	.146	$+0.63$	.453
SQuAD	$+3.78$	$< .001$	$+2.50$	.001
MuSiQue	-5.42	$< .001$	-6.04	$< .001$
NQ	$+0.71$	.483	$+0.34$	.676
HotpotQA	-2.56	.005	-2.20	.007
NewsQA	-0.23	.875	$+0.01$	.993

A.7 Fact Extraction Failure Rate

We define extraction failure as a refusal or an empty set of actionable facts. Table 11 quantifies this behavior on the supplementary H20/vLLM implementation. The count measures actionable statements, not whether those statements are true or faithfully represent internal model knowledge.

Table 11: Supplementary fact extraction statistics. The two SQuAD variants and the two MuSiQue variants are grouped as in the evaluation report.

Benchmark	Mean actionable facts	Failure rate
FaithEval	10.01	0.00%
NQ	9.98	0.00%
HotpotQA	9.99	0.00%
SQuAD / SQuAD-golden	9.96	0.23%
MuSiQue / MuSiQue-golden	9.83 / 10.03	0.23%
NewsQA	9.15	4.71%

An empty fact set disables suppression, leaving context-boosted RAG rather than preventing answer generation. As a separate check of this reduced system, the supplementary $\alpha = 0$ ablation reaches 82.0 ACC on FaithEval versus 81.8 for standard RAG. This does not guarantee baseline-level performance on the subset of failed extractions, or when nonempty but poor-quality facts are extracted. NewsQA has the highest measured failure rate (4.71%).

A.8 Lexical Overlap and Shared-Token Policies

For shared tokens, the default rule applies both offsets, giving $\alpha + \beta$. The offline overlap analysis uses lowercased content-word sets as a tokenizer-free proxy, rather than the tokenizer-ID sets used during decoding. We report their Jaccard similarity and separately compare four policies for shared token IDs in the H20/vLLM evaluation. Lexical overlap occurs in most samples, but its relative magnitude is small: mean Jaccard ranges from 0.035 to 0.155, and at most 0.07% of samples exceed 0.5.

Table 12: Supplementary content-word overlap statistics (a lexical proxy, not tokenizer-ID overlap).

Benchmark	Mean Jaccard	Zero overlap	Jaccard > 0.5
FaithEval	0.086	0.00%	0.00%
SQuAD	0.086	0.40%	0.00%
MuSiQue	0.086	0.23%	0.00%
NQ	0.119	0.24%	0.00%
HotpotQA	0.155	0.00%	0.07%
NewsQA	0.035	4.02%	0.00%

The within-benchmark ranges are 1.6, 1.6, and 1.9 ACC points, respectively. No policy is best on all three benchmarks: neutral is best on FaithEval, whereas parametric-only treatment is best on SQuAD and MuSiQue. Thus the default additive treatment is a simple convention, not an empirically optimal or necessary choice.

Table 13: Shared-token policy comparison (ACC, %; supplementary H20/vLLM evaluation). Tokens outside the intersection retain their usual treatment.

Policy	Shared-token offset	FaithEval	SQuAD	MuSiQue
net (default)	$\alpha + \beta$	84.0	85.1	83.5
context	β	82.8	84.6	83.1
parametric	α	84.0	86.2	85.0
neutral	0	84.4	85.9	84.4

A.9 Prompt Sensitivity

We vary one prompt at a time while holding the logits processors at $\alpha = -1$ and $\beta = 3$. Table 14 reports EM and ACC separately to avoid conflating exact matching with answer containment.

Table 14: Prompt sensitivity (EM / ACC, %; supplementary H20/vLLM evaluation). Removing paraphrasing is a stage ablation rather than a wording change.

Stage varied	Variant	FaithEval EM / ACC	NQ EM / ACC
—	default	82.8 / 84.0	84.4 / 90.7
Fact extraction	concise	83.1 / 84.0	83.8 / 91.1
Fact extraction	structured	82.9 / 84.2	84.7 / 91.5
Paraphrasing	concise	84.4 / 85.2	83.4 / 90.5
Paraphrasing	removed	83.7 / 84.4	81.2 / 89.1
Answer	context-strict	84.1 / 84.7	84.6 / 91.9

Among the tested wording changes, ACC differs from the default by at most 1.2 points. Removing paraphrasing reduces NQ ACC by 1.6 points and EM by 3.2 points, although it slightly improves FaithEval; the value of this stage is therefore dataset-dependent. As a separate check, the $\alpha = 0$, $\beta = 0$ ablation with the default answer prompt reaches 81.9 ACC on FaithEval, below both the default full pipeline (84.0) and the context-strict full pipeline (84.7) in Table 14. These comparisons suggest that prompt wording alone does not explain the observed gains; they do not establish robustness to arbitrary prompts.

A.10 Robustness under Retrieval Noise

We mix one gold passage with K sampled distractors and place the gold passage at rank R. Seeds 41, 42, and 43 control distractor sampling. Standard RAG is evaluated under identical noise. Table 15 reports CSRAG means and standard deviations; the standard-RAG columns retain the reported point values because corresponding standard deviations are unavailable.

Table 15: Retrieval-noise evaluation (ACC, %; H20/vLLM). CSRAG noisy-condition values are mean $\pm$ standard deviation over three distractor seeds. Clean rows are point estimates.

	HotpotQA		NQ		NewsQA	
Condition	CSRAG	Standard RAG	CSRAG	Standard RAG	CSRAG	Standard RAG
clean	92.75	91.90	90.74	87.99	93.10	91.15
$K = 1$, gold@1	92.06 ± 0.41	92.11	91.10 ± 0.16	89.32	91.99 ± 0.81	90.23
$K = 3$, gold@1	91.35 ± 0.51	91.19	90.92 ± 0.18	88.23	90.34 ± 0.41	90.11
$K = 5$, gold@1	91.73 ± 0.74	90.76	91.23 ± 0.05	88.78	90.23 ± 0.34	89.77
$K = 2$, gold@3	90.76 ± 0.12	90.48	89.53 ± 0.18	89.01	86.36 ± 0.33	89.31
$K = 4$, gold@5	90.97 ± 0.25	89.98	88.88 ± 0.32	89.25	83.98 ± 0.46	88.85

With the gold passage first, increasing distractors from one to five changes CSRAG ACC by -0.33, $+0.13$, and -1.76 points on HotpotQA, NQ, and NewsQA, respectively. CSRAG exceeds the reported standard-RAG value in eight of these nine gold-first comparisons; the exception is HotpotQA with one distractor (92.06 versus 92.11). Accuracy is not monotonically decreasing with distractor count on every dataset: HotpotQA and NQ fluctuate, whereas NewsQA declines across the tested gold-first settings.

Late-gold settings expose a more substantial failure. On NewsQA with gold@5, CSRAG reaches 83.98 versus 88.85 for standard RAG, a 4.87-point deficit. Because the boost set is built from the concatenated context, long distractor passages can dominate the set; this is a plausible explanation, not a directly established causal mechanism. The late-gold conditions also change distractor count, so they do not isolate a pure position effect. Evidence-aware passage selection or filtering for boost-set construction is future work; simply favoring top-ranked passages is not guaranteed to help when the gold passage is ranked late.

A.11 Additive versus Multiplicative Logits Processors

Let z_i be a token logit and $\delta_i = \alpha\mathbb{I}(i \in T_p) + \beta\mathbb{I}(i \in T_c)$. In the supplementary H20/vLLM study, we compare the default additive rule $z'_i = z_i + \delta_i$, a multiplicative rule $z'_i = z_i(1 + \delta_i)$, and an absolute-value-scaled rule $z'_i = z_i + \delta_i|z_i|$. Equal parameter values do not imply equal interventions: the latter two changes depend on logit magnitude, and the multiplicative rule also depends on its sign. For example, $z = -4, \delta = -1$ gives -5 additively but 0 multiplicatively, raising rather than suppressing that logit; $z = -0.5, \delta = 3$ gives 2.5 additively but -2 multiplicatively. Raw-logit scaling also depends on the arbitrary common logit offset, unlike additive offsets applied before softmax.

At the additive defaults $(-1, 3)$, multiplicative EM falls to 0.0, 0.1, and 0.0 on FaithEval, SQuAD, and MuSiQue, compared with additive EM of 82.4, 81.9, and 69.4 in this study; these collapsed scores reflect the sign-inversion failure illustrated above rather than a tuned configuration. A 21-configuration FaithEval search within the supplementary 31-run study instead selects $(-0.1, 0.1)$ for the scaled variants. Table 16 shows that recalibration makes them competitive, but the preferred rule depends on the metric: both scaled variants improve EM on all three benchmarks, whereas ACC improves only on FaithEval and decreases on the other two.

Table 16: Logits-rule comparison (EM / ACC, %; H20/vLLM). These are the rule-comparison study's scores, separate from the bootstrap comparison.

Rule	(α, β)	FaithEval	SQuAD	MuSiQue
additive	$(-1, 3)$	82.4 / 83.3	81.9 / 84.9	69.4 / 83.6
multiplicative	$(-0.1, 0.1)$	83.2 / 84.1	82.7 / 84.7	74.6 / 82.6
abs-scaled	$(-0.1, 0.1)$	83.2 / 84.1	82.8 / 84.8	75.6 / 83.2

Table 17: Multiplicative sensitivity at $\alpha = -0.1$ (FaithEval EM, %; H20/vLLM).

β	0.1	0.2	0.3	0.5	1.0
EM	83.2	82.1	74.3	5.9	0.0

The useful tested multiplicative range is narrow, around $\beta = 0.1$–0.2, with severe degradation by 0.5. Its dependence on logit scale motivates backbone-specific calibration. Additive offsets have a fixed log-odds interpretation, although their best task-level settings need not be universal. They can in principle be expressed through an OpenAI-style logit-bias interface when the API exposes the relevant tokenizer IDs and supports the required bias-map size and values. Multiplicative variants require access to raw logits. We have not validated closed-source API deployment here. We retain addition for interpretability and implementation simplicity, not because it is the only valid or uniformly best rule.

A.12 Results on Qwen3 Models

We additionally evaluate Qwen3 models using the supplementary H20/vLLM setup, comparing each backbone with and without CSRAG on identical samples. Table 18 covers model sizes from 1.7B to 32B on FaithEval; Table 19 reports all eight evaluated benchmarks for Qwen3-8B.

Qwen3-8B has four gains and four losses, all within 1.7 ACC points; these mixed results do not establish either universal improvement or statistical equivalence. Its no-CSRAG conflict scores are higher than the supplementary LLaMA-3.1-8B baseline on SQuAD (86.43 versus 79.42) and

Table 18: Qwen3-family comparison on FaithEval (ACC, %; H20/vLLM).

Backbone	CSRAG	Without CSRAG	Difference
Qwen3-1.7B	81.2	80.5	+0.7
Qwen3-8B	74.2	75.9	−1.7
Qwen3-14B	72.8	73.4	−0.6
Qwen3-32B	68.3	69.6	−1.3

Table 19: Qwen3-8B across eight benchmarks (ACC, %; H20/vLLM).

Benchmark	CSRAG	Without CSRAG	Difference
FaithEval	74.20	75.90	−1.70
SQuAD	86.26	86.43	−0.17
MuSiQue	88.15	87.08	+1.07
SQuAD-golden	98.81	98.76	+0.05
MuSiQue-golden	93.45	92.78	+0.67
HotpotQA	93.96	95.02	−1.06
NQ	91.92	91.21	+0.71
NewsQA	92.07	93.10	−1.03

MuSiQue (87.08 versus 78.56). Across the 14 evaluated LLaMA-3.1-8B/Qwen3-8B benchmark pairs, baseline ACC and CSRAG gain have a reported correlation of $r = -0.44$, consistent with, but not proving, a headroom explanation. We therefore position CSRAG as a training-free intervention for backbones prone to over-relying on parametric memory, not a guaranteed improvement for newer or larger models. Stage 1 requires instruction following; non-instruction-tuned base models are outside the evaluated scope.

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, with benchmark-provided target answers used for evaluation. 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, where the supplied context may contradict real-world knowledge and the evaluation tests whether the model follows that context. 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. 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.

Table 20: Overview of datasets used in our experiments.

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

B.2 CSRAG Implementation and Details

Hardware and inference implementations. The original submission experiments were conducted on K100AI, with CSRAG implemented using Hugging Face generation. Unless a table or subsection explicitly states otherwise, all experiments in this paper, including the main table and the remaining appendix tables, were run on K100AI with Hugging Face generation. We retain the original accuracy results rather than replacing them with measurements from a different inference stack. Additional experiments conducted during rebuttal used H20 GPUs and a vLLM reimplementation of CSRAG with continuous batching; the additional timing comparison in Table 8 used a single H20 GPU. H20 results are reported separately and are not pooled with the original K100AI measurements. Differences between the two settings may reflect both hardware and inference implementation, so we do not infer cross-platform speedups or assume numerically identical outputs.

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. The prompt requests preservation of the original semantics while increasing lexical diversity; this requirement is not enforced by a factual-equivalence verifier. The paraphrased statements are appended to the original retrieved context rather than replacing it, preserving the original evidence alongside the generated rewrites.

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 use greedy decoding and limit sampling variability within each evaluation setting. No nucleus or top-k sampling is applied. CSRAG is implemented as a 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.

Table 21: CSRAG successfully resolved a knowledge conflict in Case 1 (musique_5633c1).

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"

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. Consequently, lexical variants may be missed and a shared subword ID may be affected in unrelated words; token-set membership does not establish a semantic contradiction.

Computational Overhead. CSRAG introduces additional inference cost. 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. Wall-clock cost depends on the inference implementation, batching, and hardware, as discussed in Section A.3. No additional training, gradient computation, or iterative decoding is required.

Prompt Specification. The default prompts used for parametric fact extraction and context paraphrasing are fixed and shared across datasets and backbones, except in the explicitly identified prompt-sensitivity experiments (Appendix A.9). 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 21 and 22 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 with the displayed baseline outputs, CSRAG selects the context-supported target answers in these cases. This illustrates contextual answer selection, not a guarantee that the accompanying reasoning is valid or that the counterfactual content is true in the real world.

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

Table 22: CSRAG successfully resolved a knowledge conflict in Case 2 (musique_904144).

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 ("In 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"

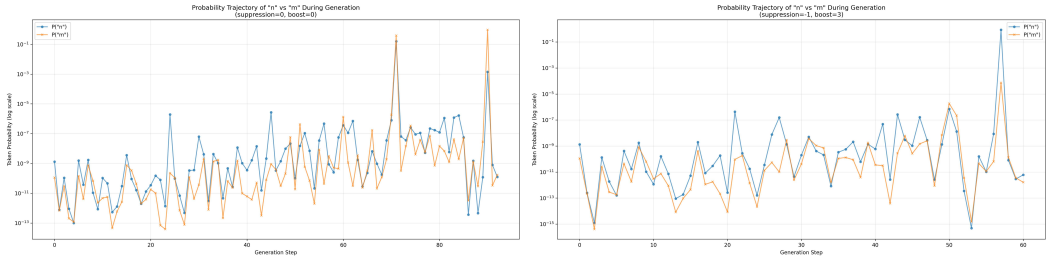

Figure 3: Token-level probability comparison between standard decoding and CSRAG (id: squad_a32214). CSRAG suppresses incorrect parametric tokens and amplifies context-grounded tokens during generation. Under standard decoding, the model assigns higher probability to m than to n, reflecting dominance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

parametric 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 displayed probabilities are consistent with the intended output-level preference shift. The two generations have different histories and the shown steps are not aligned, so this example is not a controlled comparison at an identical decoder state and does not establish a change to internal knowledge pathways. 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. The cases illustrate limitations of different pipeline components; manual inspection does not establish the prevalence or causal origin of each error type.

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." An actually empty fact set makes suppression a no-op while leaving paraphrasing and boosting active. However, the parser can retain nonempty refusal text as a fallback, so a refusal may still contribute token IDs despite containing no actionable facts. The offline failure statistic therefore does not by itself imply that the suppression set is empty. 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 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.

D Prompt Design

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.

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.

This appendix presents the prompt templates used in CSRAG, including those for parametric fact extraction, context paraphrasing, and final response generation. As illustrated in Figures 4, 5, and 6, the default prompts are fixed and shared across datasets and backbone models; Appendix A.9 explicitly varies individual stages for sensitivity analysis. 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.

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.

E Limitations and Ethical Statement

E.1 Limitations and Future Work

While CSRAG effectively mitigates knowledge conflicts in retrieval-augmented generation without requiring parameter updates, it has certain limitations that point to directions for future work:

Dependency on Parametric Fact Extraction. Suppression depends on surfacing actionable candidate facts, which need not be accurate or a faithful account of the model's internal knowledge. Supplementary extraction failure rates range from 0.00% to 4.71% (Appendix A.7). Empty sets disable suppression, while nonempty but irrelevant facts may still misdirect it; successful extraction alone does not establish robustness. Qualitative failures are discussed in Appendix C.3. More reliable extraction and validation remain future work.

Assumption of Reliable Context. CSRAG targets the generator's faithfulness under the assumption that the retrieved context is reasonably reliable. In adversarial settings (e.g., poisoned RAG or prompt injection), blindly boosting contextual signals could amplify malicious information. Therefore, CSRAG should be viewed as complementary to robust retrieval validation and filtering mechanisms.

Retrieval Noise and Late Gold Passages. In the supplementary NewsQA gold@5 condition, CSRAG obtains 83.98 ACC versus 88.85 for standard RAG, a 4.87-point deficit (Appendix A.10). A boost set derived from concatenated passages can be dominated by distractors. The tested conditions vary both rank and distractor count, so they do not isolate position as the sole cause. Evidence-aware filtering or selection for boost-set construction remains unvalidated future work.

Metric Dependence and Answer Verbosity. Better answer-containment ACC does not necessarily mean better exact answers. In the supplementary comparison, MuSiQue EM/F1 decrease by 5.42/6.04 points and HotpotQA EM/F1 by 2.56/2.20 points (Appendix A.6). Boosting is associated with 1.9–3.0

times longer outputs; 48–57% of EM failures still contain the gold span. Length-aware decoding and more precise answer extraction merit further study.

Backbone Dependence. Qwen3-8B shows four gains and four losses, each within 1.7 ACC points, across eight benchmarks (Appendix A.12). A negative correlation between baseline accuracy and gain is consistent with limited headroom, but does not prove a causal explanation. We do not claim universal improvement or statistical equivalence on Qwen3. The extraction stage requires instruction following, and non-instruction-tuned models are outside the evaluated scope.

Inference Overhead. CSRAG requires two auxiliary generation calls in addition to answer generation. These calls may be parallelized, and query-dependent extraction may be cached, but they are not free. Section A.3 separates original K100AI timings from H20 results; hardware, batching, and implementation differences prevent interpreting these measurements as intrinsic algorithmic speedups.

Scope of Evaluation. Our experiments are conducted exclusively in English. The original evaluation primarily covers 7B–8B models, while supplementary FaithEval experiments extend to Qwen3 models from 1.7B to 32B. Multilingual behavior, substantially larger or compressed models, and closed-source API deployment remain unvalidated. Supplementary uncertainty estimates belong to the H20/vLLM evaluation and must not be transferred to the original K100AI tables.

E.2 Ethical Statement

This work complies with NeurIPS ethical requirements. All experiments use publicly available academic benchmarks that contain no private or sensitive data. CSRAG requires no fine-tuning and uses the public benchmarks described above. Improved contextual faithfulness may reduce unsupported answers, but following retrieved context does not establish real-world truth. Boosting false, biased, or malicious evidence can amplify misinformation, and generated paraphrases can introduce further errors. Applications therefore require evidence validation and the safeguards appropriate to the underlying model and data. The auxiliary generation stages also increase inference compute and associated resource use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe a training-free, output-level intervention motivated by prior mechanistic findings, not a newly established internal mechanism. The contribution is qualified by mixed Qwen3 results, retrieval-noise failures, and metric-dependent gains (Appendices A.12, A.10, and A.6).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 states auxiliary-call cost, retrieval-noise failures, answer verbosity and metric dependence, and mixed backbone results. Appendix E.1 expands on extraction quality, retrieval reliability, and language/model scope, with supporting supplementary measurements and qualitative failures in Appendix C.3.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper presents an empirical method (CSRAG) and does not include theoretical results such as theorems or formal proofs. The only mathematical content is the logits adjustment formula in Section 3.3.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 specifies datasets, baselines, evaluation metrics, hyperparameters ($\alpha = -1.0$, $\beta = +3.0$), and temperature-zero decoding. Appendix B provides dataset details, preprocessing steps, and implementation specifics; Appendix D contains the default prompts. The public repository also provides the CSRAG vLLM implementation, processed evaluation data, and inference and evaluation commands; its release scope is described in Appendix **??**. Supplementary comparisons and their uncertainty are distinguished from the original results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and processed evaluation data are publicly available at `https://github.com/YeLyea/CSRAG`, with installation, inference, and ACC evaluation commands. The release covers the CSRAG vLLM pipeline, but does not include the baseline implementations or complete historical configurations for the original Hugging Face experiments. We therefore answer No to the full reproduction requirement, rather than to public availability itself (Appendix **??**).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 4.1 and Appendix B describe datasets (sample counts in Table 20), original backbones, hyperparameter selection, and temperature-zero decoding. Appendix A.5 distinguishes the supplementary evaluation, with distractor seeds and bootstrap settings in Appendix A.6 and Qwen3 results in Appendix A.12.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix A.6 reports supplementary paired-bootstrap ACC gains, 95% confidence intervals, and p-values with 10,000 resamples, plus EM/F1 tests. Appendix A.10 reports CSRAG means and standard deviations over distractor seeds 41, 42, and 43. These estimates apply only to the H20/vLLM evaluation, not the original K100AI table. Tests are against standard RAG, not FaithfulRAG or ParamMute, and no significance claim is transferred to untested comparisons.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix B distinguishes the original K100AI/Hugging Face experiments from additional H20/vLLM experiments. Section A.3 reports original runtimes and a separate single-H20 timing comparison, including implementation and batching caveats. CSRAG requires two auxiliary LLM calls per query plus final generation. The public release pins its Python dependencies, but exact accelerator memory, complete historical software versions, and full execution configurations for all reported experiments are not provided.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper includes an "Ethical Statement" that confirms conformance with NeurIPS ethical requirements. The experiments use public academic benchmarks; the Ethical Statement discusses misinformation risks, evidence validation, and additional inference compute.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Ethical Statement discusses potential benefits from reducing unsupported answers, risks from amplifying false or malicious context and erroneous paraphrases, and the resource cost of auxiliary generation. Appendix E.1 discusses dependence on evidence reliability.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work uses existing open-source LLMs and academic benchmarks and does not release a newly trained model or a high-risk scraped dataset. Risks from amplifying unreliable or malicious retrieved evidence are discussed in Appendix E.1.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The source datasets and models are credited through citations. The released project code has an MIT license, and the README states that third-party assets retain their own terms; however, individual licenses and terms for all reused datasets and models are not explicitly listed. The code license alone does not satisfy this requirement.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The public repository at `https://github.com/YeLyea/CSRAG` includes the CSRAG vLLM implementation, eight processed evaluation files, an MIT license for the project code, pinned dependencies, and a README documenting installation, data format, inference, evaluation, and release limitations (Appendix **??**). No newly trained model checkpoints are introduced.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects research is conducted in this paper.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used in CSRAG's core method for parametric fact extraction, context paraphrasing, and answer generation (Section 3); Appendix D provides default prompts. Appendix B.1 also discloses GLM-4-Flash usage for constructing distractor answer options.