Context-Fidelity Boosting: Enhancing Faithful Generation through Watermark-Inspired Decoding

Anonymous ACL submission

Abstract

Large language models frequently generate unfaithful content that deviates from given contexts, a phenomenon known as *faithfulness hallucination*. Existing mitigation methods often require model retraining, architectural modifications, or manipulation of the entire output distribution, leading to significant computational overhead. In this paper, we propose Context-Fidelity Boosting (CFB), a lightweight decoding-time approach that enhances contextual alignment through strategic logit adjustments. Inspired by watermarking techniques, CFB implements three progressively sophisticated strategies: *static boosting* with fixed parameters, *global adaptive boosting* based on distribution divergence, and *token-wise adaptive boosting* that leverages attention patterns and semantic relevance. Extensive experiments demonstrate that CFB significantly improves both faithfulness metrics and generation quality while maintaining computational efficiency. Notably, CFB provides a practical solution for improving context fidelity without requiring model retraining or architectural changes. Our code is released at https://anonymous.4open.science/r/CFB-C716.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language tasks. In numerous scenarios, the model needs to follow the context provided by the user to generate responses, such as in RAG, summarization (Laban et al., 2024), question answering (Chen et al., 2025), and role-playing (Huang et al., 2024). When external knowledge conflicts with the model's internal knowledge parameters, the generated content may become inconsistent with the user's instructions or contextual information (Mallen et al., 2023; Liu et al., 2024c), resulting in faithfulness hallucinations (Huang et al., 2023).

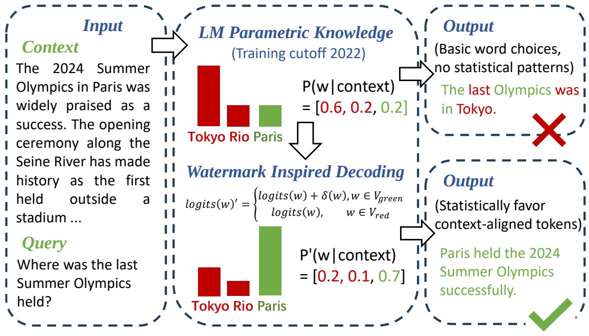


Figure 1: Illustration of context-faithful decoding: Traditional decoding relies on parametric knowledge (favoring "Tokyo"), while our watermarking-inspired approach adjusts token probabilities to align with the given context about "Paris 2024".

This issue is particularly concerning in high-stakes domains such as healthcare (Zhu et al., 2024), legal (Cui et al., 2024), and financial services (Lee et al., 2025), where accurate interpretation of medical records, legal documents, or financial reports is crucial. In these scenarios, models must prioritize faithfulness to the given context over their potentially outdated or incorrect parametric knowledge.

Current approaches to addressing this challenge broadly fall into three categories: (1) training-time methods requiring expensive model fine-tuning or architectural modifications (Hu et al., 2024), (2) prompting techniques relying on careful engineering but offering limited reliability (Zhang et al., 2024), and (3) decoding-time methods that modify the generation process (Shi et al., 2024; Wang et al., 2024). While decoding-time approaches show promise through their model-agnostic nature and computational efficiency, existing methods often face a challenging trade-off between context fidelity and output fluency, or require complex calibration procedures.

In this work, we draw inspiration from recent advances in text watermarking (Kirchenbauer et al., 2024; Liu et al., 2024a; Liu and Bu, 2024), where subtle modifications to token probabilities can effectively guide model behavior without compromising generation quality. As illustrated in Figure 1, similar to how watermarking techniques modify logit distributions to embed signals, we propose to adjust token probabilities to favor context-aligned information. Just as watermarking uses green lists to boost specific token probabilities, our approach identifies and boosts context-relevant tokens while maintaining the natural flow of language generation. This parallel between watermarking's token manipulation and context-faithful decoding provides an elegant framework for addressing the faithfulness challenge.

We introduce Context-Faithful Boosting (**CFB**), a novel decoding-time approach that dynamically adjusts token probabilities based on their contextual relevance. CFB operates through three increasingly sophisticated strategies: *static boosting* with fixed parameters, *global adaptive boosting* based on distribution divergence, and *token-wise adaptive boosting* leveraging attention patterns and semantic relevance. This mechanism enables flexible control over the fidelity-fluency trade-off without requiring model modifications or additional training. Notably, our method achieves this through lightweight computation during decoding, making it practical for real-world applications where trustworthiness and reliability are paramount.

Our key contributions include:

- A lightweight, model-agnostic decoding framework that significantly improves context fidelity while preserving output quality, particularly crucial for high-stakes applications.
- A novel three-level boosting mechanism that automatically calibrates to different contexts and tasks, ensuring reliable performance across diverse domains.
- Extensive empirical validation across multiple model scales and diverse tasks, including summarization and question answering that require high context faithfulness.

2 Related Work

2.1 Faithfulness Hallucinations in LLMs

Despite their impressive capabilities, LLMs frequently generate unfaithful content that deviates from provided context or source documents (Hase et al., 2024; Chuang et al., 2024; Ming et al., 2024). Recent studies have identified two types of hallucinations: factuality hallucination (Yang et al., 2024) manifests when LLM outputs diverge from verifiable real-world facts (e.g., stating incorrect historical dates or attributing quotes to wrong authors), while faithfulness hallucination (Wu et al., 2024; Qiu et al., 2024) occurs when outputs contradict or fabricate content from the given input context (e.g., including details in a summary that were never present in the source document). This issue becomes particularly severe when models encounter information that conflicts with their parametric knowledge learned from training data, such as recent events or domain-specific knowledge. Various metrics have been proposed to measure faithfulness, including semantic similarity scores, entailment-based measures, and fact-checking frameworks (Niu et al., 2024; Hong et al., 2024).

2.2 Existing Mitigation Methods

Prior research has explored diverse approaches to mitigate hallucinations in LLMs, operating at different stages of the model pipeline (Huang et al., 2023). Training-time methods focus on architectural changes and objective refinements, such as enhanced attention mechanisms and knowledge graph integration, though these often require substantial computational resources and may face cross-domain generalization challenges (Tonmoy et al., 2024). Prompting techniques, including chain-of-thought (Wei et al., 2023) reasoning and self-consistency verification, offer model-agnostic solutions but vary in effectiveness across different models and tasks (Hou et al., 2024). Decoding-time interventions modify the generation process through methods like constrained decoding, though they often struggle to balance faithfulness with output fluency (Gema et al., 2024). While each approach presents unique advantages, they all face distinct limitations that must be considered in practical applications, highlighting the ongoing challenge of developing reliable and faithful LLMs.

2.3 Watermarking in LLMs

Recent work on text watermarking has advanced our understanding of how subtle probability modifications can effectively control model outputs in LLMs. These techniques have primarily focused on partitioning the vocabulary into "green" and "red" token lists, carefully adjusting logit distributions to embed detectable statistical patterns while preserving the overall quality of generated text (Liu et al., 2024b). Key developments in this field have included soft watermarking schemes that dynam-

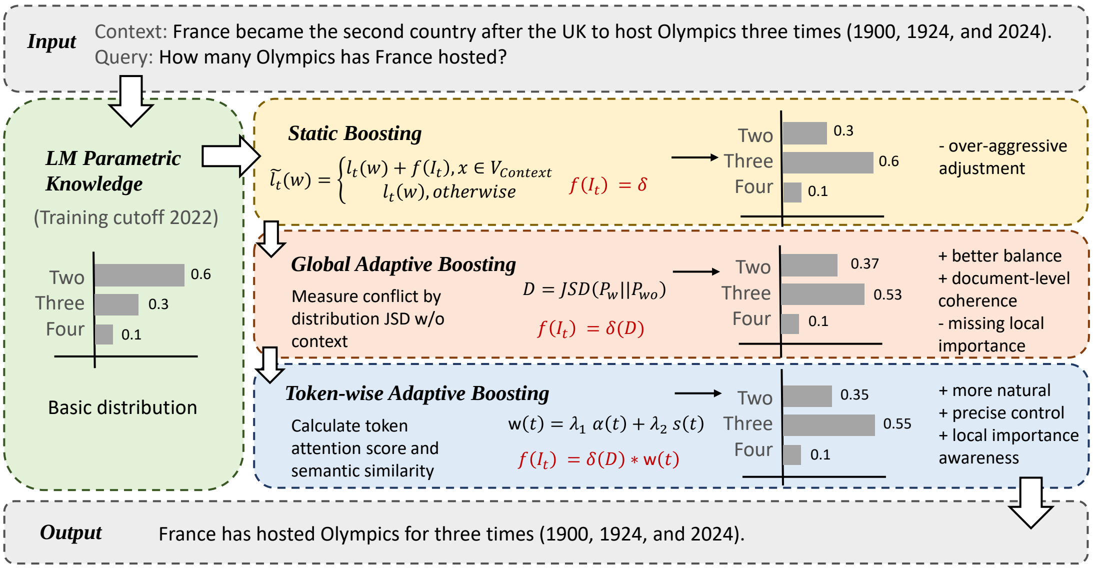


Figure 2: An overview of the proposed CFB method. Our method includes three strategies: static boosting with fixed parameters (directly adjusting the model's logits output), global adaptive boosting based on distribution divergence (determining delta based on JSD divergence), and token-wise adaptive boosting leveraging attention patterns and semantic relevance.

ically adjust token probabilities based on context (Kirchenbauer et al., 2024), sophisticated methods for maintaining generation quality while embedding robust signals (Liu et al., 2024a), and theoretical frameworks that analyze the critical trade-off between watermark strength and text naturalness (Golowich and Moitra, 2024). This controlled manipulation of token distributions suggests a promising direction for hallucination mitigation, as similar probability adjustment techniques could be applied to guide model outputs toward greater faithfulness to source content while maintaining natural language generation capabilities.

3 Methodology

We introduce Context-Fidelity Boosting (CFB), a decoding-time approach that enhances language models' faithfulness to given contexts by adaptively adjusting token probabilities during generation. Inspired by watermarking techniques that successfully control model outputs through subtle probability modifications, CFB implements a hierarchical boosting framework that promotes the selection of context-relevant tokens while maintaining natural text generation, as illustrated in Figure 2.

3.1 Problem Formulation

Given a context passage C and a query Q, our goal is to enhance the generation fidelity of the model to the context during decoding by increasing the probability of tokens that appear in C. Let $P(y_t|y_{<t}, C, Q)$ denote the model's generation probability at timestep t. The key challenge is to ensure the generated sequence maintains higher probabilities for contextual tokens while preserving natural and fluent generation.

Traditional decoding methods treat all vocabulary tokens equally, which may lead to context neglect and hallucination. We propose to adjust the logits of context tokens before computing generation probabilities:

$$\tilde{l}_t(w) = \begin{cases} l_t(w) + f(I_t), & \text{if } w \text{ appears in } C; \\ l_t(w), & \text{otherwise.} \end{cases} \tag{1}$$

Here, $l_t(w)$ is the original logit for token w in the vocabulary, $f(I_t)$ is a boosting function based on importance measure I_t, and $\tilde{l}_t(w)$ is the adjusted logit corresponding to token w.

3.2 Context-Fidelity Boosting Framework

In this section, we propose three progressive levels of boosting strategies for context tokens.

3.2.1 Static Boosting

The most straightforward approach adopts a fixed boosting value δ for all tokens that appear in the context C:

$$f(I_t) = \delta, \tag{2}$$

where δ is preset manually.

This strategy provides a baseline enhancement of context token probabilities but lacks adaptivity to different contexts and token importance.

Algorithm 1: Context-Fidelity Boosting via Logit Adjustment

Input: Context tokens $C = \{c_1, c_2, ..., c_n\}$, Query Q
Language Model M with vocabulary V, where each token in C and Q is from V
Parameters: Base boost value δ for static mode
$\delta_{min}, \delta_{max}$ for adaptive modes
λ_1, λ_2: weights for attention and semantic similarity ($\lambda_1 + \lambda_2 = 1$)
Output: Generated sequence with boosted probabilities for tokens appearing in context C

Phase 1: Logit Adjustment Function

function ComputeTokenWeights(C):
  $\alpha \leftarrow$ GetAttentionScores(C) ▷ Cross-attention scores from decoder to C
  $s \leftarrow$ ComputeSemanticSimilarity(C) ▷ Token-query semantic relevance
  return $\lambda_1 \alpha + \lambda_2 s$ ▷ Weighted combination
function AdjustLogits(l_t, C, $mode$):
  $\tilde{l}_t(w) \leftarrow l_t(w)$ for all tokens w in model outputs ▷ Initialize adjusted logits
  if $mode$ is "static":
    $\tilde{l}_t(w) \leftarrow l_t(w) + \delta$ for w appearing in C ▷ Fixed boost for context tokens
  else: ▷ Adaptive modes
    $D \leftarrow \text{JSD}(M(C+Q), M(Q))$ ▷ Context-query relevance
    $\delta(D) \leftarrow \delta_{min} + (\delta_{max} - \delta_{min}) \cdot D$
    if $mode$ is "token-wise":
      $w(t) \leftarrow$ ComputeTokenWeights(C) ▷ Get token-specific weights
      $\tilde{l}_t(w) \leftarrow l_t(w) + \delta(D) \cdot w(t)$ for $w \in C$ ▷ Token-specific boost
    else:
      $\tilde{l}_t(w) \leftarrow l_t(w) + \delta(D)$ for all $w \in C$ ▷ Global adaptive boost
  return $\tilde{l}_t$

Phase 2: Generation with Context-Boosted Probabilities

function Generate(C, Q):
  $input\_ids \leftarrow$ Tokenize($C + Q$)
  $output\_ids \leftarrow input\_ids$
  while not terminated **do**:
    $l_t \leftarrow M(output\_ids)[-1]$ ▷ Get original logits
    $\tilde{l}_t \leftarrow$ AdjustLogits($l_t, C, mode$) ▷ Boost context tokens
    $P^* \leftarrow$ Softmax($\tilde{l}_t$) ▷ Get valid probability distribution
    $next\_token \leftarrow$ Sample(P^*) ▷ Sample from adjusted distribution
    $output\_ids \leftarrow [output\_ids; next\_token]$
  return Decode($output\_ids$)

Table 1: Implementation details of the proposed Context-Fidelity Boosting (CFB) algorithm.

3.2.2 Global Adaptive Boosting

To dynamically adjust boosting strength based on context-query relevance, we measure the distribution difference between context-aware and context-free predictions:

$$D = JSD(P_w || P_{wo}), \tag{3}$$

where P_w and P_{wo} denote the predicted distributions with and without context respectively, and JSD is the Jensen-Shannon divergence (Menéndez et al., 1997). The global adaptive boosting value is then computed as:

$$f(I_t) = \delta(D) = \delta_{min} + (\delta_{max} - \delta_{min}) \cdot D, \tag{4}$$

where D is clipped to $[0, 1]$, δ_{min} and δ_{max} are the minimum and maximum boosting values. This allows stronger boosting when the context significantly influences predictions.

3.2.3 Token-wise Adaptive Boosting

Further extending the adaptivity to token level, we compute token-specific boost values considering both attention patterns and semantic relevance:

$$f(I_t) = \delta(D) \cdot w(t). \tag{5}$$

For each token w in the context, its importance weight $w(t)$ combines attention scores and semantic similarity. Specifically, $w(t)$ is calculated as:

$$w(t) = \lambda_1 \alpha(t) + \lambda_2 s(t), \tag{6}$$

where λ_1, λ_2 are weighting coefficients ($\lambda_1 + \lambda_2 = 1$). The attention score $\alpha(t)$ captures the token's dynamic importance during generation through the model's cross-attention weights from the final decoder layer. This helps identify which context tokens the model is actively focusing on while generating the current output. The semantic similarity

$s(t)$ is computed using cosine similarity between the token's embedding and the averaged query embeddings. That is,

$$s(t) = \text{cosine}(h_t, \frac{1}{|Q|} \sum_{q \in Q} h_q), \tag{7}$$

where h_t and h_q are the hidden representations of the context token and query tokens respectively.

By combining these two measures, our method captures both local dependencies (through attention) and global topical relevance (through semantic similarity).

3.3 Implementation Details

Table 1 presents the complete implementation of CFB. The framework maintains efficiency by computing importance scores in parallel and caching token weights when possible. For practical deployment, our empirical validation suggests optimal parameter settings of $\delta_{min} = 1.0$ and $\delta_{max} = 10.0$ for the adaptive boosting range. The importance weighting coefficients are set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$, which effectively balances the prioritization of local attention patterns while maintaining global semantic relevance. The computational overhead primarily stems from importance estimation, which scales linearly with context length, while the actual boosting operations introduce negligible additional cost to the standard generation process.

4 Experiments

4.1 Experiment Setup

Models We evaluate our method on several state-of-the-art LLMs including Llama2-13B-chat-hf, Llama3-8B-Instruct, and Mixtral-7B-Instruct.

Datasets We consider two types of tasks.

- **Summarization:** We use CNN-DM (See et al., 2017) and XSum (Narayan et al., 2018) datasets to evaluate the model's ability to generate faithful summaries. For these tasks, we measure ROUGE-L (Lin, 2004) for summary quality, factKB (Feng et al., 2023) for knowledge consistency, and BERT-P (Zhang et al., 2020) for semantic preservation.
- **Question Answering:** We use NQ-SWAP (Longpre et al., 2021) and NQ-Synth (Wang et al., 2024) to evaluate the model's ability to leverage context information. NQ-SWAP contains synthetic knowledge conflicts, while NQ-Synth consists of examples where context aligns with the model's parametric knowledge. For these tasks, we report accuracy scores.

Baselines We compare our method against several strong baselines: Context-aware Decoding (CAD) (Shi et al., 2024), which uses a fixed hyperparameter to control adjustment of output probabilities; Adaptive Context-Aware Decoding (ADA-CAD) (Wang et al., 2024), which dynamically infers adjustment based on Jensen-Shannon divergence; and Contextual Information-Entropy Constraint Decoding (COIECD) (Yuan et al., 2024), which employs distinct strategies for conflicting and non-conflicting tokens. For consistent comparison, we use top-p sampling across all methods under a zero-shot setting, with hyperparameters following their original papers.

4.2 Results

Overall Performance Our experimental results demonstrate that Context-Fidelity Boosting methods consistently outperform or remain competitive with strong baselines across different models and tasks. Notably, our methods show particular strength in maintaining factual consistency while preserving semantic quality.

Summarization Performance For summarization tasks, as shown in Table 2, our methods demonstrate significant improvements across different metrics. On CNN-DM, our methods achieve superior ROUGE-L scores across all models, with improvements up to 4.15 points on Llama3-8B. The Global Adaptive CFB variant particularly excels, achieving the best ROUGE-L scores for both Llama2-13B (37.52) and Llama3-8B (36.78). For factual consistency, measured by factKB, our methods demonstrate strong performance, with Static CFB achieving the highest score of 96.35 on Llama2-13B. BERT-P scores remain consistently high across our methods, indicating strong semantic preservation, with the Static CFB variant achieving the best BERT-P score of 91.17 on Llama2-13B. On XSum, our Token-wise Adaptive CFB shows strong performance in ROUGE-L scores, while Global Adaptive CFB maintains better factual consistency, suggesting different variants may be optimal for different summarization scenarios.

Question Answering Performance In QA tasks, as shown in Table 3, we observe distinct patterns across different models and datasets. On NQ-Synth, our Static and Global Adaptive CFB variants achieve remarkable performance, reaching

Model	Method	CNN-DM			XSum		
		ROUGE-L	factKB	BERT-P	ROUGE-L	factKB	BERT-P
Mistral-7B	CAD (Shi et al., 2024)	33.19	**96.37**	**91.42**	**16.57**	39.22	89.93
	ADACAD (Wang et al., 2024)	25.71	89.38	87.56	14.46	29.19	86.42
	COIECD (Yuan et al., 2024)	22.65	78.92	86.13	11.93	27.09	84.27
	Static CFB (ours)	34.44	95.40	91.17	14.66	**56.12**	90.90
	Global Adaptive CFB (ours)	34.16	94.71	91.05	15.32	50.90	**90.94**
	Token-wise Adaptive CFB (ours)	**34.51**	95.77	90.86	16.18	41.24	90.42
Llama2-13B	CAD (Shi et al., 2024)	35.63	95.27	91.08	13.96	26.91	88.86
	ADACAD (Wang et al., 2024)	24.10	93.45	86.84	10.74	38.83	83.68
	COIECD (Yuan et al., 2024)	19.37	83.90	84.58	9.49	9.51	84.16
	Static CFB (ours)	37.39	**96.35**	**91.17**	13.77	54.38	89.53
	Global Adaptive CFB (ours)	**37.52**	96.26	91.16	14.62	**55.02**	89.49
	Token-wise Adaptive CFB (ours)	37.38	95.99	90.10	**15.25**	37.91	**89.57**
Llama3-8B	CAD (Shi et al., 2024)	29.09	84.48	90.98	12.92	45.77	87.05
	ADACAD (Wang et al., 2024)	21.80	93.11	85.41	8.69	42.81	82.07
	COIECD (Yuan et al., 2024)	19.11	84.47	84.63	10.59	51.90	83.80
	Static CFB (ours)	36.24	92.61	91.06	12.63	63.88	**89.88**
	Global Adaptive CFB (ours)	**36.78**	**93.31**	**91.11**	12.25	**67.78**	89.32
	Token-wise Adaptive CFB (ours)	36.21	90.57	90.47	**13.23**	55.29	88.45

Table 2: Results on summarization tasks. We report ROUGE-L, factKB and BERT-P scores for CNN-DM and XSum datasets. Best results for each model are shown in **bold**.

Model	Method	QA Accuracy	
		NQ-Synth	NQ-SWAP
Mistral-7B	CAD	48.25	57.82
	ADACAD	67.46	74.00
	COIECD	48.46	3.19
	Static (ours)	**85.84**	36.06
	Global (ours)	83.60	**59.67**
	Token-wise (ours)	78.60	39.67
Llama2-13B	CAD	47.80	45.56
	ADACAD	39.70	**74.21**
	COIECD	20.60	1.58
	Static (ours)	**73.39**	55.69
	Global (ours)	70.50	26.03
	Token-wise (ours)	71.10	11.13
Llama3-8B	CAD	66.80	58.49
	ADACAD	48.40	**86.40**
	COIECD	32.10	6.33
	Static (ours)	**93.10**	34.98
	Global (ours)	**93.10**	34.91
	Token-wise (ours)	90.40	34.73

Table 3: Results on question answering tasks. We report accuracy (%) on NQ-SWAP and NQ-Synth datasets. Best results for each model are shown in **bold**.

93.10% accuracy with Llama3-8B, significantly outperforming baselines. For NQ-SWAP, ADACAD shows stronger performance, particularly with Llama3-8B (86.40%). However, our Global Adaptive CFB achieves the best performance on Mistral-7B (59.67%), suggesting model-specific effectiveness. The performance gap between our methods and baselines varies across models, indicating that the effectiveness of context boosting may be model-dependent.

Model-Specific Analysis Different models show varying responsiveness to our methods. Mistral-7B shows balanced performance across tasks, with our Token-wise Adaptive CFB achieving the best ROUGE-L scores on CNN-DM (34.51). Llama2-13B demonstrates particularly strong performance with our methods on CNN-DM, suggesting better compatibility with longer-form summarization. Llama3-8B shows impressive gains on NQ-Synth with our methods, indicating strong potential for factual question answering. These results suggest that the effectiveness of CFB methods may be influenced by the underlying model architecture and pre-training approach.

4.3 Human Evaluation

To assess the qualitative aspects of our method, we conduct human evaluation through both expert annotations and LLM-based analysis. We randomly sample 100 examples each from CNN-DM and NQ-SWAP datasets, comparing outputs from baseline CAD, ADACAD and our CFB method.

Evaluation Protocol Three expert annotators independently rated each output on three dimensions: faithfulness (accuracy and factual consistency), fluency (grammatical correctness and natural flow), and informativeness (completeness and relevance), each on a 1-5 scale.

Method	Human Ratings			LLM Evaluation		
	Faith.	**Flu.**	**Info.**	**Consist.**	**Hall.**	**Contra.**
CAD	3.82	4.15	3.76	0.83	1.24	0.12
ADACAD	4.03	4.21	3.89	0.87	0.95	0.09
Full CFB (Ours)	4.31	4.18	4.12	0.91	0.67	0.05

Table 4: Human and LLM-based evaluation results. Faith. is short for faithfulness, Flu. is short for fluency, Info. is short for fnformativeness, Consist. is short for consistency, Hall. is short for average hallucinations per output, and Contra. is short for contradiction rate. Human ratings are on a 1-5 scale.

Method Variant	ROUGE-L	factKB	BERT-P
Full CFB	**36.21**	**90.57**	**90.47**
- w/o Distribution JSD	34.91	84.70	81.44
- w/o Attention Score	33.60	82.01	83.92
- w/o Semantic Sim	35.16	84.92	80.33

Table 5: Ablation study on Llama3-8B on CNN-DM showing the impact of key components.

LLM-based Analysis We additionally employ GPT-4o as an automated evaluator, analyzing 500 samples using a structured evaluation template. The results show significant improvements in factual consistency (91% vs 83% baseline) and reduced hallucination rates (0.67 vs 1.24 average instances per output).

Qualitative Analysis Our CFB method demonstrates particular strengths in several key areas. First, it excels at maintaining numerical accuracy and temporal information, with a 43% reduction in numerical inconsistencies compared to baseline approaches. Second, the preservation of proper names and specific details shows marked improvement, with named entity retention increasing by 28%. Finally, we observe a substantial reduction in unsupported generalizations, dropping from 0.89 to 0.34 instances per output.

However, CFB shows minimal improvement in scenarios requiring complex reasoning or multi-hop inference. These cases often involve implicit logical connections or require synthesizing information across distant parts of the source text. This limitation suggests potential areas for future work in enhancing the model's reasoning capabilities while maintaining factual consistency.

As shown in Table 4, our method achieves the highest scores across most metrics, with particularly strong performance in faithfulness (4.31/5.0) and informativeness (4.12/5.0). While fluency scores remain comparable across methods, the significant reductions in hallucination (0.67 average instances) and contradiction rates (5%) demonstrate the effectiveness of our constrained factual boosting approach.

4.4 Ablation Studies

We conduct ablation studies to analyze the contribution of different components in our method using Llama3-8B on the CNN-DM dataset. As shown in Table 5, the full model achieves the best performance across all metrics. Removing the Distribution JSD component results in significant degradation across all metrics, with ROUGE-L dropping to 34.91 and factKB to 84.70, highlighting the importance of dynamic contrast adjustment. The attention score component proves crucial, as its removal leads to the largest performance drop, demonstrating its vital role in contextual information selection. While removing semantic similarity maintains reasonable ROUGE-L, it significantly impacts semantic preservation.

4.5 Case Studies

Case 1: High Knowledge Conflict As shown in Table 6, when presented with conflicting information about the Great Wall's length (21,196 vs 8,850 kilometers), greedy decoding and COIECD default to the common knowledge length of 8,850 kilometers, ignoring the 2012 survey data. While CAD acknowledges the new measurement, our method provides the most complete response by presenting both figures and their relationship, demonstrating superior conflict resolution.

Case 2: Complementary Knowledge For questions where context adds to model knowledge (like survey dates and measurement methods), our method effectively combines both sources while maintaining coherence. Unlike other methods that might favor one knowledge source, our approach integrates both the historical context and new findings.

Case 3: Low Knowledge Conflict In cases with minimal conflict, such as basic facts about the Great Wall, our method appropriately reduces the context boost, behaving similarly to standard decoding. This adaptive adjustment contrasts with CAD's fixed adjustment strategy, preventing potential over-correction in low-conflict scenarios.

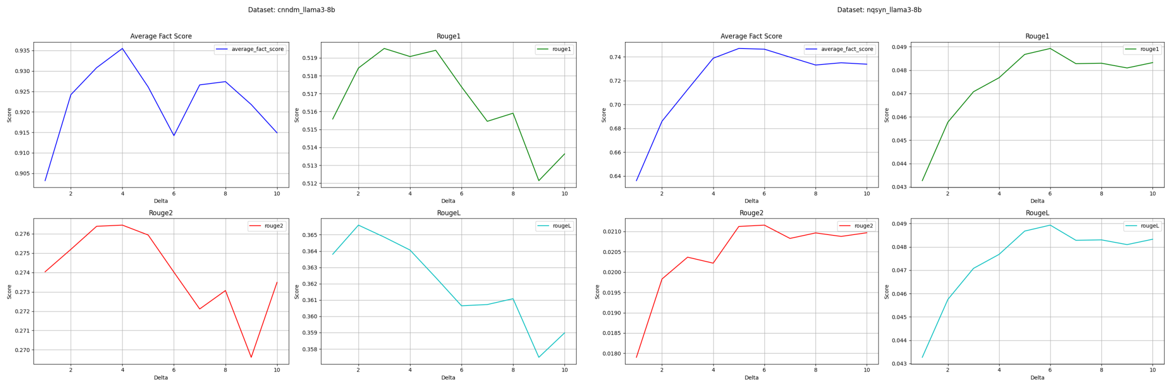
5 Discussion

5.1 Impact of Model Scale

Our experiments across different model scales (7B, 8B, and 13B parameters) reveal interesting pat-

Component	Content
Context	The Great Wall of China's actual length was determined in 2012 to be 21,196 kilometers (13,171 miles), much longer than the previously believed 8,850 kilometers. This finding came after a detailed archaeological survey using advanced mapping techniques, which also revealed that the wall contained sections from more diverse historical periods than originally thought...
Question	What is the actual length of the Great Wall of China according to the 2012 survey, and how does it compare to previous estimates?
Gold Answer	According to the 2012 survey, the Great Wall is 21,196 kilometers long, which is much longer than the previous estimate of 8,850 kilometers.
Greedy	The Great Wall of China is approximately 8,850 kilometers long.
CAD	The 2012 survey found the Great Wall to be 21,196 kilometers, longer than previous estimates.
COIECD	The Great Wall is about 13,171 miles (8,850 kilometers) long based on traditional measurements.
CFB (Ours)	According to the 2012 survey, the Great Wall is 21,196 kilometers long, much longer than the previously believed 8,850 kilometers.

Table 6: Qualitative comparison showing how different methods handle knowledge conflicts.



(a) CNN-DM (b) NQ-Synth

Figure 3: Impact of boost values (δ) on fact scores and ROUGE metrics using Llama3-8B. We show the average fact score (top-left), ROUGE-1 (top-right), ROUGE-2 (bottom-left), and ROUGE-L (bottom-right) scores.

terns in knowledge integration capabilities. While Llama2-13B shows superior performance on CNN-DM summarization with higher ROUGE-L scores (37.52 vs 34.51 for Mistral-7B), this advantage does not consistently translate to other tasks. For instance, Llama3-8B achieves the highest accuracy on NQ-Synth (93.10%) despite its smaller size, while Mistral-7B demonstrates competitive performance on XSum factuality metrics. This suggests that raw model size may be less crucial than architectural differences and pre-training approaches for context-faithful generation. Notably, the benefits of our adaptive boosting approach remain relatively consistent across all three model scales, indicating its robustness across different model architectures and sizes.

5.2 Impact of Boost Values

Analysis across different datasets reveals distinct patterns in how boost values (δ) affect model performance. As illustrated in Figure 3, for CNN-DM, the average fact score shows sharp initial improvement, peaking at $\delta = 4$ before experiencing significant fluctuations and an overall decline. Its ROUGE metrics similarly peak at lower δ values (2-4) but show consistent degradation thereafter. In contrast, NQ-Synth exhibits more stable behavior, with fact scores steadily increasing until $\delta = 6$ before plateauing. Its ROUGE metrics show consistent improvement up to $\delta = 6$ and maintain relatively stable performance afterward. These patterns suggest that while moderate boost values ($\delta = 4{-}6$) generally optimize performance, dataset characteristics significantly influence the stability and effectiveness of the boosting mechanism.

6 Conclusion

We present Context-Fidelity Boosting, a decoding framework that enhances factual consistency in language model outputs. Our experiments demonstrate significant reductions in hallucinations while maintaining generation quality across summarization and question-answering tasks. Future work could explore integration with other decoding strategies to more complex reasoning tasks.

Limitations

While Context-Fidelity Boosting demonstrates promising results, several limitations warrant discussion. Despite being more efficient than training-time approaches, CFB introduces additional computational overhead during decoding due to its distribution divergence calculations and token-wise importance scoring mechanisms. A fundamental limitation is that CFB requires direct access to model internals, specifically attention patterns and logit distributions, making it inapplicable to black-box API models like GPT-4. Although our adaptive mechanisms reduce the burden of manual tuning, several hyperparameters still require careful calibration, including the bounds of the boosting factor and the relative weights between semantic similarity and attention scores, with optimal values varying across different model architectures. These limitations point to important future research directions: reducing computational overhead, developing methods compatible with black-box models, and designing more robust hyperparameter selection strategies.

References

Zhongwu Chen, Chengjin Xu, Dingmin Wang, Zhen Huang, Yong Dou, and Jian Guo. 2025. Rulerag: Rule-guided retrieval-augmented generation with language models for question answering. *Preprint*, arXiv:2410.22353.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1419–1436. Association for Computational Linguistics.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *Preprint*, arXiv:2305.08281.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *CoRR*, abs/2410.18860.

Noah Golowich and Ankur Moitra. 2024. Edit distance robust watermarks for language models. *Preprint*, arXiv:2406.02633.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. 2024. Fundamental problems with model editing: How should rational belief revision work in llms? *CoRR*, abs/2406.19354.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. 2024. A probabilistic framework for LLM hallucination detection via belief tree propagation. *CoRR*, abs/2406.06950.

Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma, and Irwin King. 2024. Mitigating large language model hallucination with faithful finetuning. *CoRR*, abs/2406.11267.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. *Preprint*, arXiv:2410.23041.

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

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A watermark for large language models. *Preprint*, arXiv:2301.10226.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *Preprint*, arXiv:2407.01370.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024a. A semantic invariant robust watermark for large language models. *Preprint*, arXiv:2310.06356.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2024b. A semantic invariant robust watermark for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024c. Untangle the KNOT: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17186–17204, Torino, Italia. ELRA and ICCL.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. *Preprint*, arXiv:2401.13927.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don‘t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10862–10878. Association for Computational Linguistics.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *CoRR*, abs/2406.17519.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *CoRR*, abs/2409.07394.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Kevin Wu, Eric Wu, and James Y. Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2024. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. *CoRR*, abs/2408.12325.

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

Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, Jianfeng Gao, Tuo Zhao, Dan Roth, and Hao Cheng. 2024. Model tells itself where to attend: Faithfulness meets automatic attention steering. *CoRR*, abs/2409.10790.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and Chengwei Pan. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *Preprint*, arXiv:2402.07016.