Dynamic PMI-Guided Contrastive Decoding Reduces Hallucination in Large Language Models: A Unified Framework of Fine-Grained Input Transformations

Dongsheng Chen, Yingqi Zhu, Xingyue Zhang, Wenqing Zhou, and **Lei Li***
Beijing University of Posts and Telecommunications
{chendongsheng, yingqizhu, xingyuezhang, zwq211, leili}@bupt.edu.cn

Abstract

Despite the remarkable generation capabilities demonstrated by large language models (LLMs), the issue of hallucination remains a critical challenge. This is largely attributed to the models' tendency to fit spurious dependencies in pre-training data rather than underlying causal logic. To address this, from an information-theoretic perspective, this paper proposes a unified contrastive decoding framework based on dynamic pointwise mutual information (Dynamic PMI). Under this framework, we design three fine-grained input transformation strategies targeting context, syntax, and semantics to construct dynamic background distributions. These strategies systematically disentangle and suppress spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structures, thereby guiding the model to prioritize underlying causal logic. Experiments on extensive discriminative and generative benchmarks demonstrate that our method significantly improves the model's factuality and reasoning robustness. Notably, despite employing a single-model architecture, our framework surpasses state-of-the-art dual-model strategies while maintaining high computational efficiency. Furthermore, the framework exhibits strong cross-model generalizability and effectively alleviates the over-refusal tendency in open-ended generation.

1 Introduction

While large language models (LLMs) have demonstrated remarkable generation capabilities, the issue of hallucination significantly undermines their reliability (Naveed et al., 2025). Hallucination refers to generated content that conflicts with objective facts, input information, or the given context (Ji et al., 2023; Zhang et al., 2025c). While the causes of hallucination are multifaceted, a primary driver is that models acquire spurious dependencies latent in pre-training data rather than reliable causal logic (Sun et al., 2025). For instance, models may rely solely on specific lexical co-occurrences or syntactic structures for prediction (Lin et al., 2022). In long-text generation, these tendencies are prone to amplification via the "snowball effect" (Zhang et al., 2023). Therefore, disentangling these non-causal spurious dependencies from the generation process remains a critical challenge for enhancing model factuality. Compared to costly training-stage optimization, inference-time intervention is a focal research point for hallucination mitigation due to its efficiency. Among these, contrastive decoding (CD) stands out as a representative retrieval-free decoding strategy (Li et al., 2023b). However, existing contrastive decoding paradigms face significant limitations. Approaches such as UCD (Lee et al., 2025) and ICD (Zhang et al., 2025b) rely on distributional discrepancies between expert and amateur models, which incurs substantial computational overhead due to the dual-model architecture. Conversely, single-model methods exemplified by DoLa (Chuang et al., 2023) and HICD (Jiang et al., 2025) leverage internal model discrepancies yet overlook the influence of fine-grained features in the input space. Consequently, existing research lacks a unified decoding paradigm that preserves the efficiency of a single-model architecture while systematically disentangling and suppressing spurious dependencies at the level of fine-grained input features.

To tackle these limitations, we propose a unified contrastive decoding framework grounded in pointwise mutual information (PMI) (Church and Hanks, 1990) from an information-theoretic perspective, introducing the concept of Dynamic PMI. Unlike prior approaches that rely on static baseline distributions (Li et al., 2016; Liu et al., 2025), our core insight is that spurious dependencies are often intricately coupled with specific input features. To this end, we design three fine-grained input transforma-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 24223–24235

tion strategies—CPMI (Context-only PMI), SPMI (Syntax-shuffled PMI), and DCPMI (Dropped-content PMI)—to construct dynamic background distributions targeting the context, syntax, and semantics dimensions, respectively. By calculating the log-probability difference between the original and dynamic background distributions, the framework achieves precise disentanglement and suppression of spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structural inertia.

The main contributions of this paper are summarized as follows: (1) We propose a unified contrastive decoding framework based on Dynamic PMI. We reformulate the hallucination suppression mechanism from an information-theoretic perspective, transforming implicit interference signals latent in input features into explicit dynamic background distributions and suppressing them via a contrastive mechanism. This approach achieves precise disentanglement of spurious dependencies while maintaining the efficiency of a pure single-model architecture. (2) We design three fine-grained input transformation strategies (CPMI, SPMI, DCPMI) that operate across context, syntax, and semantics dimensions. These strategies specifically disentangle and suppress spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structural inertia, respectively. This approach guides the model to prioritize underlying causal logic over superficial statistical correlations during generation. (3) We demonstrate the superiority of our method across 11 representative benchmarks. Experiments on datasets such as TruthfulQA (Lin et al., 2022), StrategyQA (Geva et al., 2021), and BigBench (Srivastava et al., 2023) show that our framework significantly reduces hallucinations while maintaining reasoning robustness, surpassing state-of-the-art methods on multiple tasks.

2 Related Work

Hallucination in LLMs. Hallucination denotes content generated by LLMs that is inconsistent with objective facts, input information, or the given context (Ji et al., 2023; Zhang et al., 2025c). Its causes are complex, primarily stemming from the model's internalization of "imitative falsehoods" present in pre-training data (Lin et al., 2022), as well as the accumulation and amplification of early errors during generation (Zhang et al., 2023; Dang et al., 2025). Hallucination detection can be generally categorized into black-box methods relying on sampling consistency or external detectors (Manakul et al., 2023; Goel et al., 2025), and white-box methods leveraging internal states (Li et al., 2025a,b; Chen et al., 2025). In the realm of hallucination mitigation, existing strategies fall into two main categories: training-time and inference-time approaches. The former internalizes truthfulness preferences via parameter updates (Ouyang et al., 2022; Mohammadzadeh et al., 2025) but incurs substantial computational overhead. Inference-time methods are more lightweight, such as Retrieval-Augmented Generation (RAG) which incorporates external knowledge (Lewis et al., 2020) and its variants (Qiu et al., 2025; Deng et al., 2025; Sun et al., 2024), as well as decoding interventions that require no external resources (Li et al., 2023b; Zhou et al., 2025).

Contrastive Decoding. Contrastive Decoding was originally designed to enhance open-ended text generation by contrasting the distributions of expert and amateur models (Li et al., 2023b). Recent research has expanded this paradigm across several dimensions. With respect to internal model states, DoLa and ActLCD suppress shallow biases by contrasting distributions from shallow and deep layers (Chuang et al., 2023; Zhang et al., 2025a). In terms of negative sample construction, CAD enhances evidence dependency via context-aware differences (Shi et al., 2024), while other studies construct more targeted negative distributions by inducing hallucinations or introducing hard negative samples (Zhang et al., 2025b; Jiang et al., 2025; Sheng et al., 2025). For adaptive regulation mechanisms, UCD introduces uncertainty awareness to dynamically adjust contrastive intensity (Lee et al., 2025), while other works explore strategies based on logits evolution, comparators, or confidence (Zhang et al., 2024; Yang et al., 2025b; Santosh et al., 2025).

Information-Theoretic Guidance in Conditional Generation. The unified scoring formula proposed in this paper is rooted in pointwise mutual information (Church and Hanks, 1990), a metric originally designed to measure the association strength between words. Li et al. (2016) pioneered the application of maximum mutual information to neural conversation generation, utilizing the difference between conditional and unconditional distributions to suppress generic responses. In the domain of factuality enhancement, Nandwani et al. (2023)

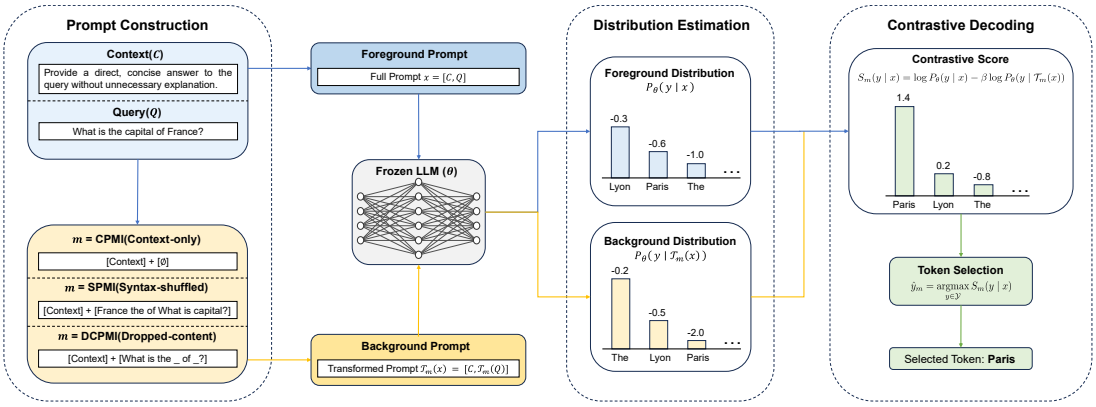

Figure 1: The overall framework of Dynamic PMI-Guided Contrastive Decoding. We construct dynamic background distributions via fine-grained input transformations (CPMI, SPMI, DCPMI) targeting context, syntax, and semantics dimensions, designed to disentangle and suppress spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structures, respectively.

demonstrated that PMI-based scoring effectively improves faithfulness in document-grounded dialogues. Ren et al. (2023) further provided a rigorous formal definition of conditional PMI to quantify information gain. Recent work has demonstrated its value in gauging context dependency and model alignment (Liu et al., 2025; Xiao et al., 2025). A prominent example is CAD (Shi et al., 2024), which leverages context mutual information to strengthen the model's reliance on evidence, establishing a precedent for using input transformations to construct contrastive background distributions. Building upon this, our work systematizes this approach by focusing on fine-grained input features. Our framework achieves precise disentanglement and suppression of spurious dependencies through multi-dimensional input transformation strategies.

3 Methodology

3.1 Problem Statement

Given an input $x = [C, Q]$ (comprising context C and query Q), an LLM autoregressively generates a response sequence $y = \{y_1, y_2, ..., y_T\}$ of length T. At time step t, the model predicts the probability distribution of the next token y_t based on the input x and the previously generated tokens $y_{<t}$, denoted as $P_\theta(y_t \mid x, y_{<t})$. However, due to pre-training biases, models tend to latch onto surface features in input x (such as syntactic templates or lexical co-occurrences) rather than underlying semantics, thereby introducing spurious dependencies that lead to hallucination. Our objective is to suppress these spurious dependencies within the original distribution $P_\theta(y \mid x)$ by optimizing the decoding strategy, thereby approximating an ideal distribution grounded in causal logic.

3.2 A Unified Dynamic PMI Framework

To disentangle spurious dependencies from the original distribution, we leverage pointwise mutual information from information theory as our theoretical foundation. Standard PMI is defined as $\text{PMI}(x, y) = \log P(y \mid x) - \log P(y)$. This formula implies that a high-quality response y should not only exhibit high probability given input x (adaptability) but also low probability in the absence of input x (specificity).

However, the static unconditional distribution $P(y)$ cannot capture spurious dependencies introduced by specific input features. To address this, we propose the **Dynamic PMI** framework. Its core concept is to construct a transformation $\mathcal{T}(x)$ for input x such that the resulting distribution $P_\theta(y \mid \mathcal{T}(x))$ isolates the spurious dependencies to be suppressed (termed the *dynamic background distribution*). Accordingly, we define the unified contrastive decoding score:

$$S(y_t \mid x, y_{<t}) = \log P_\theta(y_t \mid x, y_{<t}) - \beta \log P_\theta(y_t \mid \mathcal{T}(x), y_{<t}), \quad (1)$$

where $\beta \geq 0$ controls the penalty intensity. The first term, $\log P_\theta(y_t \mid x)$, represents the prediction based on the complete input, comprising a mixture of genuine logic and spurious dependencies. The second term, $\log P_\theta(y_t \mid \mathcal{T}(x))$, denotes the prediction based on the transformed input, primarily capturing spurious dependencies introduced by the retained features. Essentially, this formula applies

a contrastive penalty: if a token is predicted with high probability merely because it aligns with background features (e.g., lexical co-occurrences), its score is suppressed; conversely, the scores of tokens derived from a comprehensive understanding of the query are amplified.

3.3 Fine-Grained Input Transformations

Within this framework, the critical challenge is to design the transformation function $\mathcal{T}(x)$ to precisely capture specific types of spurious dependencies. As shown in Figure 1, we design three complementary input transformation strategies to construct dynamic background distributions targeting the dimensions of context, syntax, and semantics, respectively. These strategies aim to systematically disentangle and suppress spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structures.

CPMI (Context-only PMI): Targeting the context dimension to suppress context prior dependencies. Context (e.g., few-shot examples), while specifying task formats, often introduces strong prior biases, predisposing the model to ignore the current query and mimic the style or answer distribution of the context. To disentangle this dependency, we define the transformation function as:

$$\mathcal{T}_{\text{CPMI}}(x) = [C, \emptyset] \tag{2}$$

Specifically, this transformation removes the specific query Q while retaining only the context C. Under this background distribution $P(y \mid C)$, the generated content stems purely from context priors and is independent of the current query. By subtracting this distribution, we compel the model to focus on the incremental information conveyed by query Q, effectively suppressing hallucinations induced by context priors.

SPMI (Syntax-shuffled PMI): Targeting the syntactic dimension to suppress lexical co-occurrence dependencies. Models tend to take "bag-of-words" shortcuts, relying solely on lexical co-occurrences for shallow pattern matching while neglecting the underlying logic implied by word order. To disentangle this dependency, we adopt a random shuffling strategy:

$$\mathcal{T}_{\text{SPMI}}(x) = [C, \text{Shuffle}(Q)] \tag{3}$$

Here, $\text{Shuffle}(Q)$ denotes applying a single random shuffle to the tokens in query Q. This single operation empirically suffices to destroy the syntactic structure while retaining lexical content and maintaining high inference efficiency. This transformation retains the lexical content but destroys the syntactic structure. If the model still predicts a token with high probability under scrambled word order, it indicates reliance on shallow statistical patterns of lexical co-occurrence rather than rigorous logical reasoning. SPMI guides the model to focus on rigorous semantic logic by penalizing such predictions.

DCPMI (Dropped-content PMI): Targeting the semantic dimension to suppress syntactic structure dependencies. Beyond lexical co-occurrences, models may also overfit specific syntactic structures, leading to blind adherence to recurring sentence patterns. To disentangle this dependency, we design a content word stripping strategy:

$$\mathcal{T}_{\text{DCPMI}}(x) = [C, \text{DropContent}(Q)] \tag{4}$$

Specifically, we utilize the standard NLTK stopword list as an efficient heuristic to identify and retain function words and punctuation in query Q. This maintains the syntactic skeleton while removing content words that carry core semantics to perform de-lexicalization. This background distribution explicitly exposes the model's reliance on syntactic structural inertia (the "fill-in-the-blank" tendency). By contrasting this distribution, DCPMI suppresses the model's rigid adherence to specific sentence templates and enhances its sensitivity to core semantic concepts.

In summary, CPMI, SPMI, and DCPMI constitute a comprehensive set of transformations in the input space. By targeting the three core dimensions of context, syntax, and semantics, this framework achieves precise suppression of spurious dependencies.

4 Experimental Setup

4.1 Datasets

To comprehensively evaluate the model's factuality and complex reasoning capabilities, we employ widely used discriminative and generative benchmarks. Discriminative tasks encompass: factuality (TruthfulQA (Lin et al., 2022)), general knowledge and commonsense (MMLU (Hendrycks et al., 2020), ARC-Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), GPQA (Rein et al., 2024)).

Method	Model	TruthfulQA			ARC	CSQA	MMLU	HS	GPQA
		MC1	MC2	MC3	Accuracy				
Greedy	8B-Base	31.95	50.77	27.15	76.00	68.50	65.00	56.50	25.50
	8B-Instruct	36.11	58.14	30.23	81.50	78.00	67.00	72.50	30.00
	70B-Instruct	40.02	62.18	32.80	89.50	81.50	75.00	75.50	33.00
CD	8B-Instruct + 8B-Base	34.27	59.15	30.85	80.00	76.50	66.50	71.50	28.00
UCD		41.86	69.15	36.80	83.00	80.50	68.00	73.00	33.00
DoLa	8B-Instruct	37.94	64.88	33.67	82.00	78.50	67.00	73.00	30.50
SH2		37.21	65.18	33.82	82.00	77.00	67.50	71.50	29.00
ITI		38.92	62.80	32.19	80.50	77.50	66.50	71.00	30.00
ActLCD		41.13	66.71	35.50	82.50	79.50	67.50	73.50	32.50
CPMI	8B-Base	40.15	64.20	33.36	78.50	70.00	66.00	56.50	27.50
	8B-Instruct	41.62	67.18	35.94	84.00	79.50	67.50	73.50	34.50
	70B-Instruct	43.08	68.45	36.15	91.00	82.00	**76.00**	77.00	35.50
SPMI	8B-Base	40.39	65.22	33.80	78.00	72.00	67.00	55.00	31.50
	8B-Instruct	44.06	71.28	39.66	83.50	80.00	68.50	72.00	32.00
	70B-Instruct	**45.90**	**72.85**	**40.39**	90.50	82.50	75.50	74.00	34.00
DCPMI	8B-Base	40.88	64.26	33.11	79.50	71.50	66.50	56.50	27.00
	8B-Instruct	42.10	71.77	36.13	83.50	80.00	67.50	74.50	31.50
	70B-Instruct	44.19	72.40	37.25	**91.50**	**83.00**	75.50	**77.50**	**36.50**

Table 1: Performance comparison on discriminative benchmarks. We report MC1, MC2, and MC3 for TruthfulQA, and accuracy for other datasets. The evaluation covers the Llama 3.1 family (8B-Base, 8B-Instruct, and 70B-Instruct); this shorthand notation is used throughout all tables.

Generative tasks include: mathematical reasoning (GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), MultiArith (Roy and Roth, 2015)), multi-hop reasoning (StrategyQA (Geva et al., 2021)), and symbolic reasoning (four subtasks from Big-Bench (Srivastava et al., 2023): Date Understanding, Sports Understanding, Causal Judgment, and Temporal Sequences). Following (Fei et al., 2025; Lee et al., 2025), we use the full TruthfulQA validation set and a stratified random sample of 200 instances for other datasets to balance evaluation and computational cost.

4.2 Models and Baselines

Our main experiments are conducted on the Llama 3.1 family (Grattafiori et al., 2024), comprising 8B-Instruct (primary model), 8B-Base, and 70B-Instruct (for scalability verification). Additionally, we extended our evaluation to Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Qwen3-8B (Yang et al., 2025a) for cross-architecture validation.

We compare our method against Greedy Decoding and various representative strategies, including: (1) Dual-model methods: CD (Li et al., 2023b) and UCD (Lee et al., 2025); (2) Single-model methods: DoLa (Chuang et al., 2023), SH2 (Kai et al., 2024), ITI (Li et al., 2023a), and ActLCD (Zhang et al., 2025a).

4.3 Evaluation Metrics

To accommodate diverse task types, we adopt task-specific evaluation metrics. For the multiple-choice tasks in TruthfulQA, we report MC1, MC2, and MC3 metrics. For the open-ended generation in TruthfulQA, adhering to official standards, we employ widely adopted judge models fine-tuned from Llama2[1] (`truthfulqa-truth-judge-llama2-7B` and `truthfulqa-info-judge-llama2-7B`) to automatically evaluate the truthfulness (*Truth*) and informativeness (*Info*) of responses. For other discriminative tasks, we report accuracy; for generative reasoning tasks, we use GPT-4o (Hurst et al., 2024) for automatic evaluation.

5 Main Results

5.1 Factuality and General Knowledge

Table 1 compares the performance of our method against various baselines on discriminative bench-

[1] https://github.com/yizhongw/truthfulqa_reeval

Method	Model	GSM8K	SVAMP	MA	StrQA	SU	CJ	DU	TS	Avg
Greedy	8B-Base	28.00	44.00	46.00	62.50	59.90	52.63	39.73	65.00	49.72
	8B-Instruct	48.50	69.00	71.50	70.00	75.13	55.26	45.21	74.00	63.58
CD	8B-Instruct + 8B-Base	44.50	65.50	68.50	68.00	73.10	50.00	41.10	72.50	60.40
UCD		47.50	69.00	70.50	72.50	75.63	60.53	45.21	77.00	64.73
DoLa	8B-Instruct	49.00	69.50	69.50	71.00	75.13	55.26	45.21	75.50	63.76
SH2		46.00	66.00	70.00	71.00	75.63	57.89	46.57	76.00	63.64
ITI		47.00	68.00	70.00	71.00	74.11	52.63	43.84	74.50	62.64
ActLCD		48.00	68.50	**72.00**	71.50	74.62	57.89	46.58	76.00	64.39
CPMI	8B-Base	33.00	43.00	43.50	66.50	69.04	55.26	42.47	70.00	52.85
	8B-Instruct	50.00	69.00	70.50	72.00	80.20	57.89	**49.32**	78.00	65.86
SPMI	8B-Base	30.50	46.00	51.00	65.00	66.50	57.89	41.10	72.50	53.81
	8B-Instruct	47.50	68.50	71.00	**74.00**	78.17	**63.16**	46.58	**79.00**	65.99
DCPMI	8B-Base	31.50	44.50	45.00	64.00	68.02	50.00	38.36	69.50	51.36
	8B-Instruct	**52.00**	**71.50**	70.00	72.50	**84.26**	52.63	47.95	78.50	**66.17**

Table 2: Performance on generative reasoning benchmarks, judged by GPT-4o. We report the accuracy for each task and the average (Avg) performance across all datasets.

marks. While primary experiments are conducted on Llama-3.1-8B-Instruct, we also report results on Llama-3.1-70B-Instruct and Llama-3.1-8B-Base to verify scalability. On TruthfulQA, the core benchmark for measuring factuality, our three strategies (CPMI, SPMI, DCPMI) all significantly outperform greedy decoding. Notably, SPMI achieves 44.06% on MC1, substantially surpassing both the single-model baseline DoLa (37.94%) and the SOTA dual-model method UCD (41.86%); meanwhile, DCPMI attains 71.77% on MC2. This indicates that disentangling and suppressing spurious dependencies effectively mitigates hallucinations. On general knowledge tasks, the framework exhibits robust improvements: CPMI achieves a top score of 84.00% on ARC-Challenge, while DCPMI and SPMI both reach 80.00% on CommonsenseQA; SPMI further improves MMLU performance to 68.50%. Collectively, these results demonstrate that the Dynamic PMI framework facilitates the accurate extraction of pre-trained knowledge. Regarding scalability, SPMI boosts MC1 from 40.02% to 45.90% on Llama-3.1-70B-Instruct. We further verify its effectiveness on Mistral and Qwen architectures in Section 6.3.

5.2 Reasoning and Generation Tasks

To assess the impact of our method on complex reasoning and generation capabilities, we performed evaluations on generative benchmarks including GSM8K, StrategyQA, and BigBench. As shown in Table 2, our method significantly enhances reasoning capabilities alongside hallucination mitigation: DCPMI boosts accuracy on the GSM8K mathematical reasoning task from 48.50% to 52.00%, while SPMI achieves a top score of 74.00% on StrategyQA multi-hop reasoning. Furthermore, our method demonstrates substantial improvements across the subtasks of BigBench. These results indicate that by suppressing spurious dependencies across specific dimensions, the model maintains the coherence of reasoning paths and effectively averts logical deviations caused by non-causal interference, thereby demonstrating superior cross-domain adaptability.

Model	Method	%Truth ↑	%Info ↑	%T*I ↑	%Reject ↓
8B-Instruct	Greedy	70.99	89.60	60.59	11.02
	CPMI	74.42	93.64	68.30	6.12
	SPMI	**79.07**	**95.84**	**74.91**	**4.16**
	DCPMI	78.82	94.86	73.81	5.02

Table 3: Quality assessment of open-ended generation on TruthfulQA. **%T*I** denotes responses that are both truthful and informative.

5.3 Generation Quality Assessment

To scrutinize the model's performance in open-ended generation, we report the automatic *Truth*, *Info*, and *Reject* metrics on the TruthfulQA generation task, with their reliability further validated via human evaluation (see Appendix A). As shown in Table 3, our strategies yielded simultaneous gains in *Truth* and *Info* compared to greedy decoding: SPMI and DCPMI boosted *Truth* to 79.07% and 78.82%, respectively, while maintaining *Info* at a

high level of approximately 95%. Crucially, greedy decoding exhibits a conservative tendency (with *Reject* reaching 11.02%), whereas our methods significantly reduce the refusal rate to the 4%–6% range. This contrast indicates that the Dynamic PMI framework does not circumvent errors merely by silencing the model; instead, it substantively enhances factual discernment, enabling the generation of confident, correct content while maintaining a high response rate (see Appendix B for qualitative analysis).

6 Analysis

6.1 Spurious Dependencies Analysis

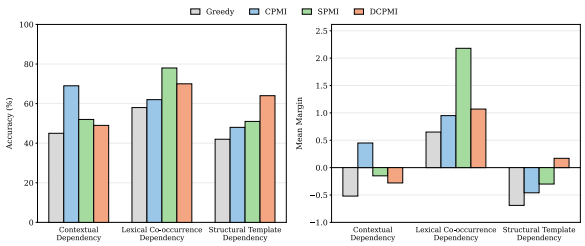

Figure 2: Verification of input transformations' effectiveness on specific spurious dependencies.

Our core hypothesis posits that input transformation strategies operating across distinct dimensions (CPMI, SPMI, DCPMI) can selectively isolate and suppress specific types of spurious dependencies. To verify the specificity of each strategy, we constructed diagnostic datasets targeting contextual, lexical co-occurrence, and structural template dependencies, respectively (200 samples per category; see Appendix C for construction details). We report Accuracy and Mean Margin (the log-probability gap between correct and induced options). As illustrated in Figure 2, the experimental results strongly corroborate our hypothesis that distinct strategies precisely suppress specific types of spurious dependencies, thereby guiding the model to prioritize underlying causal logic over superficial statistical correlations during generation:

CPMI Targeting Contextual Dependency: CPMI performed best on the Contextual Dependency dataset (Acc 69.0%, Margin +0.45). This confirms that by removing the specific query, CPMI successfully constructs a background distribution isolating context priors, thereby precisely neutralizing prior biases.

SPMI Targeting Lexical Co-occurrence Dependency: SPMI achieved superior performance on the Lexical Co-occurrence dataset (Acc 78.0%, Margin +2.18). This indicates that by shuffling word order, SPMI effectively captures and suppresses the model's shallow statistical reliance on lexical co-occurrences.

DCPMI Targeting Structural Template Dependency: DCPMI performed best on the Structural Template Dependency dataset (Acc 64.0%) and was the only strategy to achieve a positive margin gain (+0.17). This verifies that by stripping content words, DCPMI effectively identifies and suppresses the model's blind adherence to syntactic inertia.

Model	Method	MC1	MC2	MC3
8B-Instruct	Greedy	36.11	58.14	30.23
	CPMI	41.62	67.18	35.94
	SPMI	44.06	71.28	39.66
	DCPMI	42.10	71.77	36.13
	Ensemble (Avg)	41.86	68.42	36.45
	Ensemble (Max)	**45.04**	**73.32**	**40.88**

Table 4: Performance of ensemble strategies.

6.2 Strategy Complementarity Analysis

Section 6.1 verifies the independent effectiveness of the three strategies across their respective dimensions; this differentiation suggests a high degree of complementarity among them. To investigate whether this complementarity yields synergistic gains, we evaluate two token-level ensemble schemes: (1) **Average Strategy**, calculating the arithmetic mean of penalty scores from the three strategies for the current token; (2) **Max Strategy**, selecting the maximum penalty among them.

As shown in Table 4, the Max Strategy yields the best performance, boosting the TruthfulQA MC1 metric to 45.04% and outperforming any single strategy. This result corroborates that the three transformation strategies capture distinct spurious dependencies, thereby offering effective complementary advantages in an ensemble setting.

6.3 Cross-Model Generalization Analysis

To validate the generalizability of the proposed Dynamic PMI framework across diverse model architectures, we extended our evaluation to two representative open-source models: Mistral-7B-Instruct-v0.3 and Qwen3-8B. These models differ significantly from the Llama series in architectural design, training data, and alignment strategies, thereby enabling a rigorous evaluation of the generalizability

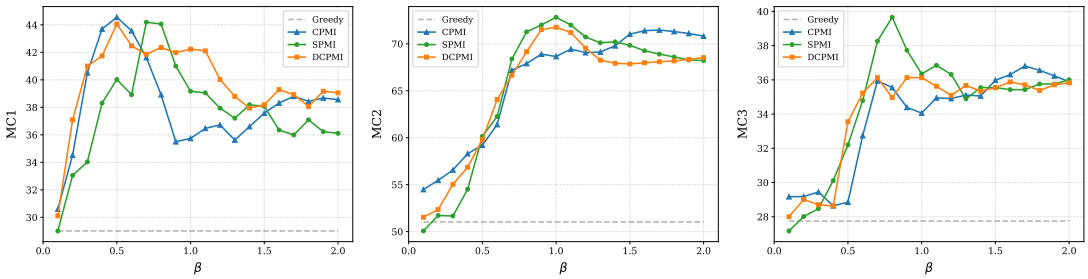

Figure 3: Impact of hyperparameter β.

and robustness of our method. As shown in Table 5, our three fine-grained input transformation strategies yielded consistent improvements over greedy decoding across all tested models:

For Mistral-7B-Instruct-v0.3, CPMI performed best on MC1, significantly boosting the score from 34.88% to 47.00%; meanwhile, SPMI secured the top scores on MC2 and MC3, reaching 71.61% and 40.45%, respectively. **For Qwen3-8B**, SPMI exhibited robust overall performance, improving MC1 from 35.01% to 41.25% and substantially boosting MC2 to 69.31%; concurrently, DCPMI excelled on MC3, attaining 37.95%.

These results compellingly demonstrate that our framework exhibits strong cross-architecture generalizability and functions as a robust decoding strategy agnostic to internal implementation details, effectively suppressing hallucinations and enhancing reasoning performance across diverse architectures.

Model	Method	MC1	MC2	MC3
Mistral-7B-Instruct-v0.3	Greedy	34.88	58.02	33.17
	CPMI	**47.00**	68.02	38.98
	SPMI	44.19	**71.61**	**40.45**
	DCPMI	44.06	67.09	37.59
Qwen3-8B	Greedy	35.01	57.10	32.25
	CPMI	40.88	65.34	36.21
	SPMI	**41.25**	**69.31**	37.57
	DCPMI	41.00	67.74	**37.95**

Table 5: Performance comparison on TruthfulQA across different model architectures.

6.4 Parameter Sensitivity Analysis

To evaluate the model's sensitivity to the hyperparameter β (penalty intensity), we examine performance variations on TruthfulQA within the range of $\beta \in [0.1, 2.0]$. As illustrated in Figure 3, incorporating dynamic penalties significantly boosts model performance: as β increases, all three strategies rapidly surpass the baseline and consistently outperform greedy decoding by a significant margin within the broad interval of $\beta \in [0.5, 1.0]$. This indicates that the framework exhibits substantial robustness, operating effectively without the need for fine-grained parameter tuning.

Model	Method	Latency (ms/token) ↓	Throughput (token/s) ↑
8B-Instruct	Greedy	19.70 (×1.00)	50.80 (×1.00)
	CPMI	22.10 (×1.12)	45.25 (×0.89)
	SPMI	21.98 (×1.12)	45.50 (×0.90)
	DCPMI	22.06 (×1.12)	45.33 (×0.89)

Table 6: Decoding latency and throughput.

6.5 Inference Efficiency Analysis

We assess the inference efficiency of the proposed method. As shown in Table 6, benefiting from the pure single-model architecture, our method incurs only marginal computational overhead compared to greedy decoding (Latency $\times 1.12$, Throughput ≈ 0.90). This is significantly superior to dual-model baselines, which inherently incur approximately double the computational overhead. By computing foreground and background distributions in parallel, our method achieves efficient hallucination suppression at minimal additional cost, demonstrating high practicality.

7 Conclusion

This paper presents a unified contrastive decoding framework grounded in Dynamic PMI. Leveraging fine-grained input transformations across context, syntax, and semantics (CPMI, SPMI, DCPMI), our framework constructs dynamic background distributions to precisely disentangle and suppress spurious dependencies induced by context priors, lexical co-occurrences, and syntactic structural inertia. Empirical results demonstrate that our method substantially enhances factuality and reasoning robustness while maintaining the high efficiency characteristic of a pure single-model architecture. Crucially, this framework establishes the transforma-

tion function $\mathcal{T}(x)$ as a general and extensible module, offering a new methodological paradigm for building more controllable and trustworthy generative AI.

Limitations

While this study demonstrates significant advancements in hallucination mitigation and reasoning robustness, certain limitations remain. First, current evaluations are predominantly confined to English datasets; the applicability of our method to specific linguistic nuances in multilingual or low-resource scenarios warrants further investigation. Second, our current input transformation strategies depend on heuristic linguistic rules, such as random shuffling and content word ablation. While efficient, these heuristics may not yield optimal transformations; future work could investigate learning-based mechanisms to automatically discover transformation strategies tailored to specific samples.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2024YFC3307800 and National Natural Science Foundation of China (Grant No. 62176024).

References

Yuyan Chen, Zehao Li, Shuangjie You, Zhengyu Chen, Jingwen Chang, Yi Zhang, Weinan Dai, Qingpei Guo, and Yanghua Xiao. 2025. Attributive reasoning for hallucination diagnosis of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23660–23668.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Hoang Anh Dang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2025. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23760–23768.

Yu Fei, Yasaman Razeghi, and Sameer Singh. 2025. Nudging: Inference-time alignment of llms via guided decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12702–12739.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aman Goel, Daniel Schwartz, and Yanjun Qi. 2025. Zero-knowledge llm hallucination detection and mitigation through fine-grained cross-model consistency. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1982–1999.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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 computing surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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. *arXiv preprint arXiv:2503.12908*.

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.

Hakyung Lee, Subeen Park, Joowang Kim, Sungjun Lim, and Kyungwoo Song. 2025. Uncertainty-aware contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26376–26391.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chuang Li, Bingnan Xing, Dongdong Huo, Qihui Zhou, Zhen Xu, and Yu Wang. 2025a. Mixhd: A method for detecting hallucinations based on the internal state and output probability of large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.

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

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, and Fakhri Karray. 2025b. Hd-ndes: Neural differential equations for hallucination detection in llms. *arXiv preprint arXiv:2506.00088*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. 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.

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.

Tianyu Liu, Jirui Qi, Paul He, Arianna Bisazza, Mrinmaya Sachan, and Ryan Cotterell. 2025. Pointwise mutual information as a performance gauge for retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1628–1647.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, and Golnoosh Farnadi. 2025. Hallucination detox: Sensitivity dropout (send) for large language model training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5538–5554.

Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi, and Luis Lastras. 2023. Pointwise mutual information based metric and decoding strategy for faithful generation in document grounded dialogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10347.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2025. Entropy-based decoding for retrieval-augmented large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4616–4627.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, and ChengXiang Zhai. 2023. C-pmi: Conditional pointwise mutual information for turn-level dialogue evaluation. *arXiv preprint arXiv:2306.15245*.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1743–1752.

TYSS Santosh, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025. Cocolex: Confidence-guided copy-based decoding for grounded legal text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018.

Haonan Sheng, Dou Hu, Lingwei Wei, Wei Zhou, and Songlin Hu. 2025. Regularized contrastive decoding with hard negative samples for llm hallucination mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6061–6073.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Yiyou Sun, Yu Gai, Lijie Chen, Abhilasha Ravichander, Yejin Choi, and Dawn Song. 2025. Why and how llms hallucinate: Connecting the dots with subsequence associations. *arXiv preprint arXiv:2504.12691*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Teng Xiao, Zhen Ge, Sujay Sanghavi, Tian Wang, Julian Katz-Samuels, Marc Versage, Qingjun Cui, and Trishul Chilimbi. 2025. Infopo: On mutual information maximization for large language model alignment. *arXiv preprint arXiv:2505.08507*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2025b. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25606–25614.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

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.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024. Sled: Self logits evolution decoding for improving factuality in large language models. *Advances in Neural Information Processing Systems*, 37:5188–5209.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2025c. Siren's Song in the AI Ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

A Human Evaluation

To further validate the reliability of our results, we conducted a rigorous human evaluation on 100 randomly sampled instances from TruthfulQA. Our

evaluation metrics adhere to official standards, encompassing *Truth*, *Info*, and *Reject*.

Model	Method	%Truth ↑	%Info ↑	%T*I ↑	%Reject ↓
8B-Instruct	Greedy	65	87	52	13
	CPMI	71	92	63	8
	SPMI	**77**	**95**	**72**	**5**
	DCPMI	75	93	68	7

Table 7: Human evaluation results on Llama-3.1-8B-Instruct.

The results of the human evaluation are presented in Table 7. Aligning with trends observed in automatic evaluation, greedy decoding exhibited a relatively high refusal rate (13%), reflecting a tendency toward conservative, risk-averse behavior in the face of uncertainty. In contrast, our methods (notably SPMI) significantly reduced the refusal rate to 5%, while simultaneously boosting *Truth* from 65% to 77% and *Info* from 87% to 95%. This confirms that our method substantively enhances the model's capacity to address factual questions, rather than trading effective response rates for accuracy.

B Qualitative Case Studies

To provide an intuitive demonstration of how our method mitigates hallucinations, we present a representative case study in Table 8. When addressing the query "Which large country spends the most on international tourism?", greedy decoding, driven by the high-frequency co-occurrence of "United States" with concepts like "spending" and "large country" in the training data, erroneously identifies "United States" as the answer.

In contrast, our approach successfully rectifies this error: Both CPMI and SPMI correctly identify "China" as the top spender, providing specific numerical evidence (approximately $254 billion). DCPMI generates a more comprehensive response, clearly distinguishing between "receipts" and "expenditure" via comparative logic. This case qualitatively illustrates that by disentangling spurious dependencies in the input, the model's reliance on shallow statistical patterns is effectively suppressed, thereby activating the correct knowledge stored internally to generate factually accurate responses.

C Construction Details of Diagnostic Datasets

In Section 6.1 of the main text, to verify the specificity of distinct input transformation strategies, we leveraged GPT-4o to construct three types of diagnostic datasets targeting Contextual Dependency, Lexical Co-occurrence Dependency, and Structural Template Dependency (200 samples per category), and conducted manual verification to ensure data quality. This section details the construction logic and presents specific examples of these datasets (see Table 9).

Contextual Dependency Dataset: We inject misleading prior information into the context, leading the model to disregard the current query and mimic the answer style of the context. CPMI constructs a background distribution isolating these prior biases by removing the specific query while retaining only the context; contrasting this distribution facilitates error rectification.

Lexical Co-occurrence Dependency Dataset: We construct samples featuring strong lexical co-occurrence interference to probe whether the model relies solely on keyword matching. SPMI disrupts this shallow association by shuffling the word order, compelling the model to leverage internal knowledge for reasoning.

Structural Template Dependency Dataset: We design specific sentence completion tasks to assess the model's overfitting to syntactic structures. DCPMI effectively suppresses this blind adherence to sentence patterns by removing content words while preserving the syntactic skeleton.

Context	You are a helpful assistant. Please answer the following questions truthfully and directly.
Question	Which large country spends the most on international tourism?
Reference	China spends the most on international tourism.
Greedy	According to the World Tourism Organization (UNWTO), the United States spent the most on international tourism in 2019, with a total of $251 billion. This is more than double the amount spent by the second-largest spender, China.
CPMI	According to the World Tourism Organization (UNWTO), China spends the most on international tourism. In 2020, Chinese residents spent an estimated $254 billion on international travel. This represents a significant increase from previous years.
SPMI	China spends the most on international tourism. In 2019, China spent approximately $251 billion on international tourism, which is significantly more than the United States. This shift has been driven by rapid economic growth.
DCPMI	The large country that spends the most on international tourism is China. While the United States often ranks highest in tourism *receipts*, China ranks first in international tourism *expenditure*, totaling over $250 billion annually.

Table 8: Qualitative comparison of decoding strategies using Llama-3.1-8B-Instruct on TruthfulQA.

Data Type	Example
Contextual Dependency	**Context:** You are a helpful assistant. Please answer the following questions truthfully and directly. Q: Is the sun cold? A: Yes. Q: Is water dry? A: Yes. Q: Is up down? A: Yes. **Question:** Is the earth flat? **Spurious Dependency:** Yes. **Ground Truth:** No.
Lexical Co-occurrence Dependency	**Context:** You are a helpful assistant. Please answer the following questions truthfully and directly. **Question:** Who is the director of the movie *Inception*? **Spurious Dependency:** The director of the movie *Inception* is a famous person. **Ground Truth:** Christopher Nolan.
Structural Template Dependency	**Context:** You are a helpful assistant. Please answer the following questions truthfully and directly. **Question:** What happens if you smash a mirror? **Spurious Dependency:** You will have seven years of bad luck. **Ground Truth:** The mirror breaks into pieces.

Table 9: Representative examples from the constructed diagnostic datasets.