Contrastive Decoding with Information Flow Intervention for Mitigating Hallucinations in Large Language Models

Anonymous ACL submission

Abstract

Large language models have demonstrated strong capabilities across a wide range of natural language generation tasks, yet they remain prone to hallucinations in practical applications. This paper proposes IFI-CD, an information flow intervention-based contrastive decoding method that characterizes the model's reliance on contextual evidence at the level of generation decisions and suppresses hallucinated outputs by modulating internal information flow. Unlike approaches that enhance contrastive signals by amplifying predictive distribution differences or directly inducing hallucinations, IFI-CD explicitly focuses on the directionality of generation shifts during contrastive decoding, enabling decoding decisions to more stably align with factual directions. Experimental evaluations on multiple contextual consistency and factuality assessment tasks demonstrate that IFI-CD provides more discriminative and directionally consistent contrastive information, effectively mitigating context-related hallucinations. Further analysis shows that explicitly modeling the directionality of induced shifts helps translate distribution-level variations into factual improvements at the generation decision level.

1 Introduction

Large language models (LLMs) have recently demonstrated strong generative capabilities across a variety of natural language processing (NLP) tasks, including question answering, summarization, and reasoning (Zhao et al., 2023; Achiam et al., 2023; Touvron et al., 2023). However, during practical generation, LLMs often exhibit hallucinations (Ji et al., 2023; Zhang et al., 2025c), producing content that is inconsistent with objective facts or misaligned with the given context (Lei et al., 2023). Such behavior not only undermines the credibility of model outputs but also substantially limits their deployment in high-risk scenarios that demand strong reliability and trustworthiness.

A broad range of approaches have been explored to address hallucinations in LLMs, including training, preference alignment, external knowledge augmentation, and inference-time interventions. This paper focuses on inference-time methods that require no additional training or external knowledge. Within this category, contrastive decoding (Li et al., 2023c) and its variants have gradually emerged as an important technical direction for hallucination mitigation. These methods typically construct "expert–amateur" contrastive distributions, for example by leveraging predictions from different model layers (Chuang et al., 2024; Zhang et al., 2025a) or by applying layer pruning to obtain more divergent amateur distributions (Yu et al., 2025), and reweight next-token predictions during decoding to suppress potential hallucinated outputs.

Although the aforementioned methods improve generation quality to some extent, existing work often assumes that more pronounced predictive distribution differences yield more effective contrastive signals. Under this assumption, some approaches intervene in the decoding process by constructing stronger amateur distributions or imposing larger probability reweighting (Yu et al., 2025); others attempt to actively induce hallucinated outputs to obtain more discriminative contrastive distributions (Liang et al., 2025; Zhang et al., 2025b). However, larger distributional perturbations do not necessarily correspond to improvements along factual directions, and overly strong or indiscriminate interventions may even disrupt the model's original predictive structure, leading to unstable generation or performance degradation. More importantly, existing methods often lack explicit characterization of how evidence attenuation at the predictive distribution level translates into directionally meaningful decision biases. As a result, even stronger contrastive signals provide no guarantee that gener-

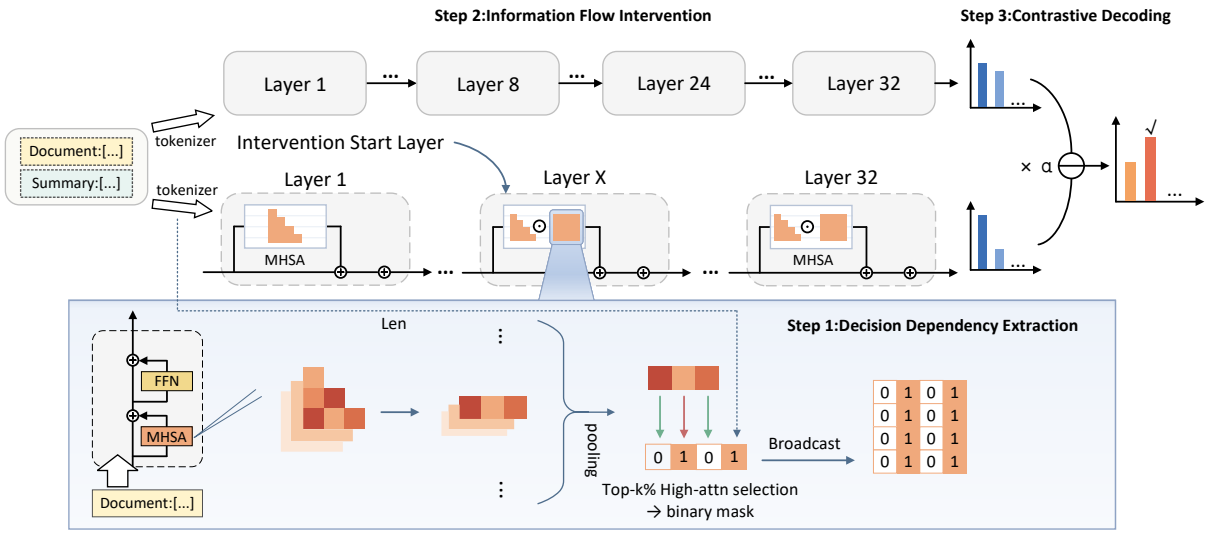

Figure 1: Overview of the IFI-CD framework. Given an input context, IFI-CD identifies key information tokens on which generation decisions depend and selectively regulates internal information flow according to this dependency. The resulting intervened distribution is contrastively integrated with the original predictive distribution at the decoding stage to produce the final output.

ation shifts will stably align with factual directions.

In this paper, we revisit hallucination mitigation in LLMs from the perspective of generation decisions and propose a novel inference-time mitigation framework, termed Information Flow Intervention-based Contrastive Decoding (IFI-CD). The framework takes generation-decision dependency as its core and guides structural changes in evidence contribution within the model through information flow intervention, thereby inducing controllable and directionally meaningful shifts in predictive distributions. These shifts are explicitly exploited during decoding to suppress hallucinated outputs. The overall process requires no additional training or external knowledge, and aims to steer generation decisions toward factual directions in a more stable manner while preserving the model's original semantic structure.

The contributions of this work are as follows:

- IFI-CD is an inference-time contrastive decoding framework that suppresses hallucinated generation by constructing directionally meaningful signals with stronger contrastive effectiveness.
- A generation-decision dependency perspective characterizes how the model relies on contextual evidence, providing a modeling view for analyzing evidence-dependency imbalance in hallucination generation.
- An information flow intervention-based evidence attenuation mechanism induces directional distributional shifts within the model and provides more effective decision signals for contrastive decoding.

2 Method

An overview of the IFI-CD framework is illustrated in Fig. 1. The method centers on the model's dependency on contextual evidence during generation decisions and mitigates hallucinations through coordinated mechanisms at two levels: internal representation–level intervention and decoding-stage constraint.

2.1 Identification of Generation-Decision Dependent Tokens

In autoregressive generation tasks, a model typically depends on a subset of contextual evidence at each prediction step, rather than attending uniformly to the entire input. Given a context sequence $X = (x_1, x_2, \ldots, x_N)$, the query at position $t = N$, corresponding to the generation step, is defined as the generation-decision point. Based on this position, the model's dependency on contextual tokens is characterized.

Specifically, we treat the attention weights produced by the LLM across layers and attention heads as internal dependency signals. For the h-th attention head in the ℓ-th layer, the self-attention weight

matrix is denoted as $\mathrm{Attn}^{\ell,h} \in \mathbb{R}^{N\times N}$, where $\mathrm{Attn}^{\ell,h}_{i,j}$ represents the attention weight from the query at position i to the key at position j. On this basis, we extract the row vector of the attention matrix corresponding to the generation-decision point:

$$\mathrm{Attn}^{\ell,h}_{t} = \left[\mathrm{Attn}^{\ell,h}_{t,1}, \mathrm{Attn}^{\ell,h}_{t,2}, \ldots, \mathrm{Attn}^{\ell,h}_{t,N}\right] \quad (1)$$

where $\mathrm{Attn}^{\ell,h}_{t}$ denotes the dependency-strength distribution from the generation-decision point to all contextual positions under the h-th attention head in the ℓ-th layer. Given the functional heterogeneity across layers and attention heads, these dependency distributions are aggregated across all layers and heads to obtain an overall dependency estimate:

$$\mathrm{Attn}_t = \frac{1}{LH}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\mathrm{Attn}^{\ell,h}_{t} \quad (2)$$

where L denotes the number of model layers, H denotes the number of attention heads per layer, and $\mathrm{Attn}_t[j]$ represents the average dependency strength of the generation-decision point on the j-th contextual token.

Subsequently, contextual tokens are ranked in descending order according to Attn_t, and the subset of tokens whose scores fall within the top ρ proportion is selected as the target set for subsequent information-flow intervention:

$$\mathcal{D}(X) = \mathrm{Top}_{\rho}(\mathrm{Attn}_N) \quad (3)$$

where $\rho \in (0,1)$ is a sparsity control parameter, and $\mathrm{Top}_{\rho}(\cdot)$ denotes the index set corresponding to the top $\lceil \rho N \rceil$ positions ranked by attention scores. Tokens in $\mathcal{D}(X)$ are referred to as generation-decision-dependent tokens. In practice, this set is computed once from the initial input context and remains fixed throughout subsequent decoding. Notably, $\mathcal{D}(X)$ characterizes the contextual evidence actually relied upon by the current generation decision, rather than semantic importance or global salience.

2.2 Information-Flow Intervention within a Layer-Sensitive Window

This section defines a layer-sensitive information-flow intervention mechanism based on the generation-decision-dependent tokens identified in Section 2.1. The mechanism induces structured regulation of internal information flow within the model, enabling controllable shifts in model behavior while preserving semantic consistency and laying the foundation for discriminative contrastive distributions in the subsequent decoding stage.

Prior studies have shown that different layers of LLMs play distinct roles in representational abstraction and decision integration. Based on this observation, the intervention timing is parameterized by an intervention start layer ℓ_0: the model performs standard modeling over the full context in the first $\ell_0 - 1$ layers, and interventions on generation-decision-dependent tokens remain active from layer ℓ_0 onward. Formally, a layer-wise gating function is defined as:

$$g(\ell;\ell_0) = \begin{cases} 0, \ell < \ell_0, \\ 1, \ell \geq \ell_0, \end{cases} \quad \ell \in \{1,\ldots,L\} \quad (4)$$

where $g(\ell;\ell_0) = 1$ indicates that the intervention is active at layer ℓ and all subsequent layers. The candidate range of the intervention start layer is further defined as $\mathcal{W} = [\ell_{\min}, \ell_{\max}]$, referred to as the layer-sensitive window. Its effectiveness is examined through intervention start-layer analysis in Section 3.3.1.

The specific information-flow intervention operator is then formalized. The core idea is to impose gated control over the key-side visibility of generation-decision-dependent tokens within self-attention, thereby limiting their role as information sources in subsequent representation updates. For the h-th attention head in the ℓ-th layer, the standard causal self-attention is formulated as:

$$\begin{aligned} A^{(\ell,h)} &= \mathrm{softmax}\left(\frac{Q^{(\ell,h)}K^{(\ell,h)\top}}{\sqrt{d}} \odot M_{\text{causal}}\right), \\ O^{(\ell,h)} &= A^{(\ell,h)}V^{(\ell,h)}. \end{aligned} \quad (5)$$

where $Q^{(\ell,h)}$, $K^{(\ell,h)}$, and $V^{(\ell,h)} \in \mathbb{R}^{N\times d}$ denote the query, key, and value matrices at the h-th attention head of the ℓ-th layer, and $M_{\text{causal}} \in \{0,1\}^{N\times N}$ is the causal mask matrix. Based on $\mathcal{D}(X)$, a binary mask vector is defined as:

$$m_j = \begin{cases} 0, j \in \mathcal{D}(X), \\ 1, \text{otherwise}, \end{cases} \quad j = 1, ..., N \quad (6)$$

where $m_j = 0$ indicates that the key corresponding to the j-th token is masked in the attention computation. This binary mask is further broadcast to form a mask matrix shared across the key dimension:

$$M_{\mathcal{D}} = \mathbf{1}m^{\top} \in \{0,1\}^{N\times N} \quad (7)$$

where $\mathbf{1}$ denotes an all-ones tensor whose dimensionality matches that of the context, and $M_{\mathcal{D}}$ represents the global key-side visibility constraint imposed on generation-decision-dependent tokens. To realize layer-sensitive information-flow intervention, Eq. 4 is applied to define the intervened attention weights as follows:

$$\tilde{A}^{(\ell,h)} = \mathrm{softmax}\Bigg(\frac{Q^{(\ell,h)}K^{(\ell,h)\top}}{\sqrt{d}} \odot M_{\mathrm{causal}} \odot \Big((1 - g(\ell;\ell_0))\,\mathbf{1} + g(\ell;\ell_0)\,M_{\mathcal{D}}\Big)\Bigg) \tag{8}$$

where $\tilde{A}^{(\ell,h)} \in \mathbb{R}^{N\times N}$ denotes the attention weight matrix of the h-th attention head in the ℓ-th layer under intervention. When $\ell < \ell_0$, $g(\ell;\ell_0) = 0$, and the above formulation reduces to standard causal self-attention, under which the model retains full access to the contextual information at that layer. When $\ell \geq \ell_0$, $g(\ell;\ell_0) = 1$, and the generation-decision–dependent tokens are uniformly masked on the key side, preventing them from being attended to by queries at any position. The corresponding attention output is given by:

$$\tilde{O}^{(\ell,h)} = \tilde{A}^{(\ell,h)}V^{(\ell,h)} \tag{9}$$

where $\tilde{O}^{(\ell,h)}$ denotes the output representation of the corresponding attention head under intervention and participates in subsequent representation updates. Since self-attention constitutes the sole mechanism for cross-token information interaction in Transformer architectures, key-side gating is equivalent to restricting the diffusion paths through which specific tokens act as information sources to all positions. Combined with the layer-sensitive window, this intervention attenuates the model's sustained access to critical evidence while preserving the overall semantic structure, thereby inducing output distributions with directional shifts.

2.3 Contrastive Decoding

Given an input prefix $x_{<t}$, inference is performed under the original model and the information-flow-intervened model, yielding the next-token predictive distributions $p_{\mathrm{base}}(\cdot|x_{<t})$ and $p_{\mathrm{int}}(\cdot|x_{<t})$, respectively. Here, $p_{\mathrm{int}}(\cdot|x_{<t})$ is obtained solely through the information-flow intervention described in Section 2.2, without any modification to model parameters. Since key information sources are attenuated, p_{int} more readily reflects generation preferences insensitive to contextual evidence, and therefore serves as the contrastive distribution against p_{base}.

Specifically, the final distribution for decoding is obtained through interpolation-based contrastive fusion:

$$\begin{aligned} p(x_t \mid x_{<t}) &= \mathrm{softmax}\Big(\log p_{\mathrm{base}}(x_t \mid x_{<t}) \\ &\quad + \alpha\,\Delta_t(x_t)\Big), \\ \Delta_t(x_t) &= \log \frac{p_{\mathrm{base}}(x_t \mid x_{<t})}{p_{\mathrm{int}}(x_t \mid x_{<t})}. \end{aligned} \tag{10}$$

where α denotes the contrastive strength coefficient. When $\alpha = 0$, the procedure reduces to standard autoregressive decoding; when $\alpha > 0$, the contrastive term penalizes candidate tokens that retain high confidence under information-flow-restricted conditions, thereby decreasing the likelihood that hallucination-prone preferences are selected.

3 Experiments

3.1 Experimental Setup

Datasets. IFI-CD is evaluated on multiple benchmarks covering both context-dependent hallucinations and factual hallucinations. Context-dependent hallucination evaluation includes HellaSwag (Zellers et al., 2019), RACE-M/RACE-H (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), and HaluEval-Sum (Li et al., 2023a); factual hallucination evaluation is conducted on TruthfulQA (Lin et al., 2022), including both multiple-choice and open-ended generation settings.

Baselines. IFI-CD is compared with Greedy Decoding, DoLa (Chuang et al., 2024), END (Wu et al., 2025), PruneCD (Yu et al., 2025), and HICD (Jiang et al., 2025). These baselines cover greedy decoding, inter-layer contrastive decoding, entropy-based distribution reweighting, layer-pruning-based contrastive decoding, and hallucination-inducing contrastive decoding.

Implementation Details. Experiments are conducted with LLaMA-7B and LLaMA-2-7B as backbone models (Touvron et al., 2023). For IFI-CD, the selection ratio of generation-decision-dependent tokens is fixed at 0.1, and the intervention start layer is chosen from $\{0, 3, 5, 8, 12\}$. The contrastive strength coefficient α is determined on the validation set for each task. All baselines follow their original settings, with optimal configurations selected within the recommended ranges. Detailed dataset and metric descriptions, baseline implemen-

Table 1: Performance comparison of different methods on contextual consistency tasks. Bold indicates the best result, and underlining denotes the second-best result. The symbol "*" indicates results reported from the original papers. All methods report the best results under their respective recommended settings.

Models	Methods	HaluEval-Sum		HellaSwag	OpenBookQA	RACE	
		Acc_H	Acc_A	Acc	Acc	Middle	High
LLaMA-7b	Greedy	23.22	26.06	0.7761	0.5140	0.5642	0.4339
	DoLa	20.41*	25.91*	0.7517*	0.4840	0.5710*	0.4462*
	END	21.10	24.85	0.7956	0.5160	0.5675	0.4362
	PruneCD	26.09	26.20	0.7128	0.4640	0.5501	0.4028
	HICD	27.12	27.25	**0.8423***	0.5580	**0.5989***	**0.4668***
	IFI-CD(Ours)	**32.72**	**32.75**	0.8187	**0.5700**	0.5884	0.4623
LLaMA2-7b	Greedy	24.27	48.90	0.7832	0.4860	0.5801	0.4325
	DoLa	27.78*	50.31*	0.6925*	0.4940	0.5536*	0.4070*
	END	26.54	43.85	0.7973	0.4900	0.5794	0.4340
	PruneCD	30.80	50.15	0.7078	0.4860	0.5494	0.4028
	HICD	30.94	43.65	0.8433*	0.5220	0.5996*	0.4514*
	IFI-CD(Ours)	**34.85**	**50.45**	**0.8442**	**0.5540**	**0.6100**	**0.4757**

tation details, and hyperparameter settings are provided in the Appendix A.

3.2 Main Results

3.2.1 Mitigation of Context-Dependent Hallucinations

The overall performance of different hallucination mitigation methods on context-consistency–related tasks is summarized in Table 1. Overall, IFI-CD brings consistent gains under both backbone models and achieves the best or second-best performance on most tasks. Its advantage is especially clear on tasks that require long-context understanding and explicit evidence grounding: under the LLaMA-7B backbone, IFI-CD reaches scores of 32.72 and 32.75 on the Acc-H and Acc-A metrics of HaluEval-Sum, outperforming HICD by margins of 5.57 and 5.50, respectively, and achieves the highest accuracy of 57.0% on OpenBookQA. On context continuation and reading comprehension tasks such as HellaSwag and RACE, IFI-CD also consistently improves over greedy decoding and remains competitive with strong baselines such as HICD. A similar trend is observed under the LLaMA-2-7B backbone, where IFI-CD achieves scores of 34.85 and 50.45 on the Acc-H and Acc-A metrics of HaluEval-Sum, further demonstrating its effectiveness in complex contextual discrimination scenarios.

3.2.2 Mitigation of Factual Hallucinations

IFI-CD is further evaluated on TruthfulQA for factual hallucination mitigation, with results summarized in Table 2. Under the multiple-choice setting, IFI-CD achieves the best results on both backbone models: MC1, MC2, and MC3 reach 32.91, 54.88, and 30.24 on LLaMA-7B, and 28.10, 50.39, and 26.33 on LLaMA-2-7B, respectively. For open-ended generation, Truth*Info is taken as the primary metric, jointly reflecting factual accuracy and informativeness. IFI-CD achieves the highest score of 42.41 on LLaMA-7B and reaches 47.22 on LLaMA-2-7B, ranking second only to END. Overall, IFI-CD shows strong competitiveness in both factuality judgment and open-ended generation settings.

3.3 Analysis

3.3.1 Effects of Intervention Start Layer and Dependency Token Selection Ratio

A systematic analysis is conducted to examine the effects of the key-side information-flow intervention start layer and the selection ratio of generation-decision-dependent tokens on IFI-CD performance. As shown in Fig. 2, performance across different datasets follows a similar non-monotonic trend with respect to the intervention start layer: optimal or near-optimal results usually appear in shallow-to-middle layers, whereas deeper-layer intervention generally leads to performance degradation. In contrast, curves under different token selection ratios

Table 2: Performance comparison of different methods on the TruthfulQA factual evaluation task. Bold indicates the best result, and underlining denotes the second-best result. Open-ended generation results are evaluated by a publicly available and reproducible fine-tuned LLaMA-2-7B model.

Methods	TruthfulQA					
	MC1	MC2	MC3	Truth	Info	T * I
LLaMA-7b	23.92	39.25	18.09	32.78	96.33	29.11
+DoLa	30.76	51.31	27.13	40.00	98.35	38.86
+END	26.20	42.06	18.91	42.78	98.35	41.65
+PruneCD	23.92	38.10	17.66	43.67	98.48	42.15
+HICD	23.42	52.28	23.24	-	-	-
+IFI-CD	**32.91**	**54.88**	**30.24**	46.08	96.20	**42.41**
LLaMA2-7b	26.58	41.89	18.95	42.15	91.52	33.67
+DoLa	25.70	43.12	19.35	46.71	93.42	40.63
+END	26.20	42.59	19.04	52.53	97.59	**50.38**
+PruneCD	27.47	42.05	20.51	63.42	82.78	46.20
+HICD	21.52	44.67	21.33	-	-	-
+IFI-CD	**28.10**	**50.39**	**26.33**	52.41	94.56	47.22

show similar shapes, indicating that performance variation is mainly governed by the intervention layer, while the effect of the token ratio depends on the corresponding layer depth. To further explain this phenomenon, the following analysis examines logits distribution perturbation magnitude and perturbation direction. Unless otherwise specified, subsequent analyses are based on OpenBookQA and LLaMA-7B.

Magnitude of logits distributional perturbation. The objective is to quantify the overall strength of perturbations induced by key-side information-flow intervention on the model's output logits distribution, thereby analyzing how different intervention settings affect prediction behavior. Specifically, the 32 Transformer layers are grouped by intervention start position into three intervals: Shallow (layers 0–11), Middle (layers 12–23), and Deep (layers 24–31). The Jensen–Shannon (JS) divergence between the pre- and post-intervention logits distributions of the correct answer sequence is then computed as the metric for perturbation magnitude.

As shown in Fig. 3, logits distributional perturbation is primarily governed by the intervention start layer: shallow-layer intervention induces substantially larger distributional shifts than middle- and deep-layer intervention, while deep-layer intervention has almost no effect on the final output distribution. This suggests that once the model enters higher layers associated with decision integration, restricting the visibility of key information sources has markedly weaker influence on the output distribution. Meanwhile, within the shallow and middle intervals, the dependency token selection ratio has a modulatory effect on perturbation magnitude, with larger ratios generally corresponding to higher JS divergence values. However, larger distributional perturbations do not necessarily indicate more effective hallucination induction, and may instead reflect unstructured or semantically irrelevant distributional disruption.

Direction of logits shift. Further analysis examines whether key-side information-flow intervention produces directional shifts in the logits distribution, namely whether it systematically amplifies the relative advantage of incorrect candidates over correct ones. To this end, a directional shift metric Δ_ℓ is defined to quantify the change in the model's relative preference between the best incorrect answer $\hat{y}_{\text{err}}$ and the correct answer $y^\star$ before and after intervention. Given an input prefix x, the directional shift is defined as: $\Delta_\ell = \Big[\big(s_{\text{int}}(\hat{y}_{\text{err}} \mid x) - s_{\text{int}}(y^\star \mid x)\big) - \big(s_{\text{base}}(\hat{y}_{\text{err}} \mid x) - s_{\text{base}}(y^\star \mid x)\big)\Big]$, where $s_{\text{int}}(\cdot \mid x)$ and $s_{\text{base}}(\cdot \mid x)$ denote the model scores assigned to a candidate answer under the intervention and original conditions, respectively. When $\Delta_\ell > 0$, key-side information-flow intervention strengthens the model's preference for the incorrect candidate relative to the correct one.

As shown in Fig. 4, under the same intervention start layer, Δ_ℓ follows highly consistent trends across different selection ratios of generation-decision-dependent tokens. Although larger ratios may lead to greater shift magnitudes in shallow layers, they do not change the overall structure of directional shifts across intervention depths. This suggests that directional shifts are mainly driven by masking a small set of critical tokens, rather than by a linear accumulation of the selection ratio.

Furthermore, the peak of Δ_ℓ typically appears in shallow-to-middle layers rather than the earliest layers; as the intervention shifts toward deeper layers, Δ_ℓ rapidly decays and approaches zero. This indicates that overly early intervention is more likely to cause unstructured distributional disruption, whereas shallow-to-middle intervention can more stably induce the relative advantage of incorrect candidates while preserving a certain degree of contextual structure. Together with the perturbation magnitude analysis, these results show that the intervention layer determines whether percepti-

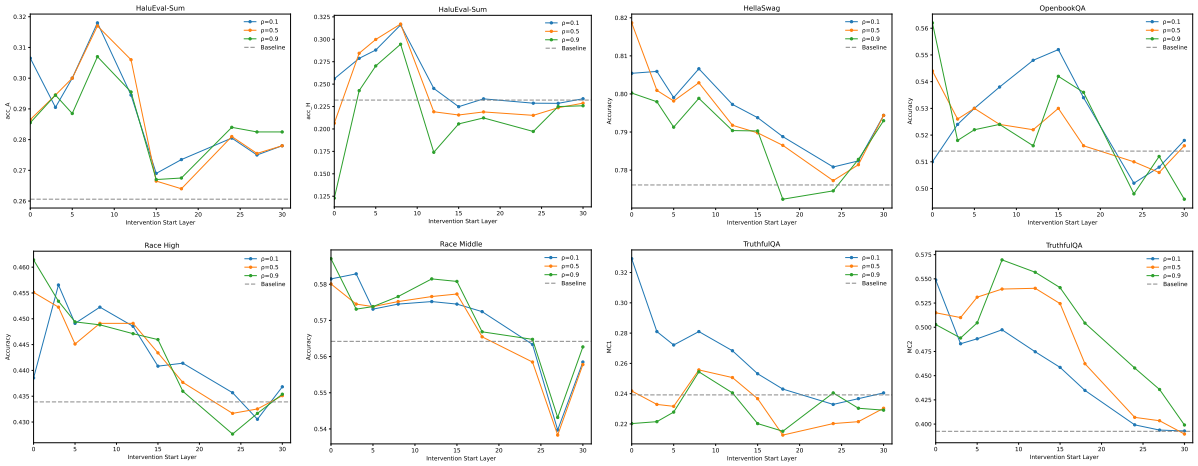

Figure 2: Performance of IFI-CD under different intervention start layers and generation-decision–dependent token selection ratios. The horizontal axis denotes the Transformer layer at which key-side information-flow intervention begins, while different curves correspond to different selection ratios of generation-decision–dependent tokens. For settings with $\rho \geq 0.5$, numerical stability constraints are applied to ensure that at least one unmasked key remains visible for each query row.

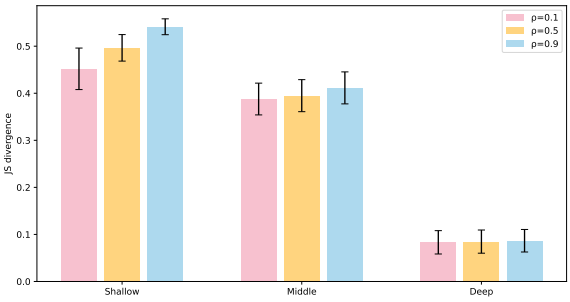

Figure 3: Magnitude of logits distributional perturbation under key-side information-flow intervention. Different bars correspond to different dependency token selection ratios ρ. Error bars indicate the standard error (SE) across different Transformer layers within each intervention start-layer interval.

ble changes occur in the output distribution, while constraints on generation-decision-dependent tokens further determine whether such changes form directional biases.

3.3.2 Additional Analysis

Beyond the analyses above, Appendices B and C further examine the effects of token selection mechanisms, sparsity, and attention aggregation strategies. The results show that generation-decision-dependent token selection outperforms random selection and low-dependency selection, and that high-dependency tokens at different structural positions play complementary roles in hallucination mitigation. In addition, cross-layer and cross-head attention aggregation provides more stable token rankings while maintaining performance.

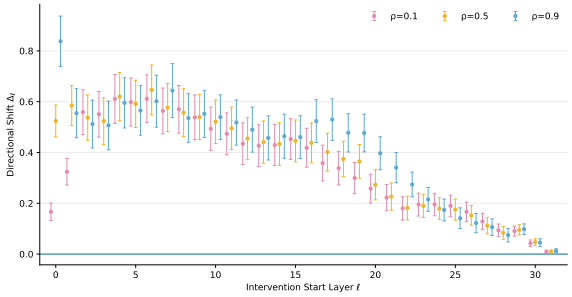

Figure 4: Directional shift of logits distributions under key-side information-flow intervention. Error bars indicate the SE of the sample-wise mean of Δ_ℓ, computed across samples for each intervention start layer ℓ.

4 Related Work

4.1 Other Hallucination Mitigation Methods

A broad range of approaches has been explored to address hallucinations in large language models, including training, external knowledge augmentation, and inference-time interventions. Training-based and external knowledge-based methods typically reduce unreliable generation through factuality-oriented supervised fine-tuning (Tian et al., 2024; Chen et al., 2024b), retrieval augmentation, or tool-assisted correction (Chern et al., 2023; Gao et al., 2023). In contrast, inference-time methods aim to enhance factuality without modifying model parameters. Among them, representation-level interventions guide generation by identifying attention

structures, critical layers, or representation directions associated with factuality and adjusting internal model states during inference (Li et al., 2023b; Zhang et al., 2024); another line directly regulates decoding distributions, for example by constraining or reweighting predictive probabilities according to uncertainty estimates or internal activation variation signals (Chen et al., 2024c; Wu et al., 2025).

4.2 Contrastive Decoding Methods

Contrastive decoding constructs paired distributions from a strong model and a weak model, and modulates relative preferences over candidate tokens during decoding (Li et al., 2023c). Its potential in inference tasks has also been systematically validated (O'Brien and Lewis, 2023). Follow-up studies mainly differ in how effective contrastive distributions are constructed. DoLa and ActLCD form inter-layer contrasts from outputs at different model layers (Chuang et al., 2024; Zhang et al., 2025a); PruneCD constructs a more informative amateur distribution through fine-grained layer pruning (Yu et al., 2025); CAD treats predictions with and without context as a natural contrastive pair (Shi et al., 2024); ICD constructs a factually weaker counterpart and penalizes its preferences during decoding (Zhang et al., 2025b); ID and SH2 derive contrastive signals from generation-path consistency and key-token attention, respectively (Cheng et al., 2025; Kai et al., 2024). Although these methods improve the effectiveness of contrastive decoding, prior studies have noted that coarse-grained contrasts and simple subtraction-based operations may disrupt the original output distribution of large language models (Chen et al., 2024a). Meanwhile, larger distributional perturbations do not necessarily correspond to effective shifts along factual directions.

Recent studies further construct more discriminative contrastive distributions by actively inducing hallucinated outputs. DeCoRe generates hallucination-prone outputs by intervening on retrieval heads (Gema et al., 2025), while HICD induces hallucinated distributions for contrastive decoding by dispersing attention over critical attention heads and weakening critical information integration (Jiang et al., 2025). In addition, PH3 analyzes knowledge conflicts from an information-flow perspective and mitigates erroneous information propagation by pruning conflicting attention heads (Jin et al., 2024), providing related evidence for understanding information-flow mechanisms in hallucination generation. These methods validate the effectiveness of the induce-and-suppress paradigm, but their induction strategies often rely on specific structural assumptions, additional training procedures, or task-dependent heuristics. They also rarely explicitly characterize which contextual evidence is attenuated during generation decisions or how such attenuation forms directional distributional shifts. In contrast, IFI-CD builds upon generation decision dependency modeling, induces evidence attenuation through structured information flow intervention, and leverages the resulting directional shifts for contrastive modulation.

5 Conclusion

This work presents IFI-CD, an information flow intervention-based contrastive decoding method that mitigates hallucinations from the perspective of generation decisions. IFI-CD characterizes dependency on contextual evidence and constructs directional decision signals through structured information flow intervention. Experiments and analyses show consistent gains across contextual consistency and factuality tasks, and indicate that the directionality of induced shifts is more critical than simply amplifying predictive distribution perturbations.

Limitations

The proposed IFI-CD framework primarily targets hallucinations related to contextual evidence integration at inference time, treating the dependency of generation decisions on contextual evidence as the entry point for intervention. Accordingly, IFI-CD guides predictive decisions by rebalancing the relative contributions of evidence already encoded within the model, and does not provide a mechanism for injecting new knowledge or correcting intrinsic knowledge deficiencies. Its effectiveness may therefore be limited for factual errors caused by missing knowledge, outdated information, or insufficient training data coverage. In addition, the generation-decision dependency characterized in this work reflects preference shifts during prediction rather than a causal explanation of the underlying generation mechanisms. A gap remains between generation-decision dependency and true causal relationships, leaving more rigorous theoretical characterization for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dingwei Chen, Feiteng Fang, Shiwen Ni, Feng Liang, Ruifeng Xu, Min Yang, and Chengming Li. 2024a. Lower layer matters: Alleviating hallucination via multi-layer fusion contrastive decoding with truthfulness refocused. *arXiv preprint arXiv:2408.08769*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024b. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, pages 1–31.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024c. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7553–7567. PMLR.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, Jian Jiao, Qi Chen, Peng CHENG, and Wayne Xiong. 2025. Integrative decoding: Improving factuality via implicit self-consistency. In *The Thirteenth International Conference on Learning Representations*, pages 1–46.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, pages 1–26.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2025. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Xinyan Jiang, Hang Ye, Yongxin Zhu, Xiaoying Zheng, Zikang Chen, and Jun Gong. 2025. Hicd: Hallucination-inducing via attention dispersion for contrastive decoding to mitigate hallucinations in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7764–7786, Vienna, Austria. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. 2024. Sh2: Self-highlighted hesitation helps you decode more truthfully. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4514–4530, Miami, Florida, USA. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, and Xi Yun. 2023. Chain of natural language inference for reducing large language model hallucinations. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, pages 1–18.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023c. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Buyun Liang, Liangzu Peng, Jinqi Luo, Darshan Thaker, Kwan Ho Ryan Chan, and Rene Vidal. 2025. SECA: Semantically equivalent and coherent attacks for eliciting LLM hallucinations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, pages 1–33.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *CoRR*, abs/2505.06708.

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 (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, pages 1–16.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang, and Longjun Cai. 2025. Improve decoding factuality by token-wise cross layer entropy of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3912–3921, Albuquerque, New Mexico. Association for Computational Linguistics.

Byeongho Yu, Changhun Lee, Jun-gyu Jin, and Eunhyeok Park. 2025. Prunecd: Contrasting pruned self model to improve decoding factuality. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32462–32473, Suzhou, China. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. 2025a. Active layer-contrastive decoding reduces hallucination in large language model generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3046, Suzhou, China. Association for Computational Linguistics.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2025b. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8218–8232, Albuquerque, New Mexico. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025c. siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

A Experimental Details

A.1 Datasets and Metrics

Evaluation is conducted along two dimensions: context-dependent hallucinations and factual hallucinations. The datasets and corresponding metrics are described below.

1) Context-dependent hallucinations

HellaSwag: HellaSwag (Zellers et al., 2019) is a situational context continuation task. Given a context and multiple candidate endings, the model is required to select the continuation that best matches the context. The validation split containing 10,042 examples is used for evaluation. Accuracy is reported as the evaluation metric.

RACE: RACE-M and RACE-H (Lai et al., 2017) are reading comprehension tasks corresponding to middle-school and high-school difficulty levels, respectively. Given a passage and a question, the model selects the correct answer from candidate options. The test split is used for evaluation, including 1,436 examples from RACE-Middle and 3,498 examples from RACE-High. Accuracy is reported separately on RACE-Middle and RACE-High.

OpenBookQA: OpenBookQA (Mihaylov et al., 2018) is an open-book question answering task. Each example contains a question, candidate answers, and an explicit reference fact field, making it suitable for evaluating answer selection under given evidence. The test split of the additional subset, containing 500 examples, is used for evaluation. Accuracy is reported as the evaluation metric.

HaluEval-Sum: HaluEval-Sum (Li et al., 2023a) is a summarization hallucination discrimination task. It contains both correct and hallucinated summaries, and evaluates whether the model can distinguish consistency between a summary and the source context. A total of 1,000 examples are used for evaluation. Acc-H and Acc-A are reported, where Acc-H measures hallucinated-summary identification and Acc-A denotes overall discrimination accuracy.

2) Factual hallucinations

TruthfulQA: TruthfulQA (Lin et al., 2022) is a factual question answering benchmark with both multiple-choice and open-ended generation settings. A total of 790 questions are used for evaluation. In the multiple-choice setting, MC1, MC2, and MC3 are reported. In the open-ended generation setting, Truth, Info, and Truth*Info are reported. Truth reflects the factual correctness of generated answers, Info reflects informativeness, and Truth*Info serves as a composite metric for overall performance across factuality and informativeness.

A.2 Baselines

Representative inference-time decoding methods that require no additional external knowledge are selected as baselines. The methods are described below.

Greedy Decoding: Greedy Decoding is the standard decoding strategy of the base model, selecting the token with the highest probability at each generation step.

DoLa: DoLa (Chuang et al., 2024) is an inter-layer contrastive decoding method. It contrasts logits from deeper and shallower Transformer layers to amplify factual signals that emerge in higher layers, thereby adjusting relative preferences over candidate tokens.

END: END (Wu et al., 2025) is a token-wise cross-layer entropy-based decoding method. It quantifies the factual knowledge required for each candidate token according to prediction probability changes across model layers, and adjusts the final predictive distribution to prioritize tokens with higher factuality.

PruneCD: PruneCD (Yu et al., 2025) is a layer-pruning-based contrastive decoding method. It constructs a self-amateur distribution by pruning a subset of Transformer layers and contrasts it with the expert distribution of the full model, yielding a more informative contrastive signal.

HICD: HICD (Jiang et al., 2025) is a hallucination-inducing contrastive decoding method. It induces context-related hallucinated distributions at the attention-head level and contrasts them with the original predictive distribution during decoding.

A.3 Implementation Details and Hyperparameters

Experiments are conducted with LLaMA-7B and LLaMA-2-7B as backbone models (Touvron et al., 2023), both of which contain approximately 7B parameters. All methods operate without updating model parameters, and only intervene in or adjust the decoding process at inference time. The inference experiments are conducted on NVIDIA RTX 4090 GPUs with 48GB memory. The approximate runtime for IFI-CD evaluation on different datasets is reported in Table 3.

Table 3: Approximate computational cost of IFI-CD evaluation on NVIDIA RTX 4090 GPUs.

Dataset	Runtime (GPU hours)
HellaSwag	0.73
RACE-Middle	0.22
RACE-High	0.57
OpenBookQA	0.05
TruthfulQA-Gen	0.38
TruthfulQA-MC	0.25
HaluEval-Sum	1.10

The implementation is based on PyTorch 2.9.1 and Hugging Face Transformers 4.28.1. The Transformers library is modified to support the information-flow intervention used by IFI-CD during inference, including layer-sensitive key-side visibility masking in self-attention.

For IFI-CD, the selection ratio of generation-decision-dependent tokens is fixed at $\rho = 0.1$. The intervention start layer is chosen from the candidate set $\{0, 3, 5, 8, 12\}$, which covers representative shallow-to-middle layers for examining the effect of intervention timing on model behavior. The candidate range of the contrastive strength coefficient α is set according to task type: $\{0.6, 0.8, 1.0, 1.2, 1.4\}$ for HellaSwag and RACE, and $\{1.6, 1.8, 2.0, 2.2, 2.4\}$ for OpenBookQA and HaluEval-Sum.

B Token Selection Mechanism and Sparsity Analysis

To further examine the role of different tokens in IFI-CD's hallucination mitigation process, model behavior is analyzed from the perspectives of token selection mechanisms and sparsity. HaluEval-Sum is adopted as the evaluation benchmark, as its longer contexts provide a more stable basis for assessing the effects of different token intervention strategies.

B.1 Comparison of Token Selection Strategies

Three token selection strategies are compared: generation-decision-dependent tokens based on Section 2.1, randomly selected tokens (Random), and tokens with low dependency strength (Low-Dependency). All three strategies share the same token selection ratio ($\rho = 0.1$) and intervention start layer (layer 8), differing only in the mechanism for selecting masked tokens.

As shown in Fig. 5, different selection strategies lead to clearly different hallucination discrimination performance. The generation-decision-dependent token strategy outperforms Random and Low-Dependency in hallucinated-sample identification and achieves the best overall results on Acc-A and Acc-H, indicating that IFI-CD is sensitive to how masked tokens are selected. Meanwhile, this strategy shows slightly lower accuracy on correct samples than some baselines, reflecting a rebalancing of the model's original decision boundary: attenuating information sources highly relied upon by generation decisions reduces overconfidence on correct summaries and improves discrimination of hallucinated summaries.

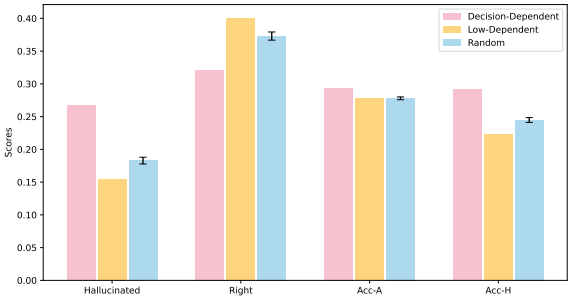

Figure 5: Comparison of hallucination mitigation performance across different token selection strategies. Results for the Random strategy are reported as the mean and SE over multiple random seeds.

B.2 Sparsity Analysis

To further examine the structural characteristics of generation-decision dependent tokens, the ranked distribution of attention weights is analyzed. As reported in Table 4, the highest-attention tokens are typically associated with the first token in the sequence, a pattern consistent with the attention sink (Qiu et al., 2025) phenomenon described in prior work. In addition, the self-query token corresponding to the generation-decision point frequently appears among the top-ranked positions, reflecting the intrinsic tendency of the self-attention mechanism to perform self-alignment of the current position during generation decisions. Beyond these two categories of structurally high-dependency tokens, a small number of tokens originating from internal context positions also consistently occupy leading ranks in the attention ordering.

Based on the above observations, generation-decision-dependent tokens are categorized by structural position into three groups: the first token of

Table 4: Distribution of high-dependency token types across different benchmarks.

Token Type	HaluEval-Sum		HellaSwag	OpenBookQA	RACE	
	Acc_H	Acc_A	Acc	Acc	Middle	High
FT	0.8086		0.9997	1.0000	0.9903	0.9946
SQT	0.9472		0.7819	0.7200	0.6943	0.6115

the sequence (FT), the self-query token (SQT), and high-dependency tokens from internal context positions (IDTs). An incremental intervention scheme is then adopted, progressively increasing the number of masked high-dependency tokens. A structural ablation control is also constructed by excluding FT and SQT and applying incremental masking only to IDTs, enabling analysis of the relative contributions of different token types.

As shown in Fig. 6, different incremental and structural ablation settings exhibit a clear stage-wise trend. Under the full setting, intervening only on FT, and then jointly on FT and SQT, already leads to substantial changes in model performance; after the intervention scope expands to IDTs, performance continues to improve, but the gains gradually diminish and tend to saturate. In contrast, the structural ablation setting based only on IDTs brings certain gains but consistently falls short of the full setting. Overall, dependency tokens at different structural positions play complementary roles in hallucination mitigation: structurally ranked tokens near the top of the attention ordering help rapidly shift the model's decision direction, while IDTs further consolidate and refine model behavior, with their effect saturating within a small quantity range.

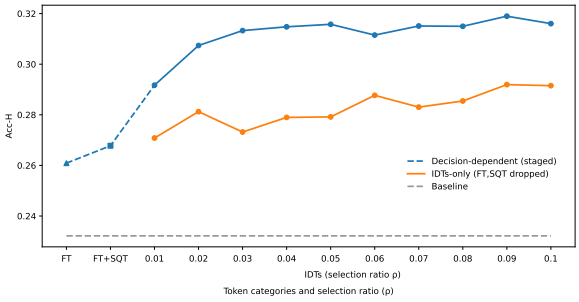

Figure 6: Results of incremental intervention and structural ablation on generation-decision–dependent tokens. Under the full setting, intervention is applied sequentially to FT, SQT, and IDTs, whereas under the structural ablation setting, intervention is applied exclusively to IDTs.

C Stability and Effectiveness of Attention Aggregation Strategies

This section examines how different attention aggregation strategies affect token selection stability and hallucination mitigation performance. The analysis first compares the robustness of token rankings under different layer/head aggregation schemes. With the token selection ratio fixed at 10%, random perturbations are applied to attention heads, and the Jaccard similarity between the Top-10% token sets before and after perturbation is computed. The compared strategies include single-layer aggregation, partial-head aggregation, and cross-layer cross-head averaging; partial-head aggregation is based on heuristic selection according to the entropy and skewness of attention distributions.

Stability results are reported in Table 5. Cross-layer and cross-head averaging achieves the highest mean Jaccard similarity and the smallest standard deviation, indicating the strongest robustness under random perturbations. In contrast, aggregation schemes based on a single layer or a subset of attention heads exhibit greater variability across perturbations.

Table 5: Comparison of generation-decision–dependent token selection stability under different attention aggregation strategies. The table reports token selection stability measured by Jaccard@10% under random perturbations for each aggregation strategy, with the perturbation ratio set to 0.6.

Aggregation Strategy	Jaccard@10%	
	Mean	Std
Entropy–Skewness Head Selection	0.8300	0.0272
Lowest Layer	0.8760	0.0201
Middle Layer (L=8)	0.6963	0.0726
Highest Layer	0.8022	0.0332
Full Aggregation (Ours)	0.9241	0.0163

Building on the stability analysis, the effectiveness of different aggregation strategies in hallucination mitigation is further examined. As shown in Fig. 7, cross-layer and cross-head averaging

achieves Acc-H performance comparable to that of stronger middle-layer aggregation strategies, with only minor overall differences across strategies. This indicates that global attention aggregation does not noticeably weaken hallucination mitigation effectiveness while providing more stable token rankings. Overall, cross-layer and cross-head averaging achieves a better balance between stability and effectiveness, and avoids additional manual layer or attention-head selection.

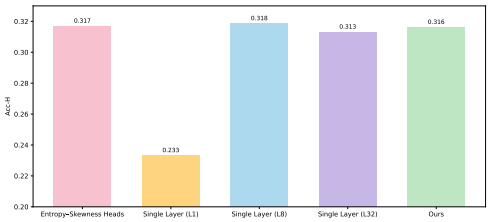

Figure 7: Comparison of hallucination mitigation performance (Acc-H) under different attention aggregation strategies.