

## Article

# ISFJ-RAG: Interventional Suppression of Hallucinations via Counter-Factual Joint Decoding Retrieval-Augment Generation

Yuezhao Liu <sup>1</sup>, Wei Li <sup>1,2,\*</sup> , Yijie Wang <sup>1</sup>, Ningtong Chen <sup>1</sup> and Min Chen <sup>3,4</sup> 

<sup>1</sup> School of Information Engineering, Nanchang University, Nanchang 330031, China; liuyuezhao@email.ncu.edu.cn (Y.L.); 419100240132@email.ncu.edu.cn (Y.W.); chenningtong@email.ncu.edu.cn (N.C.)

<sup>2</sup> School of Artificial Intelligence, Nanchang University, Nanchang 330031, China

<sup>3</sup> School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China; minchen@ieee.org

<sup>4</sup> Pazhou Laboratory, Guangzhou 510640, China

\* Correspondence: liwei.cs@ncu.edu.cn

## Abstract

Although retrieval-augmented generation (RAG) technology mitigates the hallucination issue in large language models (LLMs) by incorporating external knowledge, and combining reasoning models can further enhance RAG system performance, retrieval noise and attention bias still lead to the diffusion of factual errors in problems such as factual queries, multi-hop questions, and unanswerable questions. Existing methods struggle to effectively suppress “high-confidence hallucinations” in long-chain reasoning due to their failure to decouple knowledge bias effects from causal reasoning paths. To address this, this paper proposes the ISFJ-RAG framework, which dynamically intervenes in hallucinations through counterfactual joint decoding. First, a structural causal model (SCM) reveals three root causes of hallucinations in RAG systems: irrelevant knowledge interference, reasoning path bias, and spurious correlations in self-attention mechanisms. A dual-decoder architecture is further designed: the total causal effect decoder models the global relationship between user queries and knowledge, while the knowledge bias effect decoder captures potential biases induced by external knowledge. A dynamic modulation module converts the latter’s output into a proxy measure of hallucination bias. By computing individual treatment effects (ITEs), the bias component is removed from the full generation distribution, achieving simultaneous suppression of knowledge-irrelevant and reasoning-irrelevant hallucinations. Ablation experiments validate the robustness of average token log-probability as a confidence metric. Experiments demonstrate that on the RAGEval benchmark, ISFJ-RAG improves generation completeness to 86.89% (+5.49%) while reducing hallucination rates to 10.39% (−2.5%) and irrelevance rates to 4.44% (−2.99%).

**Keywords:** hallucination; RAG; causal inference; counterfactual reasoning; LLM



Academic Editor: Francesco Archetti

Received: 8 December 2025

Revised: 31 January 2026

Accepted: 3 February 2026

Published: 9 February 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

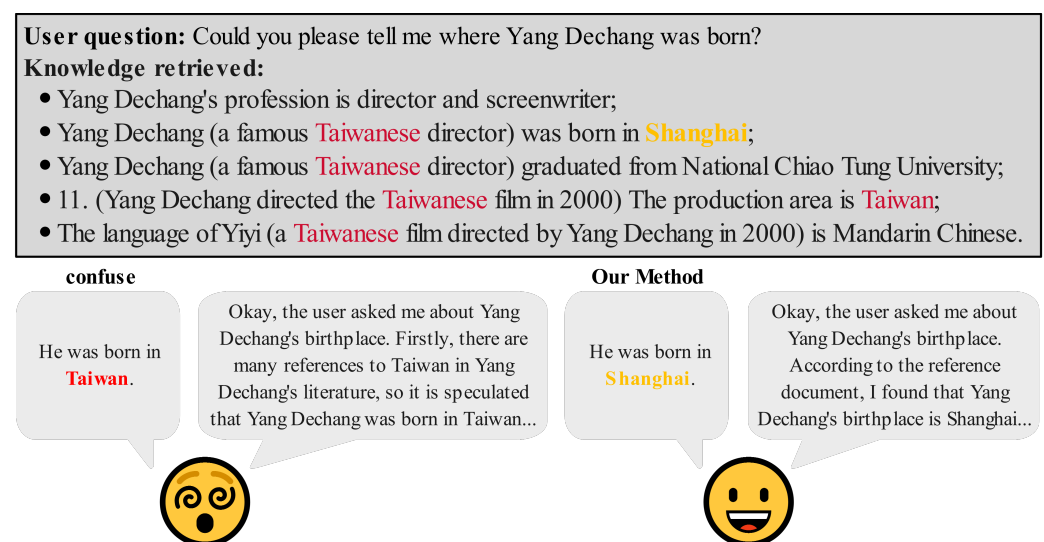
With the rapid development of artificial intelligence technology, large language models (LLMs) have demonstrated powerful reasoning capabilities and outstanding performance in various natural language processing (NLP) tasks, thanks to their massive parameters and extensive pre-training [1,2]. However, the knowledge of LLMs primarily relies on the data from the pre-training phase, which cannot cover all domains of knowledge. When faced with domains not sufficiently covered by the training corpus [3], these models may

still generate semantically coherent but factually incorrect responses [4,5], a phenomenon collectively referred to as hallucinations in large language models.

The recently proposed DeepSeek R1 and other long-chain reasoning models have demonstrated profound capabilities in complex domains such as mathematics and coding [6,7]. However, empirical studies indicate that when the reasoning chain length exceeds a critical threshold [8], the models are prone to fall into the “over-reasoning” trap, leading to deviations in logical paths and accumulation of factual errors [9]. RAG [10] refers to retrieving relevant information from external knowledge bases before using LLMs to answer questions and combining the retrieved external knowledge to respond to user queries, but existing systems still face challenges such as retrieval noise interference.

The reliability of the system can be improved through RAG technology combined with long-reasoning models [11]. However, even in RAG systems utilizing long-reasoning models, the incidence of hallucinations remains significant when dealing with complex problems [12], and due to the enhancement of model performance, hallucinations may become even harder to detect. Combining previous research [13,14], we argue that the hallucination issue in RAG tasks combined with long-chain reasoning models may stem from the influence of external knowledge noise and reasoning process noise. Additionally, in the Transformer [15] architecture, the model employs a self-attention mechanism to compute weighted representations of the input sequence, aiming to capture dependencies between different elements. However, the calculation of attention scores relies on interactions among all elements in the input sequence [16,17], and when external knowledge noise and reasoning process noise are introduced, these interactions may introduce bias.

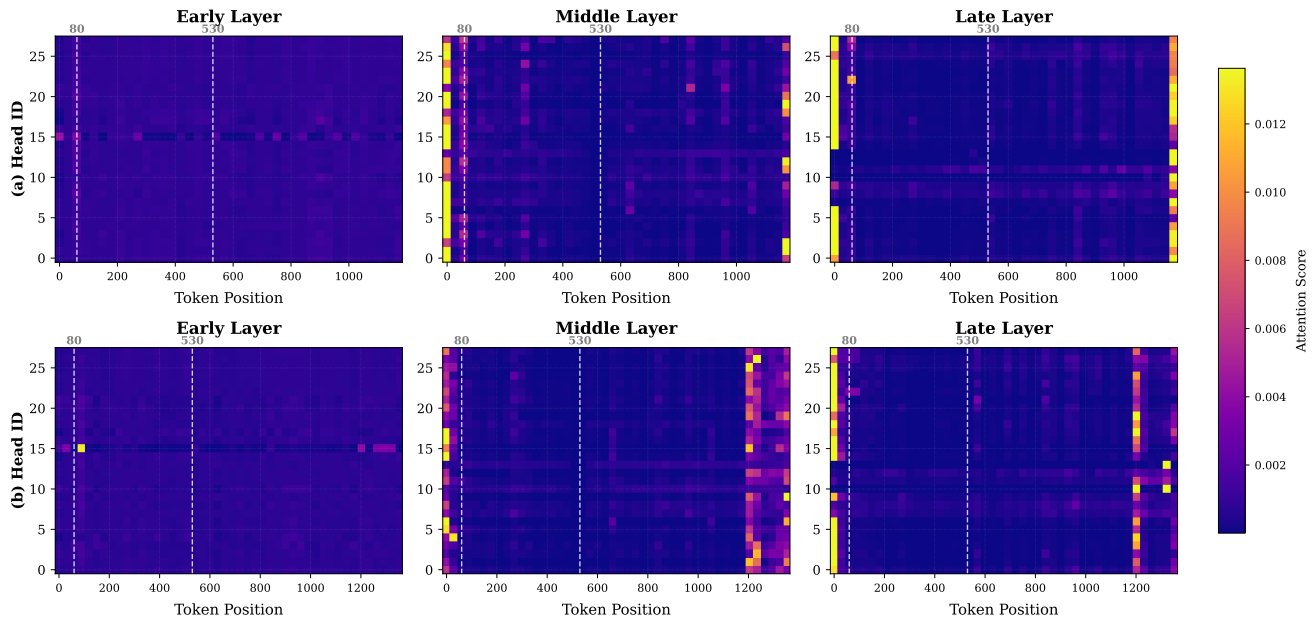
An example of hallucination in large language models is shown in Figure 1, where the retrieved information contains descriptions about Edward Yang and Taiwan. Due to the presence of excessive noise, the model conflates these details and produces incorrect, irrelevant reasoning, ultimately leading to hallucinated output.



**Figure 1.** An example of hallucination in LLM. In retrieved knowledge, Yang Dechang and Taiwanese are frequent co-occurrences leading to pseudo-relevance; the model fails to recognize this irrelevance.

Figure 2 illustrates the attention distribution across the reasoning (a) and response generation (b) phases, where the latter includes the generated Chain-of-Thought context. Beyond the typical attention sinks at initial tokens [18], we observe a general sparsity of knowledge-related attention. Critically, the causal mechanism of hallucination is isolated in phase (b) by monitoring specific “copy heads” (e.g., L26 H26) identified via [19] during a multi-entity query on EMS adoption. As shown in the comparison, a decisive contrast

explains the failure: while the model effectively anchors to the ground truth “Nov 2017” (index ~80) for the correct Sky Aviation response, it fails to attend to the relevant evidence “Dec 2021” at index ~530 for Sunshine Media. This “attention failure” leads the model to hallucinate “Nov 2021” based on adjacent spurious correlations, confirming that the model ignores available external knowledge in favor of flawed internal priors.



**Figure 2.** (a) Attention visualization of “copy heads” (e.g., L26, H26) identified via [19] for the query regarding EMS adoption dates across the reasoning phases; (b) Attention visualization across the response generation phases. Vertical lines mark the ground-truth tokens: “Nov 2017” for Sky Aviation (idx ~80) and “Dec 2021” for Sunshine Media (idx ~530). A critical contrast appears in (b): the model successfully activates attention at ~80 to retrieve the correct date for Sky Aviation. However, for Sunshine Media, it fails to attend to the ground truth “Dec 2021” at ~530 (dark region), causing it to hallucinate “Nov 2021” based on adjacent spurious context.

To address the aforementioned issues, we propose the ISFJ-RAG framework (Interventional Suppression of Hallucinations via counter-Factual Joint Decoding for Retrieval-Augmented Generation), which dynamically suppresses hallucinations through counterfactual joint decoding. Specifically, ISFJ-RAG incorporates a dual-decoder architecture:

- **Fully Causal Effect Decoder:** Its input includes both external knowledge and user questions, designed to model the holistic causal relationship between user questions and external knowledge;
- **Knowledge Bias Effect Decoder:** Its input contains only external knowledge, aimed at generating latent bias representations related to external knowledge.

Based on this, we designed a dynamic adjustment module according to the previous idea [20,21], adjusting the output of the second decoder to serve as a proxy that maximizes the representation of hallucination bias effects. Subsequently, we compute the individual treatment effect (ITE) based on the counterfactual reasoning framework to obtain the final token distribution after subtracting the hallucination bias effect. This process of subtracting the two decoders continues until the model completes its response, so the input to the knowledge-biased effect decoder also includes the model’s reasoning process. Therefore, our architecture can simultaneously mitigate hallucinations caused by knowledge-irrelevant and reasoning-irrelevant factors.

The contributions of this paper are as follows:

1. The paper systematically analyzes the main causes of hallucination issues in reasoning models for RAG tasks from a casual perspective and proposes the hypothesis that hallucinations primarily stem from low relevance among user queries, retrieved knowledge, and reasoning processes, as well as biases introduced by self-attention mechanisms during the encoding phase.
2. We design a dual-decoder framework ISFJ-RAG based on counterfactual reasoning which could effectively reduce hallucinations, and we open-source the related code at <https://github.com/1784669877/code/>.
3. We conduct extensive experiments on the RAGEval benchmark, and validate that ISFJ-RAG we proposed is superior to the existing methods from various evaluation metrics. Specifically, ISFJ-RAG improves generation completeness to 86.89% (+5.49%), while reducing hallucination rates to 10.39% (−2.5%) and irrelevance rates to 4.44% (−2.99%).

## 2. Related Work

Temperature sampling [22] influences the creativity and faithfulness of the output by controlling the randomness of the generation process, but excessively low temperatures can significantly suppress diversity. To address this issue, Chang et al. [23] have proposed KL-Divergence Guided Temperature Sampling, which employs a dual-decoder architecture to generate probability distributions based on inputs with and without background knowledge, respectively. The method dynamically adjusts the temperature parameter by quantifying the impact of background knowledge on the generation results through the KL divergence between the two distributions, thereby optimizing the trade-off between diversity and knowledge faithfulness. Additionally, research based on the Copying Heads theory has introduced the AARF framework [19,24], which dynamically intervenes in the outputs of copying heads and Knowledge FFNs to regulate the model's reliance on external versus internal knowledge.

Causal reasoning [25] and counterfactual thinking provide new methodological [26] support for the optimization of RAG systems. Sun et al. [27] have effectively reduced the dependency of generative models on potentially biased knowledge by constructing a Structural Causal Model (SCM) framework, using retrieved knowledge as an intervention variable. Addressing the unique characteristics of RAG systems, researchers have further proposed a knowledge debiasing mechanism based on counterfactual learning. A representative work is the two-stage decoding architecture designed by [28], which first filters low-confidence retrieval results through a relevance discriminator and then employs a counterfactual generator to simulate the output distribution under unbiased knowledge conditions, significantly enhancing the objectivity of generated content.

Leveraging a dual-decoder architecture to mitigate hallucinations in large language models is a research direction. Trustful LLMs [29] introduce a parallel decoding framework, where the Guided Context Decoder interacts with the prompt decoder through a Cross-Attention mechanism to enforce the model to strictly adhere to the retrieved knowledge. However, this method has a strong dependence on the quality of the knowledge source and lacks robustness. Similarly, the Lower Layer Matters framework [30] filters non-factual tokens by comparing the distributions of the base model and the fine-tuned “amateur model” and fusing the outputs of the final layer and lower layers. Nevertheless, this method requires manual debugging, and there is subjectivity and uncertainty in the selection of lower layers. In addition, these methods are often accompanied by significant deployment overhead.

In research on mitigating hallucinations in large language models within RAG systems, existing work mainly focuses on two aspects: retrieval enhancement and post-generation

verification. On the retrieval side, knowledge graphs [31] methods or hybrid retrieval strategies that integrate BM25 and Embedding are typically used to improve retrieval accuracy at both semantic and literal levels. On the post-processing side, the focus is on evaluation and correction; for example, the Re-KGR [32] method performs triple verification by identifying high-risk entities and combining them with domain knowledge graphs. In contrast, our proposed ISFJ-RAG method is based on a counterfactual reasoning framework and eliminates hallucination biases through automatic parameter adjustment. A major advantage of this method is its plug-and-play capability, allowing for deployment without the need for model fine-tuning, and it can seamlessly integrate with the aforementioned retrieval optimization and post-processing technologies, working synergistically to maximize the suppression of hallucinations.

### 3. Methodology

#### 3.1. The Structural Causal Model of KGD

As shown in Figure 3, we construct a structural causal model (SCM) [33] for the reasoning model-based RAG system. This model granularly decomposes the question-answering process into four core variable nodes, delineating the causal path from user queries to reasoning and responses in the reasoning model-based RAG system. Specifically, KGD formalizes the current task as follows:

**Node Q:** Node Q represents the natural language query initiated by the user.

**Node C:** We define C as the retrieved external knowledge and reasoning process input to the large language model, denoted as  $C = K, R$ , where K represents the external knowledge relevant to the current dialog topic, and  $R = r_{k-1}^{|K|}$  denotes the model's output of the first  $k - 1$  tokens in the current round of dialog.

**Node H:** This node describes the vectorization process of text within the model. The model inputs the text into a generative pre-trained model, where the input text is internally transformed into  $H_1, \dots, H_n$ , with  $n$  being the number of input tokens.

**Node P:** Node P implements the mapping from the representation space to the vocabulary space. Specifically, the model takes nodes Q and C as inputs, processes them through the internal hidden representation node H, and ultimately generates a probability distribution  $P(t_k | H_1, H_2, \dots, H_n)$ , where  $t_k$  represents the next token predicted by the model at the current step.

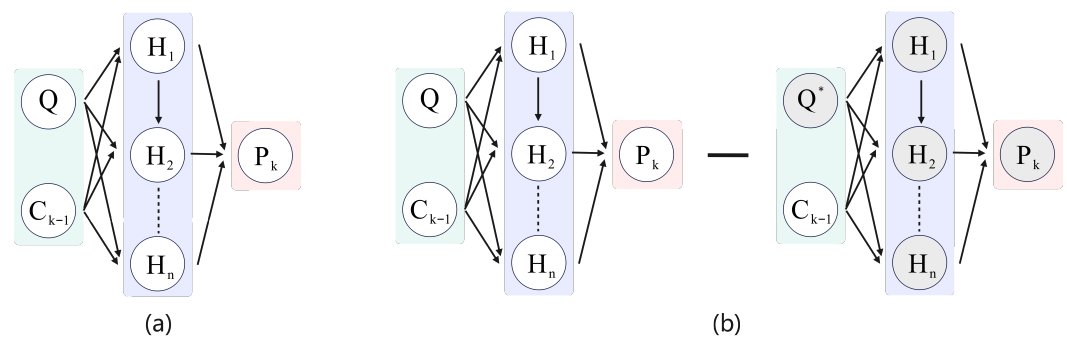
**Q→H, C→H and H→H:** The paths Q→H and C→H represent the encoding process of the model; H→H denotes the transmission of hidden states within the model. In these processes, issues such as knowledge noise and high-frequency word bias can lead to hallucinatory outputs. Additionally, we find that Q and C, as confounding variables [33], introduce spurious correlations in the encoding process of H, and this interactive computation further exacerbates hallucinations.

**H→P:** This path achieves a probabilistic mapping from the semantic space to the lexical space. Since the hidden state H may contain hallucinatory content introduced by biases or spurious correlations in the training data, such content can be erroneously amplified during decoding.

In summary, we propose three hypotheses:

1. **Irrelevance of external knowledge:** During the knowledge retrieval phase, retrieval algorithms based on similarity matching (such as BM25 or vector retrieval) [34] may retrieve irrelevant noisy content.
2. **Irrelevance of the reasoning process:** When noisy knowledge is injected into the reasoning pipeline, the reasoning model may deviate from the problem itself during deep reasoning. When answering questions, the model tends to overly focus on reasoning that may have biases.

3. **Limitations of the self-attention:** In the Transformer architecture, the model captures dependencies between different elements through the self-attention mechanism. However, there exist spurious correlations among different hidden representations.



**Figure 3.** (a) The causal graph of KGD; (b) Counterfactual reasoning: Traditional KGD (left) minus counterfactual KGD (right), where the gray nodes represent counterfactual variables.

### 3.2. Counterfactual Reasoning Analysis

Counterfactual reasoning focuses on events that could have occurred under certain conditions but did not actually happen. Assuming the individual causal effect we need to estimate is  $ITE$ , we construct two models for the indirect estimation of individual causal effects:  $f_0(X)$  and  $f_1(X)$ , where  $f_0(X)$  is used to predict the outcome  $\hat{Y}(0)$  under the intervention  $T = 0$ , and  $f_1(X)$  is used to predict the outcome  $\hat{Y}(1)$  under the control  $T = 1$ . For each individual  $i$ , the output results of the two models serve as the indirect estimation of the individual causal effect:  $ITE_i = \hat{Y}_i(1) - \hat{Y}_i(0)$  [35].

In summary, following the approach of counterfactual reasoning, we define two distinct models within the SCM for the KGD task. Let  $f_0(\cdot) = P(p|Q, C)$  denote the factual outcome used to calculate the Total Effect (TE) that encompasses both valid reasoning and potential hallucinations.

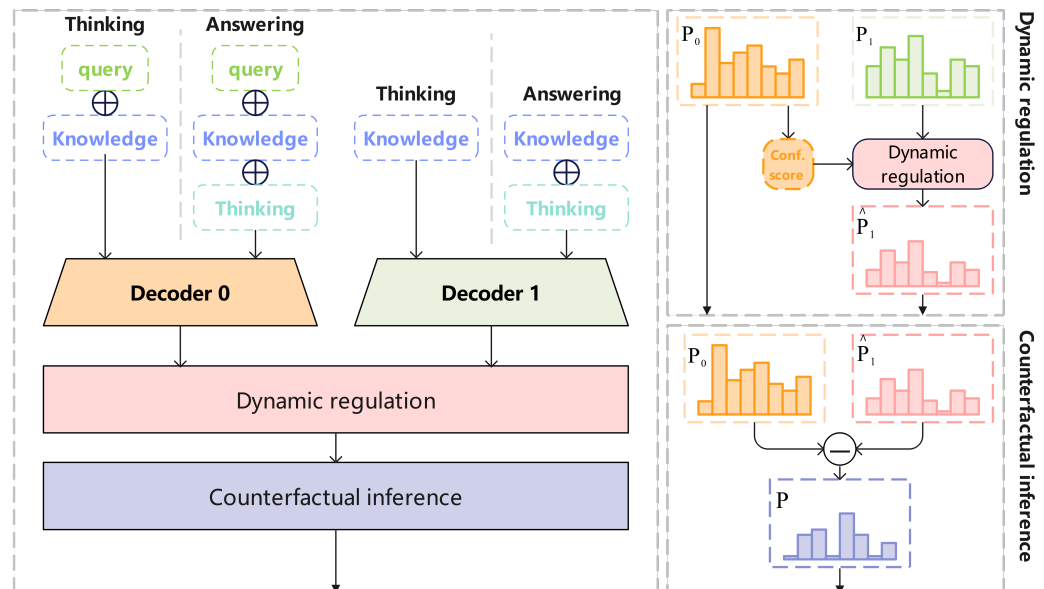
In contrast, we define the interventional model as  $f_1(\cdot) = P(p|do(Q = Q^*), C)$ , where  $Q^*$  denotes a ‘nearly empty’ state. From a causal perspective, the intervention  $do(Q = Q^*)$  effectively severs the direct causal path from the user’s specific intent ( $Q$ ) to the model’s response ( $P$ ). Consequently, the output of  $f_1(\cdot)$  is no longer driven by the query’s semantics but is dominated by the spurious correlations inherent in the context  $C$  and the model’s prior biases. Thus,  $f_1$  serves as an estimator for the Spurious Effect (SE), representing the hallucination noise we aim to isolate.

Our goal is to recover the Direct Causal Effect (DCE), which reflects the true contribution of  $Q$  to the response, conceptually defined as  $DCE \approx TE - SE$ . A straightforward approach is to subtract the output of  $f_1(\cdot)$  from  $f_0(\cdot)$ . However, direct subtraction in the probability space often leads to distribution confusion [36], where the operation might not only remove the SE but also mistakenly diminish the valid DCE. To address this issue, we introduce a dynamic adjustment mechanism designed to regulate the estimated SE from  $f_1(\cdot)$  into a proxy value that best represents the hallucination bias effect. Specifically, we define the following probability distributions:

- **Factual probability  $P_0$  (Total Effect):** The probability distribution generated by the model under the complete influence of user queries, external knowledge, and reasoning processes. This generation process includes both the valid DCE and the hallucination-prone SE.
- **Bias effect  $P_1$  (Spurious Effect):** A probability distribution estimated by the interventional decoder, which prominently includes the SE derived solely from external knowledge and reasoning processes without the guidance of the user query.

- **Target probability  $P$  (Direct Causal Effect):** The distribution obtained by eliminating SE to retain only DCE:  $P = P_0 - \hat{P}_1 = f_0(\cdot) - g(f_1(\cdot))$ , where  $g(\cdot)$  can adjust the output of  $P_1 = f_1(\cdot)$  to a proxy value that best represents the effect of hallucination bias.

As shown in Figure 4, after analyzing hallucination elimination strategies from a counterfactual causal perspective, we propose the ISFJ-RAG framework. Decoder 0 takes user queries, external knowledge, and reasoning processes as input, while Decoder 1 only receives external knowledge and reasoning processes as input.



**Figure 4.** Model framework: During the inference phase, Decoder 1 takes user queries and knowledge as input, while Decoder 2 takes knowledge as input. Counterfactual reasoning is employed to mitigate interference from external knowledge. In the response phase, Decoder 1 takes user queries, knowledge, and reasoning processes as input, while Decoder 2 takes knowledge and reasoning processes as input. The entire reasoning process is dynamically adjusted through a dynamic confidence adjustment module.

### 3.3. Dynamic Regulator Mechanism

The goal of the dynamic regulator is to adjust the probability distribution of the model's output, reducing the bias effect caused by irrelevant external knowledge while preserving the positive causal contribution of the dialog context. To achieve this, we define a dynamic parameter  $\alpha$  obtained from numerical calculations based on  $f_0(\cdot)$ , which can dynamically adjust the output of  $f_1(\cdot)$  based on the model's confidence level, thereby adjusting the output of  $P_1 = f_1(\cdot)$  to a proxy value that best represents the effect of hallucination bias. The confidence level reflects the model's uncertainty about its output, and we use the following confidence metrics proposed in previous work [37–39]:

**Average Token Log-Probability:** Calculate the average negative logarithmic probability of each token in the sentence, is given by

$$\text{Average}(p) = -\frac{1}{L_i} \sum_{j=1}^{L_i} \log(p_{ij}) \quad (1)$$

where  $L_i$  is the number of tokens in sentence  $i$ , and  $p_{ij}$  is the conditional probability of the  $j$ -th token.

**Entropy:** Measure the uncertainty of the probability distribution for each token, defined as the maximum entropy across all tokens in the sentence, given by

$$\text{Entropy} = - \sum_{w \in D} p_{ij}(w) \log p_{ij}(w) \quad (2)$$

where  $j$  represents the  $j$ -th token in the generation process.  $D$  is the dictionary of all possible words, and  $w$  is a word in  $D$ .

**Maximum Token Log-Probability:** Capture the token with the lowest conditional probability in the sentence, given by

$$\text{Maximum}(p) = \max_j (-\log(p_{ij})) \quad (3)$$

where  $p_{ij}$  is the conditional probability of the  $j$ -th token.

**Max Probability:** Measure the highest conditional probability among all tokens in the sentence, given by

$$\text{Max Probability} = \max_j p_{ij} \quad (4)$$

where  $p_{ij}$  is the conditional probability of the  $j$ -th token.

**Perplexity:** Represent the exponential of the average negative logarithmic probability of the tokens in the sentence, given by

$$\text{Perplexity}(p) = \exp \left( -\frac{1}{L_i} \sum_{j=1}^{L_i} \log(p_{ij}) \right) \quad (5)$$

where  $L_i$  is the number of tokens in sentence  $i$ , and  $p_{ij}$  is the conditional probability of the  $j$ -th token.

For the metrics Average Token Log-Probability, Entropy, Perplexity, and Maximum Token Log-Probability, since they quantify uncertainty, we invert the uncertainty values to obtain the confidence score. The confidence score is defined as

$$\text{Conf} = \frac{1}{1 + \text{Uncertainty}} \quad (6)$$

where Uncertainty represents the respective uncertainty metric (e.g., Average ( $p$ ), Entropy, Perplexity ( $p$ ), or Maximum ( $p$ )).

We define the dynamic adjuster  $g(\cdot)$  as

$$g(f_1(\cdot)) = \alpha \cdot f_1(\cdot) \quad (7)$$

where  $\alpha$  is a dynamic parameter whose value is adjusted based on the model's confidence level. Specifically, in our dual-decoder framework, to regulate  $f_1(\cdot)$  based on the output of  $f_0(\cdot)$ ,  $\alpha$  is calculated as

$$\alpha = 1 - \beta \cdot \text{Conf}(f_0(\cdot)) \quad (8)$$

where  $\beta$  represents the scaling factor.  $\text{Conf}(f_0(\cdot))$  represents the confidence level of the probability distribution generated by the model under the complete action of user questioning, external knowledge, and reasoning process, we use it to regulate the output distribution of encoder 1.

### 3.4. Algorithm Implementation

Directly using logits for subtraction may lead to generation instability, as the subtraction operation on logits is equivalent to the division operation on probabilities, which

can excessively amplify tokens predicted by the bias model with extremely low probabilities. Therefore, we perform the subtraction operation in the probability space to ensure numerical stability. Specifically, we first normalize the outputs of the two decoders using Softmax, then weight the bias distribution  $P_1$  using the dynamic coefficients calculated by Equation (8), and finally select tokens based on greedy search. The complete decoding process is shown in Algorithm 1.

---

**Algorithm 1** ISFJ-RAG Decoding Algorithm (Probability Space)
 

---

**Require:** User query  $Q$ , retrieved context  $C$ , model  $\mathcal{M}$ , max length  $T$ , scaling factor  $\beta$

**Ensure:** Generated response  $y$

```

1: Initialize generated sequence  $y_{<t} \leftarrow \emptyset$ 
2: while  $t < T$  and  $y_{t-1} \neq \text{EOS}$  do
3:   1. Dual forward pass (Based on SCM)
4:    $L_0 \leftarrow \mathcal{M}(y_{<t} \mid Q, C)$  {Logits for  $f_0(\cdot)$ : Query + Knowledge}
5:    $L_1 \leftarrow \mathcal{M}(y_{<t} \mid \emptyset, C)$  {Logits for  $f_1(\cdot)$ : Knowledge only (Counterfactual)}
6:   2. Normalize to probability space
7:    $P_0 \leftarrow \text{Softmax}(L_0)$  {Factual probability}
8:    $P_1 \leftarrow \text{Softmax}(L_1)$  {Bias effect distribution}
9:   3. Dynamic regulation
10:  Uncertainty  $\leftarrow \text{CalculateMetric}(P_1)$  {e.g., Entropy of  $f_1(\cdot)$ }
11:  Conf  $\leftarrow 1/(1 + \text{Uncertainty})$  {Confidence score}
12:   $\alpha \leftarrow 1 - \beta \cdot \text{Conf}$  {Dynamic parameter calculation}
13:  4. Bias mitigation
14:   $\hat{P}_1 \leftarrow \alpha \cdot P_1$  {Adjusted bias proxy value}
15:   $S_{\text{final}} \leftarrow P_0 - \hat{P}_1$  {Target probability  $P$ }
16:  5. Greedy selection
17:   $y_t \leftarrow \text{argmax}(S_{\text{final}})$  {Select token with highest score}
18:   $y_{<t+1} \leftarrow y_{<t} \oplus y_t$ 
19: end while
20: return  $y$ 

```

---

### 3.5. Illustrative Example

To intuitively demonstrate the effectiveness of ISFJ-RAG in mitigating hallucinations caused by retrieval noise, we present a specific case study analyzing the probability changes during the decoding process.

#### Input Scenario:

- **User Query (Q):** “When did Sky Aviation Ltd. and Sunshine Media Group Ltd. adopt the environmental management system?”
- **Retrieved Context (C):** “... In November 2021, the company optimized logistics. ... Finally, in December 2021, the company adopted the environmental management system...”
- **Standard RAG Failure:** “Sunshine Media Group adopted the system in November 2021.” (Confusing the logistics date with the adoption date).

#### Step-by-Step Analysis of ISFJ-RAG Strategy:

Phase 1: Total Effect Estimation (Decoder 0:  $Q + C$ )

The standard decoder attends to the context. However, due to the proximity of “November” and “December” in the text, the model becomes confused. Since “November” appears earlier and is semantically prominent in the preceding sentence, Decoder 0 assigns a relatively high probability ( $P_0 \approx 0.48$ ) to the token “November” when generating the critical month token, while the correct token “December” receives a lower probability ( $P_0 \approx 0.20$ ). This leads to a potential hallucination.

### Phase 2: Bias Estimation (Decoder 1: $\emptyset + C$ )

In this phase, we input the retrieved context  $C$  into the counterfactual decoder while masking the specific query  $Q$ . Operationally, this means providing the model with the context followed by the standard prompt structure (e.g., “Context: ... \n User Question: ”), but leaving the query slot explicitly empty. Without the semantic constraint of the specific question (“When... adopted system?”), Decoder 1 acts as a language model predicting based purely on context inertia. Due to recency bias and the tendency to repeat salient terms, the bias model assigns a very high probability ( $P_1 \approx 0.62$ ) to “November”. In this unconstrained scenario, the support for “December” is significantly lower ( $P_1 \approx 0.10$ ).

### Phase 3: Causal Intervention ( $P_{final} \propto P_0 - \alpha P_1$ )

Our mechanism subtracts the bias distribution from the total distribution to isolate the reasoning path.

- **For “November”:** The high bias score indicates that this token is driven by retrieval noise rather than query-based reasoning. The subtraction operation significantly penalizes it ( $0.48 - \alpha \cdot 0.62 \rightarrow$  Suppressed).
- **For “December”:** Although its initial probability was lower, the bias support is also minimal. The subtraction preserves its score relative to the penalized token ( $0.20 - \alpha \cdot 0.10 \rightarrow$  Boosted).

As shown in Table 1, after normalization, “December” becomes the top-ranked token. Consequently, the final output is correctly generated as “Sunshine Media Group... in December 2021...”.

**Table 1.** Trace of token probabilities for the critical month generation in the case study.

| Candidate Token         | $P_0(Q, C)$<br>(Standard RAG) | $P_1(\emptyset, C)$<br>(Context Bias) | Operation<br>( $P_0 - \alpha \cdot P_1$ ) | Final Result            |
|-------------------------|-------------------------------|---------------------------------------|-------------------------------------------|-------------------------|
| November(Hallucination) | 0.48 (Dominant)               | 0.62 (Noise)                          | $0.48 - \text{Penalty} \rightarrow$ Low   | $\downarrow$ Suppressed |
| December (Ground Truth) | 0.20 (Suppressed)             | 0.10 (Low Bias)                       | $0.20 - \text{Minor} \rightarrow$ High    | $\uparrow$ Selected     |
| Others (Long Tail)      | 0.32                          | 0.28                                  | -                                         | -                       |

## 4. Experimental Setup

### 4.1. Dataset

In terms of evaluation methodology, we adopted the DragonBall benchmark dataset proposed by RAGEval [40], which encompasses various question types including factual, multi-hop, and unanswerable questions. RAGEval assesses answer quality through three key dimensions: completeness, hallucination, and irrelevance. We randomly sampled 100 entries from each domain, resulting in a total of 300 evaluation samples. To specifically target the hallucination phenomenon in Retrieval-Augmented Generation (RAG) systems, our experiment bypasses the retrieval stage analysis and directly employs the BGE-Large embedding model, as recommended by the paper, for document retrieval (TopK = 5). Based on the keypoints extracted from ground truth answers, we define the following core metrics, where  $A$  denotes the generated answer,  $K = \{k_1, k_2, \dots, k_n\}$  represents the set of keypoints, and  $I[\cdot]$  serves as the indicator function:

- **Completeness:** Measures the proportion of keypoints covered by the generated answer, reflecting the comprehensiveness of information:

$$\text{Comp}(A, K) = \frac{1}{|K|} \sum_{i=1}^n I[A \text{ covers } k_i], \quad (9)$$

where  $I[A \text{ covers } k_i] = 1$  if and only if the answer correctly includes keypoint  $k_i$ .

- **Hallucination Rate:** Detects the proportion of contradictions between the generated answer and the keypoints to identify fabricated content:

$$\text{Hallu}(A, K) = \frac{1}{|K|} \sum_{i=1}^n I[A \text{ contradicts } k_i], \quad (10)$$

where  $I[A \text{ contradicts } k_i] = 1$  if and only if there is a semantic conflict between the answer and  $k_i$ .

- **Irrelevance:** Evaluates the proportion of keypoints that are neither covered nor contradicted, measuring the degree of deviation in the answer:

$$\text{Irr}(A, K) = 1 - \text{Comp}(A, K) - \text{Hallu}(A, K). \quad (11)$$

Considering the significant impact of the reasoning process on generating answers, we designed an evaluation framework for the model's reasoning process in RAG tasks. By analyzing the correlation between the model's thought process  $T$  and  $R$  (references),  $K$  (keypoints), we proposed three fine-grained metrics:

- **Problem Focus Degree:** Measures whether the reasoning strictly revolves around the problem requirements, avoiding irrelevant reasoning:

$$\text{Focus}(T, Q) = I[T \text{ aligns with } Q], \quad (12)$$

where  $I[\cdot] = 1$  if and only if  $T$  is directly related to the question  $Q$ .

- **Citation Accuracy:** Evaluates whether the reasoning is based on the reference document content without introducing external assumptions or fictional content:

$$\text{Accuracy}(T, R) = I[T \text{ quoted } R], \quad (13)$$

where  $I[\cdot] = 1$  if and only if all  $R$  are correctly cited during the reasoning process  $T$ .

- **Key Point Coverage Degree:** Detects whether the reasoning comprehensively addresses all annotated keypoints:

$$\text{Coverage}(T, K) = I[K \text{ addressed in } T], \quad (14)$$

where  $I[\cdot] = 1$  if and only if all annotated keypoints  $K$  are explicitly handled during the reasoning process  $T$ .

Experiments utilize RAGEval via DeepSeek-v3. To ensure statistical robustness, we randomly sample 300 instances. Significance against baselines is verified using paired  $t$ -tests ( $p < 0.05$ ).

#### 4.2. Baseline

In this study, we selected the 1.5 B and 7.0 B models of DeepSeek-R1-Distill-Qwen as the primary experimental subjects. To evaluate our dynamic causal dual decoder approach, we used Temperature Sampling as the baseline and compared it with sampling strategies and other decoder strategies, as follows:

- **Temperature:** Sampling with a temperature of 0.6 and top-p of 0.95.
- **DoGe [41]:** A dynamic sampling method that switches decoding strategies based on the model's factual confidence.
- **KL-divergence sampling [23]:** A dynamic temperature sampling method guided by KL divergence.

- **AARF [19]:** By intervening in attention heads and FFN modules, it dynamically adjusts the attention weights of the copy heads to external contexts and the parameter knowledge contribution of knowledge FFNs to the residual stream during generation.
- **ISFJ-RAG (Ours):** Dynamic confidence adjustment causal dual decoder based on counterfactual reasoning. In the 1.5B model,  $\beta$  is set to 1, and in the 7B model,  $\beta$  is set to 1.25. We will explain the reasons for hyperparameter selection in Section 5.3.

## 5. Results and Discussion

### 5.1. Evaluation of Hallucination

This paper adopts the average token log-probability as the confidence parameter (see ablation experiments in Section 5.3) and evaluates generation quality from three dimensions: completeness, hallucination rate, and irrelevance. We report results as mean  $\pm$  standard deviation over three random seeds (42, 123, 1337) to ensure statistical significance. For the ablation experiment, we used a random seed of 42 to extract 300 data entries. As shown in Table 2, the ISFJ method achieves optimal performance in both the 1.5B/7B dual-base models: in the 7B model, completeness (85.17%) improves by 5.49 percentage points compared to the baseline RAG, while hallucination rate (10.39%) and irrelevance (4.44%) decrease by 2.5 and 2.99 percentage points, respectively, and ROUGE-L (0.2506) significantly outperforms other methods. This indicates that the framework based on dual decoders and counterfactual reasoning can effectively suppress knowledge noise and enhance generation reliability through the bias effect elimination mechanism.

**Table 2.** Answer quality comparison of various methods on different models (1.5B and 7B). Metrics include: completeness rate (Comp.), hallucination rate (Hallu.), irrelevance rate (Irrel.), and ROUGE-L. The best results are highlighted in bold.

| Model | Method      | Comp.                   | Hallu.                  | Irrel.                  | ROUGE-L                  |
|-------|-------------|-------------------------|-------------------------|-------------------------|--------------------------|
| 1.5B  | Temperature | 65.12 $\pm$ 1.4%        | 17.78 $\pm$ 0.9%        | 17.10 $\pm$ 1.1%        | 0.1867 $\pm$ 0.02        |
|       | AARF        | 67.27 $\pm$ 1.2%        | <b>14.42</b> $\pm$ 0.8% | 18.31 $\pm$ 1.3%        | 0.1765 $\pm$ 0.02        |
|       | DoGe        | 66.40 $\pm$ 1.5%        | 15.12 $\pm$ 1.0%        | 18.48 $\pm$ 1.4%        | 0.1702 $\pm$ 0.01        |
|       | KL          | 47.89 $\pm$ 2.1%        | 25.63 $\pm$ 1.8%        | 26.48 $\pm$ 1.9%        | 0.1570 $\pm$ 0.02        |
|       | ISFJ        | <b>71.49</b> $\pm$ 0.7% | 16.97 $\pm$ 0.9%        | <b>11.54</b> $\pm$ 0.6% | <b>0.2294</b> $\pm$ 0.01 |
| 7B    | Temperature | 79.68 $\pm$ 1.7%        | 12.89 $\pm$ 1.3%        | 7.43 $\pm$ 0.9%         | 0.2386 $\pm$ 0.02        |
|       | AARF        | 82.75 $\pm$ 1.1%        | <b>10.34</b> $\pm$ 0.6% | 6.91 $\pm$ 0.4%         | 0.2355 $\pm$ 0.01        |
|       | DoGe        | 79.89 $\pm$ 1.2%        | 12.35 $\pm$ 0.9%        | 7.76 $\pm$ 0.7%         | 0.2334 $\pm$ 0.02        |
|       | KL          | 71.10 $\pm$ 1.9%        | 15.27 $\pm$ 1.4%        | 13.63 $\pm$ 1.5%        | 0.2193 $\pm$ 0.03        |
|       | ISFJ        | <b>85.17</b> $\pm$ 1.4% | 10.39 $\pm$ 1.0%        | <b>4.44</b> $\pm$ 0.8%  | <b>0.2506</b> $\pm$ 0.01 |

Compared to other methods, although AARF ranks second in model integrity (82.75%) for the 7B model, it is sensitive to hyperparameters and overly relies on the “copy head” mechanism, which may amplify noise in the knowledge base (e.g., 18.31% irrelevance for the 1.5B model). DoGe suffers from logical discontinuities due to frequent switching of decoding strategies (ROUGE-L of only 0.1702 for the 1.5B model), and its KL temperature adjustment strategy conflicts with the optimal inference model temperature (0.6) [6], resulting in a comprehensive lag in key metrics.

### 5.2. Evaluation of Reasoning

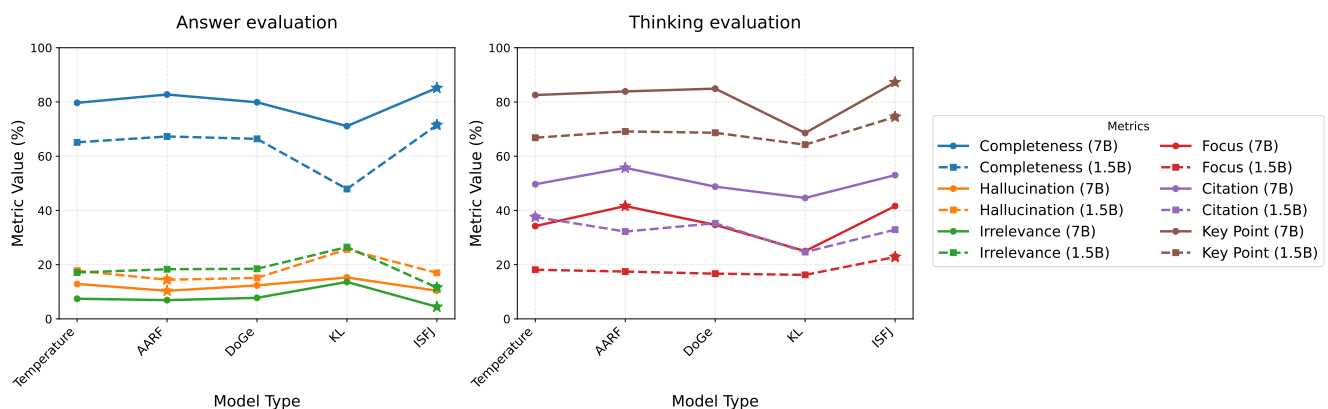
For the effectiveness of the reasoning process in QA tasks, this paper constructs a ternary evaluation system comprising question focus, citation accuracy, and key point coverage. Table 3 show that ISFJ achieves dual optimal performance in the 7B model with a question focus of 41.61% and key point coverage of 87.25% (improving by 7.38 and 4.69 percentage points over the baseline RAG, respectively).

**Table 3.** Thinking process evaluation of various methods on different models (1.5B and 7B). Metrics include: Problem Focus Degree (Focus), Citation Accuracy (Accu.), and Key Point Coverage Degree (Cov.). The best results are highlighted in bold.

| Model | Method      | Focus                   | Accu.                   | Cov.                    |
|-------|-------------|-------------------------|-------------------------|-------------------------|
| 1.5B  | Temperature | 18.12 $\pm$ 1.4%        | <b>37.58</b> $\pm$ 1.6% | 66.81 $\pm$ 1.5%        |
|       | AARF        | 17.45 $\pm$ 1.1%        | 32.21 $\pm$ 1.3%        | 69.13 $\pm$ 1.4%        |
|       | DoGe        | 16.68 $\pm$ 1.2%        | 35.26 $\pm$ 1.5%        | 68.64 $\pm$ 1.3%        |
|       | KL          | 16.23 $\pm$ 1.9%        | 24.63 $\pm$ 1.8%        | 64.24 $\pm$ 2.1%        |
|       | ISFJ        | <b>22.82</b> $\pm$ 0.7% | 32.89 $\pm$ 0.9%        | <b>74.50</b> $\pm$ 0.8% |
| 7B    | Temperature | 34.23 $\pm$ 1.3%        | 49.66 $\pm$ 1.5%        | 82.56 $\pm$ 1.1%        |
|       | AARF        | <b>41.61</b> $\pm$ 0.9% | <b>55.70</b> $\pm$ 1.0% | 83.89 $\pm$ 1.2%        |
|       | DoGe        | 34.67 $\pm$ 1.1%        | 48.77 $\pm$ 1.3%        | 84.92 $\pm$ 1.2%        |
|       | KL          | 24.98 $\pm$ 1.7%        | 44.60 $\pm$ 1.6%        | 68.56 $\pm$ 1.9%        |
|       | ISFJ        | <b>41.61</b> $\pm$ 0.6% | 53.02 $\pm$ 0.8%        | <b>87.25</b> $\pm$ 0.5% |

It is worth noting that regarding citation accuracy, ISFJ (53.02%) shows a slight decrease compared to AARF (55.70%). We attribute this to ISFJ's suppression of excessive citation—the baseline model has a tendency to mechanically add citation tags even to common or redundant phrases. A qualitative analysis (see Appendix A for details) confirms that the higher citation metrics in AARF often reflect “citation noise” rather than better grounding. ISFJ focuses on semantic correctness and removes these redundant structural tags to ensure that the answers are factually accurate and reliable, indicating that a lower citation rate does not mean lower factual completeness.

Additionally, as shown in the Figure 5, it can be observed that the metrics of Answer and Thinking are generally positively correlated, with a calculated Pearson correlation coefficient of 0.7179. This aligns with our second hypothesis regarding the causes of hallucinations in reasoning models, namely that hallucinations are inherently linked to the reasoning process.



**Figure 5.** The evaluation results of the 7B model.

In summary, the reasoning advantage of ISFJ stems from its two-stage correction mechanism: the first-stage decoder reduces noise bias through external knowledge comparison, while the second-stage decoder performs causal reasoning based on contexts where noise bias effects have been mitigated.

### 5.3. Ablation Experiment

We systematically evaluate the performance of five confidence metrics (Average Token Log-Prob., Perplexity, Entropy, Max Log-Prob., Max) on 1.5B and 7B models. Additionally, to isolate the confidence component's contribution, we perform ablation studies comparing

fixed weighting (conf = 0 to conf = 1.0) versus dynamic weighting with different confidence metrics. This comparison clarifies that the ‘conf = 0’ setting represents a direct static subtraction baseline, whereas other confidence values demonstrate the efficacy of progressive dynamic weighting.

As shown in Table 4, Average Token Log-Prob. still maintains the best performance on the 7B model, with the highest completeness (85.17%) among all dynamic confidence metrics and the lowest hallucination rate (10.39%). Notably, Average Log Prob demonstrates significant advantages in two core metrics, likely because Average Token Log-Prob. directly corresponds to the cross-entropy loss function during model training [15]. This alignment ensures consistency between the evaluation criteria and optimization objectives. Moreover, the current task is challenging with mixed external knowledge. Relying solely on local confidence measures like Max may yield poor results, while the Average Token Log-Prob. serves as a global output confidence metric, better capturing overall information. It is also noteworthy that a confidence score of conf. = 0.8 exhibits performance comparable to dynamic metrics in both the 1.5B and 7B models.

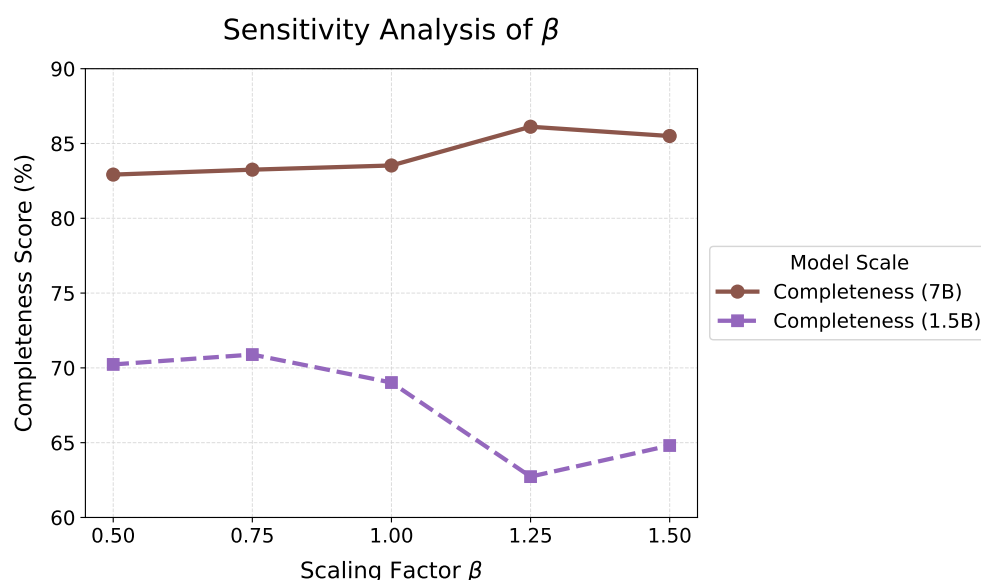
**Table 4.** Confidence calibration performance across model sizes. Lower values are preferred for hallucination and irrelevance metrics. When the confidence level is fixed at 0, according to Equation (8), it is equivalent to directly subtracting the two probability distributions without any processing. The best results are highlighted in bold.

| Metric                  | 1.5B Model    |               |               |               | 7B Model      |              |              |               |
|-------------------------|---------------|---------------|---------------|---------------|---------------|--------------|--------------|---------------|
|                         | Comp.         | Hallu.        | Irrel.        | ROUGE-L       | Comp.         | Hallu.       | Irrel.       | ROUGE-L       |
| Average Token Log-Prob. | <b>71.49%</b> | <b>16.97%</b> | <b>11.54%</b> | 0.2294        | <b>85.17%</b> | 10.39%       | <b>4.44%</b> | <b>0.2506</b> |
| Perplexity              | 64.16%        | 20.81%        | 15.03%        | <b>0.2350</b> | 80.02%        | 13.86%       | 6.12%        | 0.2211        |
| Entropy                 | 65.54%        | 19.15%        | 15.32%        | 0.2246        | 83.77%        | 11.32%       | 4.92%        | 0.2487        |
| Maximum Log-Prob.       | 68.83%        | 17.27%        | 13.90%        | 0.2284        | 82.93%        | 10.10%       | 6.97%        | 0.2491        |
| Max                     | 64.86%        | 20.89%        | 14.25%        | 0.2222        | 79.53%        | 14.48%       | 5.98%        | 0.2469        |
| conf. = 0               | 42.13%        | 31.48%        | 26.39%        | 0.1189        | 56.64%        | 25.48%       | 17.88%       | 0.1675        |
| conf. = 0.2             | 67.13%        | 17.38%        | 15.49%        | 0.2097        | 80.55%        | 12.00%       | 7.45%        | 0.2424        |
| conf. = 0.4             | 66.03%        | 19.85%        | 14.13%        | 0.1939        | 79.06%        | 14.23%       | 6.71%        | 0.2479        |
| conf. = 0.6             | 68.75%        | 18.29%        | 12.96%        | 0.1945        | 81.93%        | 13.13%       | 4.94%        | 0.2480        |
| conf. = 0.8             | 69.16%        | 17.60%        | 13.24%        | 0.2108        | 83.39%        | <b>9.98%</b> | 6.63%        | 0.2462        |
| conf. = 1.0             | 65.12%        | 17.78%        | 17.10%        | 0.1867        | 79.68%        | 12.89%       | 7.43%        | 0.2386        |

The ROUGE-L metric of Perplexity exhibits significant fluctuations, indicating that its performance is highly influenced by model scale. This may be due to the fact that as the number of parameters increases, the model capacity exceeds task requirements, leading to increased diversity in generated text and consequently reducing literal overlap with reference texts. Further analysis reveals that the Max metric, due to its over-reliance on local extremes, shows significantly lower completeness in the 1.5B model (64.86%) and the 7B model (79.53%) compared to the average. Although Max Log Prob. demonstrates lower irrelevance in the 1.5B model (13.90%), its irrelevance in the 7B model (6.97%) is notably higher than the average, suggesting its fluctuating effectiveness in hallucination control with model scale, potentially making it more suitable for smaller models. Finally, the baseline without confidence constraints (Direct) still performs the worst, confirming the necessity of confidence mechanisms.

Regarding the selection of the scaling factor  $\beta$ , our decision is based on empirical observations. As shown in Table 4, a confidence level of approximately 0.8 achieves the optimal balance between suppressing hallucinations and ensuring the completeness of responses. To further justify the theoretical basis for  $\beta$ , we conducted a sensitivity analysis using the dataset and evaluation framework defined in (Section 4.1). As visualized in

Figure 6, we observe a clear correlation between model size and the optimal  $\beta$ . The 1.5B model achieves peak completeness at a lower value ( $\beta = 0.75$ ), while the 7B model performs best at a higher value ( $\beta = 1.25$ ). This phenomenon can be explained by intrinsic model confidence. Larger models (like 7B) tend to be more “confident” in their predictions (producing sharper probability distributions). Therefore, a larger  $\beta$  is required to apply enough force to correct their bias. Conversely, smaller models (1.5B) are less confident and more sensitive; using a large  $\beta$  on them risks over-penalizing valid information, leading to the sharp drop in completeness shown in the figure.



**Figure 6.** Sensitivity analysis of the scaling factor  $\beta$  on response completeness across model sizes. The 1.5B model (dashed purple) peaks at a lower  $\beta$  ( $\approx 0.75$ ), whereas the 7B model (solid brown) requires a higher  $\beta$  ( $\approx 1.25$ ) to achieve optimal performance. This confirms that the penalty strength should be scaled according to the model’s intrinsic confidence.

Based on our sensitivity analysis, we propose a rule of thumb for selecting  $\beta$  on new architectures:

$$\beta \approx 1.0 + 0.25 \times \log_2 \left( \frac{\text{Parameter Size (B)}}{1.5(B)} \right) \quad (15)$$

Additionally, for optimal performance, we recommend evaluating  $\beta$  values within the range  $[0.5, 1.5]$  on a small validation set (e.g., 10% of RAGEval [40]).

Overall, the results from dual-decoder counterfactual reasoning outperform those of traditional RAG, and even fixed-value intervention achieves effects close to dynamic metrics under specific parameters (e.g.,  $\text{conf.} = 0.8$ ), validating the effectiveness of our overall architecture. Moreover, thanks to its inherent connection with the loss function, Average Token Log-Prob., through dynamic adjustment, further enhances the accuracy of proxy values for bias effects.

#### 5.4. Generalization Analysis

To verify the robustness of ISFJ under different models, we extended our evaluation beyond DeepSeek to include Llama-3.1, Mistral-7B-v0.3 and Qwen2.5-7B-Instruct. We report results as mean  $\pm$  standard deviation over three random seeds (42, 123, 1337) to ensure statistical significance. Performance was assessed on RAGEval and a sampled subset (300 instances) of the RAGTruth dataset [13], utilizing a DeepSeek-based judge.

As shown in Table 5, ISFJ consistently outperforms Standard RAG across all four model families: **(1) Model Generalization:** ISFJ yields significant gains not only on reasoning-

specific models but also on general-purpose models (Qwen) and those with distinct attention mechanisms (Mistral). For instance, Mistral sees a 3.4% reduction in hallucination rate on RAGEval. **(2) Statistical Rigor:** The improvements are statistically stable across seeds, confirming that the benefits are not artifacts of specific initializations. On RAGTruth, ISFJ effectively suppresses hallucination rate (HR) across the board, notably reducing the HR of Llama-3.1 by over 5 percentage points.

**Table 5.** Robustness analysis across four model families. ISFJ consistently improves hallucination rate (HR) on RAGTruth and reduces hallucinations (Hallu.) on RAGEval across diverse attention architectures. The best results are highlighted in bold.

| Model                    | Method   | RAGEval (DragonBall)   |                        |                       | RAGTruth               |
|--------------------------|----------|------------------------|------------------------|-----------------------|------------------------|
|                          |          | Comp.                  | Hallu.                 | Irrel.                | HR                     |
| Llama-3.1<br>8B-Instruct | Std. RAG | 80.34 $\pm$ 1.4        | 9.48 $\pm$ 0.8         | 10.17 $\pm$ 1.1       | 56.67 $\pm$ 2.1        |
|                          | ISFJ     | <b>86.66</b> $\pm$ 1.2 | <b>6.75</b> $\pm$ 0.5  | <b>6.60</b> $\pm$ 0.9 | <b>51.33</b> $\pm$ 1.8 |
| Mistral<br>7B-v0.3       | Std. RAG | 77.15 $\pm$ 1.6        | 13.20 $\pm$ 1.1        | 9.65 $\pm$ 1.3        | 68.90 $\pm$ 1.3        |
|                          | ISFJ     | <b>82.40</b> $\pm$ 1.3 | <b>9.80</b> $\pm$ 0.9  | <b>7.80</b> $\pm$ 1.0 | <b>59.33</b> $\pm$ 1.2 |
| Qwen2.5<br>7B-Instruct   | Std. RAG | 76.25 $\pm$ 1.9        | 14.85 $\pm$ 1.5        | 8.90 $\pm$ 1.4        | 84.33 $\pm$ 2.1        |
|                          | ISFJ     | <b>81.35</b> $\pm$ 1.5 | <b>12.15</b> $\pm$ 1.2 | <b>6.50</b> $\pm$ 1.1 | <b>81.93</b> $\pm$ 1.6 |
| DeepSeek<br>R1-Distill   | Std. RAG | 79.68 $\pm$ 1.7        | 12.89 $\pm$ 1.3        | 7.43 $\pm$ 0.9        | 83.74 $\pm$ 1.9        |
|                          | ISFJ     | <b>85.17</b> $\pm$ 1.4 | <b>10.39</b> $\pm$ 1.0 | <b>4.44</b> $\pm$ 0.8 | <b>79.93</b> $\pm$ 1.5 |

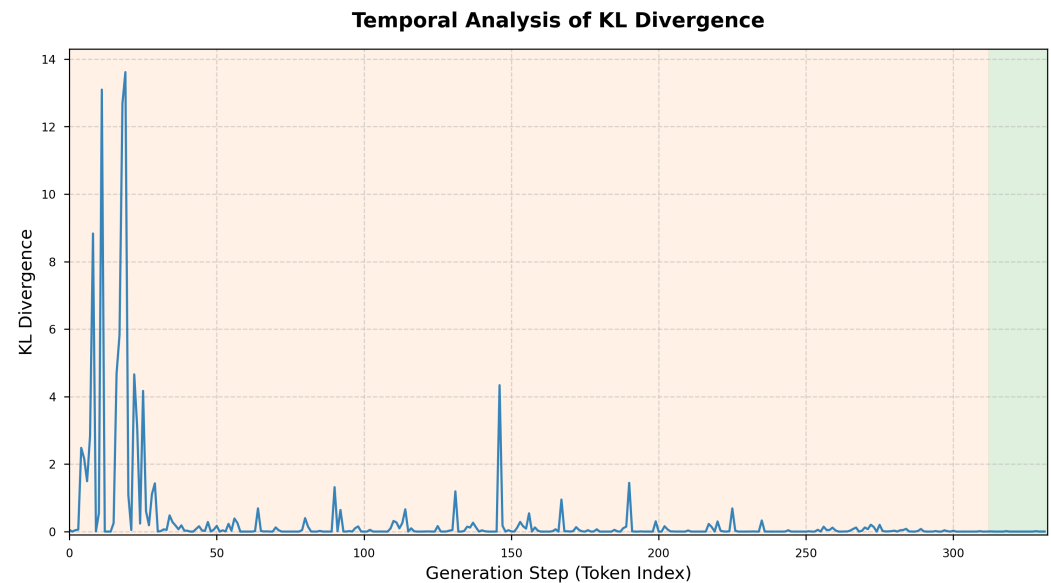
### 5.5. Analysis of Distribution Shift and Intervention Dynamics

To investigate potential distributional shifts between the main decoder  $f_0(\cdot)$  and the knowledge-only counterfactual decoder  $f_1(\cdot)$ , we performed a temporal analysis of the Kullback–Leibler (KL) divergence during the generation process, defined as  $D_{\text{KL}}(P(f_0) \parallel P(f_1))$ .

As illustrated in Figure 7, this divergence exhibits a distinct spike-and-stabilize pattern:

- **Early-Stage High Divergence:** In the initial generation steps (approximately 0–40 tokens), we observe a sharp peak in KL divergence (peak  $> 10$ ). This indicates that, due to the lack of specific query constraints,  $f_1$  diverges significantly from  $f_0$  when the model is performing critical reasoning and path selection. This confirms that our subtraction mechanism actively suppresses “knowledge-biased” paths when they conflict with query-specific reasoning.
- **Low Divergence Phase:** During the remainder of the generation process (40+ tokens), the KL divergence drops and stabilizes at a very low level ( $< 1.0$ ). Empirically, this suggests that for the majority of text generation, both decoders operate within a comparable semantic space, thereby ensuring syntactic consistency.

This analysis validates that the subtraction operation in Equation (8) does not introduce fundamental distributional mismatches; rather, it functions as a targeted intervention mechanism that plays a critical role only when necessary.



**Figure 7.** The blue curve represents  $D_{KL}(f_0||f_1)$  at each generation step, the x-axis represents the generation time steps (token indices), and the divergence is calculated on the final output probability distributions. The high divergence in the early stage is because the beginning is the initial stage of generation, and the model is still determining the sentence structure. The low divergence in the later stage indicates that there is no systematic distribution shift between the two decoders.

### 5.6. Inference Efficiency Analysis

In this chapter, we analyze the theoretical and practical costs of the dual-decoder architecture. Since  $f_0$  and  $f_1$  share the same backbone parameters, they can be batched during the inference process. Therefore, in terms of time cost, with the help of the KV caching mechanism, under parallel settings, the inference latency is comparable to that of the single-decoder benchmark model. However, because parallel computing requires the deployment of two independent large language models, the GPU memory usage is approximately twice that of the standard RAG.

For actual deployment, we recommend that ISFJ be specifically used for high-risk question-answering scenarios (such as medical or legal advice), where the importance of accuracy outweighs twice the time or space costs. In terms of hardware, assuming FP16 precision is adopted, deploying a 7B model with ISFJ requires approximately 28 GB of video memory, while a standard RAG requires about 14 GB.

## 6. Conclusions

In this paper, we study how external knowledge and self-attention affect hallucinations in large language models (LLMs) from a causal perspective. We introduce a method that uses counterfactual reasoning to dynamically adjust the decoder's output, reducing hallucinations without altering the LLM architecture. Experiments on hallucination benchmarks show that this approach effectively mitigates hallucinations while maintaining generation quality and diversity.

Regarding the trade-off issue of citation accuracy mentioned in Section 5.2, future iterations of ISFJ can explore “selective masking”—applying the subtraction mechanism only to semantic markers while bypassing structural citation markers—to suppress hallucinations while preserving the citation format.

**Author Contributions:** Conceptualization, Y.L., W.L. and M.C.; Methodology, Y.L., W.L., Y.W., N.C. and M.C.; Software, Y.L., Y.W. and N.C.; Validation, Y.L., W.L., Y.W., N.C. and M.C.; Formal analysis, Y.L., Y.W. and N.C.; Investigation, Y.L.; Resources, Y.L.; Data curation, Y.L.; Writing—original draft,

Y.L., W.L., Y.W., N.C. and M.C.; Writing—review and editing, Y.L., W.L., Y.W., N.C. and M.C.; Supervision, W.L.; Funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (No. 62106094), Jiangxi Provincial Natural Science Foundation (No. 20242BAB20071) and Nanchang University Innovation Fund Project (No. XX202506040053).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this article are available in RAGEval at <https://github.com/OpenBMB/RAGEval>. The source code in the study are openly available at <https://github.com/l784669877/code/>.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Qualitative Analysis of Citation Accuracy Trade-Off

To investigate the nature of the slight drop in citation accuracy, we conducted a granular case study comparing the baseline (AARF) and ISFJ. As shown in Table A1, the “missed” citations predominantly fall into two benign categories: (1) **Redundant Structural Citations**, where the baseline mechanically cites generic transition phrases; and (2) **Over-Citation**, where the baseline appends citation tags even when hallucinating content. Crucially, the suppression of these structural tokens by ISFJ does not compromise the factual correctness of the answer.

**Table A1.** Case study of citation behavior: AARF vs. ISFJ. This table illustrates typical instances where ISFJ suppresses citations, primarily targeting redundant ones without compromising factual correctness. **Q** means question and **Ctx** means context.

| Case Type and Analysis                                                                                                                                                                                                                                                                                                                    | Q and Key Context                                                                                                                                                                                                                                                                | AARF Output                                                                                                                                                                       | ISFJ Output                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| <b>Case 1: Over-Citation</b><br>(Redundant sourcing)<br><b>Analysis:</b> Positive Suppression. AARF mechanically appends citation [2] even when hallucinating the content (“Nov” vs. “Dec”). ISFJ suppresses the explicit citation tag but correctly retrieves the specific date (“Dec 2021”), prioritizing factual accuracy over format. | <b>Q:</b> When did Sky Aviation and Sunshine Media adopt EMS?<br><b>Ctx:</b> [1] Sky Aviation Co., Ltd. adopted the environmental management system (EMS) in November 2017... [2] Sunshine Media Group Co., Ltd. adopted the environmental management system in December 2021... | “Sky Aviation adopted the system in Nov 2017 [1], and Sunshine Media adopted it in Nov 2021 [2].”<br><i>(Note: AARF hallucinates “Nov” for Sunshine Media despite citing [2])</i> | “Sky Aviation adopted the system in 2017 (specifically Nov 2017). Sunshine Media adopted it in 2021 (specifically Dec 2021).”         |
| <b>Case 2: Structural Redundancy</b><br><b>Analysis:</b> Benign Suppression. ISFJ removes the generic “According to...” preamble and associated tags. The core reasoning (comparing 2017 vs. 2018) and factual conclusion remain identical to the baseline, proving the suppression is harmless.                                          | <b>Q:</b> Whose management change happened earlier?<br><b>Ctx:</b> [3] In March 2017, Financial Advantage Ltd. underwent a senior management change... [6] Additionally, Happy Housekeeping Service Co., Ltd. conducted a senior management change in March 2018...              | “According to the documents, Financial Advantage Ltd. changed earlier in Mar 2017 [3], while Happy Housekeeping changed in Mar 2018 [6].”                                         | “Financial Advantage Ltd.’s management change was earlier, occurring in Mar 2017, whereas Happy Housekeeping’s occurred in Mar 2018.” |

## References

1. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
2. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. DeepSeek-V3 Technical Report. *arXiv* **2024**, arXiv:2412.19437.
3. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. 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)*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 9802–9822.
4. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. Lima: Less is more for alignment. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 55006–55021.
5. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55.
6. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
7. Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. OpenAI O1 System Card. *arXiv* **2024**, arXiv:2412.16720.
8. Chen, Q.; Qin, L.; Wang, J.; Zhou, J.; Che, W. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *Proceedings of the Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada, 10–15 December 2024.
9. Jin, M.; Yu, Q.; Shu, D.; Zhao, H.; Hua, W.; Meng, Y.; Zhang, Y.; Du, M. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 1830–1842.
10. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.; Li, Q. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 6491–6501.
11. Wang, X.; Amayuelas, A.; Zhang, K.; Pan, L.; Chen, W.; Wang, W. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, 21–27 July 2024; Volume 235, pp. 50026–50042.
12. Zhou, L.; Schellaert, W.; Martínez-Plumed, F.; Moros-Daval, Y.; Ferri, C.; Hernández-Orallo, J. Larger and more instructable language models become less reliable. *Nature* **2024**, *634*, 61–68.
13. Niu, C.; Wu, Y.; Zhu, J.; Xu, S.; Shum, K.; Zhong, R.; Song, J.; Zhang, T. 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)*; Association for Computational Linguistics: Bangkok, Thailand, 2024.
14. Chyzyhyk, D.; Varoquaux, G.; Milham, M.; Thirion, B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* **2022**, *11*, giac014.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
16. Hahn, M. Theoretical limitations of self-attention in neural sequence models. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 156–171.
17. Chiang, D.; Cholak, P. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 7654–7664.
18. Xiao, G.; Tian, Y.; Chen, B.; Han, S.; Lewis, M. Efficient streaming language models with attention sinks. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria, 7–11 May 2024.
19. Sun, Z.; Zang, X.; Zheng, K.; Song, Y.; Xu, J.; Zhang, X.; Yu, W.; Li, H. ReDeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv* **2024**, arXiv:2410.11414.
20. Jesson, A.; Beltran, Velez, N.; Chu, Q.; Karlekar, S.; Kossen, J.; Gal, Y.; Cunningham, J. P.; Blei, D. Estimating the hallucination rate of generative ai. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 31154–31201.
21. Abbasi Yadkori, Y.; Kuzborskij, I.; György, A.; Szepesvari, C. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 58077–58117.
22. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* **2019**, arXiv:1904.09751.
23. Chang, C.; Reitter, D.; Aksitov, R.; Sung, Y. KL-divergence guided temperature sampling. *arXiv* **2023**, arXiv:2306.01286.

24. Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread. 2021. Available online: <https://transformer-circuits.pub/2021/framework/index.html> (accessed on 30 January 2026).
25. Pearl, J. Causal inference. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*; PMLR: Whistler, BC, Canada, 2010; pp. 39–58.
26. Halpern, J.Y.; Pearl, J. Causes and explanations: A structural-model approach. Part I: Causes. *Br. J. Philos. Sci.* **2005**, *56*, 843–887.
27. Sun, Z.; Du, L.; Ding, X.; Ma, Y.; Zhao, Y.; Qiu, K.; Liu, T.; Qin, B. Causal-Guided Active Learning for Debiasing Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 14455–14469.
28. Wang, X.; Zhang, H.; Zhao, S.; Chen, H.; Ding, Z.; Wan, Z.; Cheng, B.; Lan, Y. Debiasing Counterfactual Context With Causal Inference for Multi-Turn Dialogue Reasoning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *32*, 1125–1132.
29. Zhu, X.; Mandivarapu, J.K. Trustful LLMs: Customizing and Grounding Text Generation with knowledge bases and Dual Decoders. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 156–166.
30. Chen, D.; Fang, F.; Ni, S.; Liang, F.; Xu, R.; Yang, M.; Li, C. Lower Layer Matters: Alleviating Hallucination via Multi-Layer Fusion Contrastive Decoding with Truthfulness Refocused. *arXiv* **2024**, arXiv:2408.08769.
31. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
32. Niu, M.; Li, H.; Shi, J.; Haddadi, H.; Mo, F. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *arXiv* **2024**, arXiv:2405.06545.
33. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
34. Robertson, S.E.; Jones, K.S. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129–146.
35. Devriendt, F.; Moldovan, D.; Verbeke, W. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* **2018**, *6*, 13–41.
36. Stahlberg, F.; Byrne, B. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3356–3362.
37. Shorinwa, O.; Mei, Z.; Lidard, J.; Ren, A. Z.; Majumdar, A. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv* **2024**, arXiv:2412.05563.
38. Fadeeva, E.; Rubashevskii, A.; Shelmanov, A.; Petrakov, S.; Li, H.; Mubarak, H.; Tsymbalov, E.; Kuzmin, G.; Panchenko, A.; Baldwin, T. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 9367–9385.
39. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It’s Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Singapore, 2023; pp. 967–976.
40. Zhu, K.; Luo, Y.; Xu, D.; Yan, Y.; Liu, Z.; Yu, S.; Wang, R.; Wang, S.; Li, Y.; Zhang, N. RagEval: Scenario Specific RAG Evaluation Dataset Generation Framework. *arXiv* **2024**, arXiv:2408.01262.
41. Yang, C.; Lin, Z.; Tian, C.; Pang, L.; Wang, L.; Tong, Z.; Ho, Q.; Cao, Y.; Wang, W. A Factuality and Diversity Reconciled Decoding Method for Knowledge-Grounded Dialogue Generation. *arXiv* **2024**, arXiv:2407.05718.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.