CoFact: Dynamic Coordination of Attention Heads for Improving Factual Consistency in LLMs

Shike Li,[1] Xiaokai Wang,[1] Xiaofeng Liu,[2] Xin Tong,[3] Hu Zhang[1]*

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China
[2]College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China
[3]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
{lsk, wangxiaokai}@sxu.edu.cn, liuxiaofeng03@tyut.edu.cn, tongxin@bupt.edu.cn, zhanghu@sxu.edu.cn

Abstract

Large language models (LLMs) frequently generate fluent yet factually inaccurate content, a phenomenon known as hallucination. Recent inference-time approaches aim to improve truthfulness by steering model activations toward semantically meaningful directions. While effective to some extent, these methods typically process activations independently, neglecting the internal coordination structure of multi-head attention (MHA), where attention heads interact to form semantic representations. In this work, we propose CoFact, an adaptive inference-time mechanism that improves factual consistency by dynamically coordinating attention head behaviors. Inspired by cooperative game theory, CoFact conceptualizes attention heads as collaborative agents. It models the semantic utility and redundancy of each head and adaptively modulates their contributions to the final attention output. Notably, rather than directly altering intermediate representations, CoFact performs token-level coordination to encourage diverse and complementary attention patterns across heads. CoFact is plug-and-play compatible with mainstream LLM architectures and requires no additional supervision or model retraining. Experimental results across multiple standard factuality benchmarks demonstrate that CoFact consistently enhances factual accuracy while maintaining generation fluency.

Introduction

Large language models (LLMs) have demonstrated impressive capabilities in generating coherent and contextually rich text (Yao et al. 2024). Despite their remarkable success, they remain prone to a persistent limitation known as hallucination, producing fluent but factually inaccurate content (Huang et al. 2025; Tonmoy et al. 2024). This issue is particularly critical in high-stakes domains such as healthcare, law, and education, where factual consistency is essential (Bai et al. 2024). Although substantial progress has been made to mitigate hallucination through external knowledge augmentation (Zhang et al. 2024; Sriramanan et al. 2024), post-hoc verification (Zhao et al. 2024; Yu, Jalaian, and Bastian 2024), or training time regularization (Li et al. 2024; Arteaga, Schön, and Pielawski 2024), these methods typically incur significant inference overhead, require additional supervision, or demand architectural modification factors that limit their applicability in black-box or resource-constrained settings.

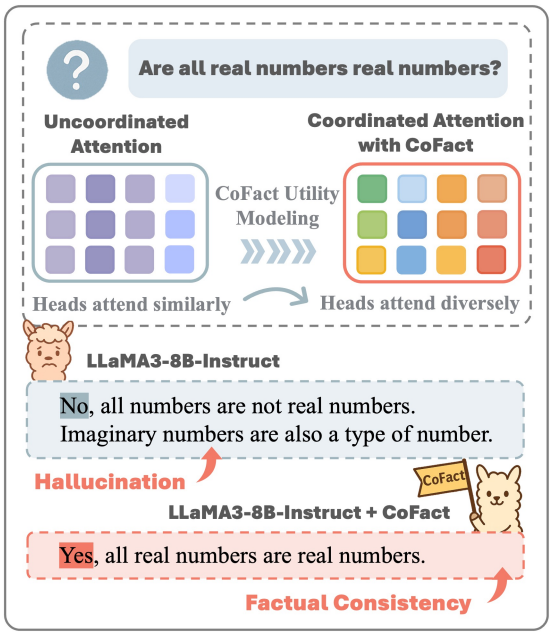


Figure 1: An illustration of how uncoordinated attention leads to hallucination. In LLaMA3-8B-Instruct, multiple heads attend to similar tokens, resulting in redundant focus and an incorrect answer ("No"). CoFact promotes diverse and complementary attention, yielding the correct fact ("Yes").

In contrast, inference-time intervention strategies (Li et al. 2023; Zhang, Yu, and Feng 2024; He et al. 2024) provide a more adaptive and modular alternative. However, their integration with the internal mechanisms of LLMs, particularly the collaborative behavior of multi-head attention (MHA) modules (Vaswani et al. 2017), remains relatively underexplored. Ideally, MHA is designed to enable attention heads to learn complementary and diverse representations of the input (Wang et al. 2022; Cordonnier, Loukas, and Jaggi 2020; Li et al. 2018). Yet through experimental analysis of attention behaviors (see Figure 1), we observe a striking pattern: Attention heads frequently focus on overlapping regions, even when processing semantically complex inputs. This behav-

*Corresponding author

ior leads to redundant information flow, reduces semantic diversity, and weakens inter-head collaboration, ultimately increasing the likelihood of hallucinated content.

We attribute this phenomenon to the absence of an adaptive mechanism that regulates the contribution of each attention head during inference. Without such coordination, heads with low semantic utility may amplify irrelevant signals, while redundant heads waste capacity by focusing on overlapping content. These observations motivate our central hypothesis: *Hallucination can be mitigated by adaptively modulating attention head influence based on their semantic utility and redundancy*. This hypothesis is further supported by a theoretical analysis, as detailed in Appendix A.

Motivated by these observations, we propose CoFact, an adaptive inference-time mechanism that improves the factual consistency of generated content by dynamically coordinating the behavior of attention heads. Unlike prior methods that intervene on isolated activations (Li et al. 2023), prune heads entirely (Voita et al. 2020), or inject fixed perturbations into the representation space (Zhang, Yu, and Feng 2024; He et al. 2024), CoFact preserves the full attention structure and introduces a structured redistribution scheme. This mechanism is inspired by cooperative game theory (Owen 2013), viewing attention heads as collaborative agents whose contributions should be adaptively modulated based on their semantic utility and redundancy. To operationalize this principle, CoFact comprises three coordinated components: a semantic valuator that estimates each head's contribution to the current token, a conflict arbiter that measures inter-head redundancy, and a coalition coordinator that integrates both signals to redistribute attention activations in a utility-aware manner. Extensive experiments on multiple open-source language models such as LLaMA2, LLaMA3, Mistral, and Gemma across factuality benchmarks including TruthfulQA, TriviaQA, and Natural Questions demonstrate that CoFact integrates effectively into existing models and consistently improves factual consistency without compromising fluency.

Summary of Contributions. (1) We propose CoFact, an adaptive inference-time mechanism that improves factual consistency by dynamically coordinating attention head behaviors. Its design is inspired by cooperative game theory, which models attention heads as collaborative agents guided by their semantic utility and redundancy. (2) We introduce a utility-aware redistribution framework composed of three components: a semantic valuator, a conflict arbiter, and a coalition coordinator. Together, these modules enable fine-grained modulation of attention during generation. (3) Experiments on multiple open-source LLMs and standard factuality benchmarks demonstrate that CoFact consistently enhances factual accuracy without compromising efficiency.

Related Work

Inference-Time Hallucination Mitigation

Existing approaches to mitigating hallucinations in LLMs can be broadly categorized into external knowledge augmentation (Agrawal et al. 2023; Andriopoulos and Pouwelse 2023; Guan et al. 2024; Mentzelopoulou 2024) and internal intervention methods (Li et al. 2023; Chuang et al. 2023; Zhou et al. 2024; Duan et al. 2025). Our work belongs to the latter, focusing on inference-time strategies that require no retraining or architectural modification. State-editing methods such as ITI (Li et al. 2023), TruthX (Zhang, Yu, and Feng 2024), LLM Factoscope (He et al. 2024), and TrFr (Chen et al. 2024b) steer generation through activation modification or projection constraints, but often depend on auxiliary supervision. Contrastive techniques like DoLa (Chuang et al. 2023) and ICD (Zhang et al. 2023) enhance semantic alignment but may degrade fluency. In contrast, CoFact introduces a coordination-based mechanism that adaptively redistributes attention activations, promoting diverse and complementary attention patterns to improve factual consistency.

Coordination and Redundancy in MHA

Prior studies have shown that attention heads often attend to similar patterns, resulting in redundancy, reduced expressiveness, and poorer generalization (Bian et al. 2021; Devlin et al. 2019; Xu et al. 2021; McGrath et al. 2023; Li, Chen, and Tong 2025). To address these issues, pruning-based methods (Cordonnier, Loukas, and Jaggi 2020; Michel, Levy, and Neubig 2019; Voita et al. 2020) remove redundant heads to improve efficiency, albeit at the cost of reduced model capacity. Diversity-promoting regularization (Li et al. 2018; Cui et al. 2019) introduces auxiliary losses to encourage differentiation across heads, though they may cause unstable training dynamics. Alternatively, cooperative optimization strategies (Wang et al. 2022; Peng et al. 2020; Shazeer et al. 2020; Tjandra et al. 2020; Peters et al. 2019) explicitly model inter-head interaction to enhance global attention expressiveness. Building on these insights, we develop an adaptive coordination mechanism that mitigates inter-head redundancy and encourages diverse, specialized attention, without modifying model architecture or training dynamics.

The CoFact Framework

In this section, we formalize the problem and introduce the CoFact framework, which enhance factual consistency through attention modulation. CoFact consists of three components: a semantic valuator that estimates token-wise head utility, a conflict arbiter that penalizes inter-head redundancy, and a coalition coordinator that redistributes attention based on both signals. Figure 2 illustrates the overall architecture.

Problem Setup

We draw inspiration from cooperative game theory and treat attention heads as collaborative agents that collectively contribute to token-level representations during generation. Let $H = \{h_1, h_2, \ldots, h_m\}$ denote the set of attention heads in a given layer. At each decoding step t, the model receives a hidden representation $X_{t-1} \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the hidden dimension. Each head $h_i \in H$ produces an intermediate output:

$$O_i^{(t)} = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \tag{1}$$

where $Q_i = X_{t-1} W_q^i$, $K_i = X_{t-1} W_k^i$, and $V_i = X_{t-1} W_v^i$ are the query, key, and value projections for head h_i, respec-

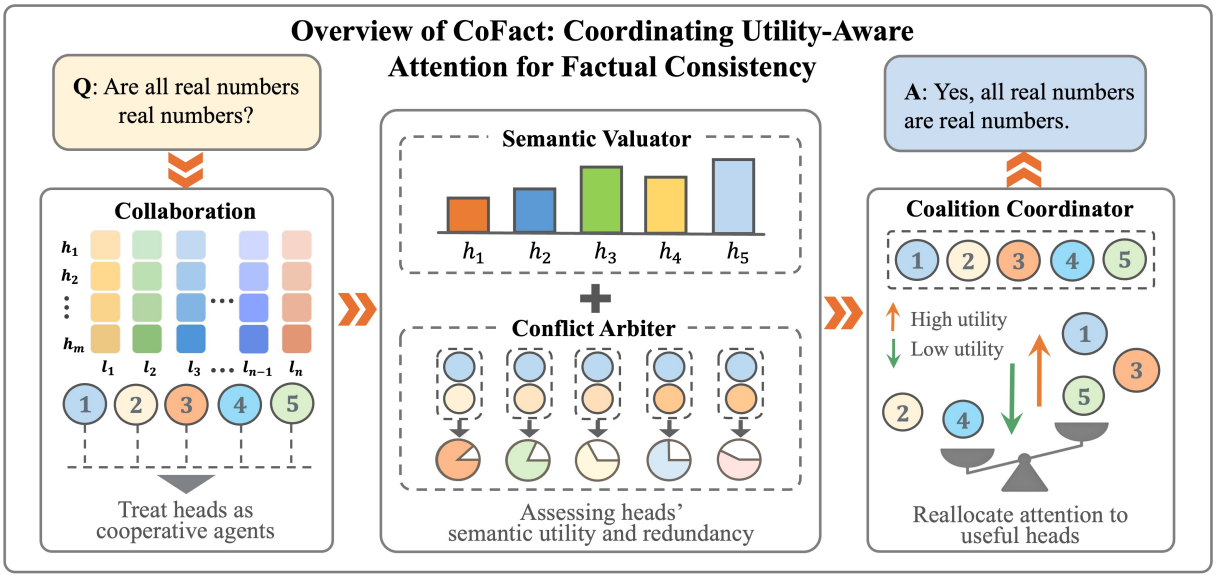


Figure 2: Overview of the CoFact framework. Given a user query, attention heads are treated as cooperative agents. CoFact jointly evaluates the semantic utility and redundancy of each head, and adaptively reallocates attention to promote diverse and complementary focus. This coordinated mechanism leads to factually consistent generation.

tively. Our goal is to compute a dynamic activation profile $\boldsymbol{\alpha}^{(t)} = \{\alpha_1^{(t)}, \ldots, \alpha_m^{(t)}\}$, where each $\alpha_i^{(t)} \in [0, 1]$ reflects the semantic utility and redundancy of head h_i at step t. The refined attention output is obtained by weighted aggregation:

$$O_t^{\text{CoFact}} = \sum_{i=1}^{m} \alpha_i^{(t)} \cdot O_i^{(t)}, \quad \text{with} \sum_i \alpha_i^{(t)} = 1. \tag{2}$$

Unlike static or learned weights fixed during training, CoFact infers $\boldsymbol{\alpha}^{(t)}$ adaptively at each step based on token-specific attention behaviors. The objective is to suppress redundant activations and promote semantic diversity, thereby improving factual consistency without modifying the model parameters or requiring retraining.

Semantic Valuator

To quantify the semantic utility of each attention head during inference, we define a contribution payoff that evaluates the marginal impact of a head on the layer's output representation. Specifically, for a given head h_i in layer l at decoding step t, its utility is computed as:

$$U_{\text{con}}^{(l,t)}(h_i) = \left\| O_t^{(l)} - O_{t,(-i)}^{(l)} \right\|_2^2, \tag{3}$$

where $O_t^{(l)}$ denotes the aggregated output of all heads in layer l at step t, and $O_{t,(-i)}^{(l)}$ is the output when h_i is excluded. This value captures how much the removal of h_i perturbs the representation. This formulation aligns with the notion of marginal contribution (Young 1985) in cooperative settings, where an agent's value is measured by its added utility to the group. It also echoes the intuition of mutual information $I(h_i; x)$, which quantifies how much information a head encodes about the input. Although we do not compute $I(h_i; x)$ explicitly, U_{con} serves as a tractable proxy for estimating a head's semantic distinctiveness. A higher $U_{\text{con}}^{(l,t)}(h_i)$ implies that h_i carries unique information that influences the token-level output. Conversely, a low score suggests redundancy or irrelevance, which may correlate with hallucination-prone behaviors. This dynamic, token-level evaluation lays the foundation for CoFact's attention reallocation.

Conflict Arbiter

While semantic contribution captures the individual utility of each head, effective coordination also requires discouraging excessive overlap among heads. To this end, we define a redundancy payoff function that quantifies the similarity between a given head and its peers in the same layer. Let $A_i^{(l,t)} \in \mathbb{R}^{n \times n}$ denote the attention matrix of head h_i at layer l and decoding step t, where n is the sequence length. The redundancy score for head h_i is computed as the average pairwise cosine similarity between its attention pattern and those of the remaining heads:

$$U_{\text{red}}^{(l,t)}(h_i) = \frac{1}{|H| - 1} \sum_{j \neq i} C_{i,j}^{(l,t)}, \tag{4}$$

where $C_{i,j}^{(l,t)}$ denotes the cosine similarity between $A_i^{(l,t)}$ and $A_j^{(l,t)}$, defined as the dot product of their vectorized forms:

$$C_{i,j}^{(l,t)} = \frac{\langle \text{vec}(A_i^{(l,t)}), \text{vec}(A_j^{(l,t)}) \rangle}{\|A_i^{(l,t)}\|_F \cdot \|A_j^{(l,t)}\|_F}, \tag{5}$$

$\text{vec}(\cdot)$ denotes the vectorization operator and $\|\cdot\|_F$ the Frobenius norm. Cosine similarity is used as it captures directional

alignment between attention patterns while being invariant to their scale, making it well-suited for identifying redundant attention behavior. A high redundancy score indicates that h_i attends to similar regions as other heads, suggesting limited contribution to representation diversity. By penalizing such redundancy, CoFact encourages more specialized and complementary attention behaviors across heads.

Coalition Coordinator

To consolidate semantic utility and redundancy into a unified activation scheme, we define a utility function for each attention head h_i in layer l at decoding step t as:

$$U^{(l,t)}(h_i) = \alpha \cdot U_{\text{con}}^{(l,t)}(h_i) - \beta \cdot U_{\text{red}}^{(l,t)}(h_i), \tag{6}$$

where α and β are non-negative hyperparameters balancing contribution and redundancy. This combined utility reflects the overall contextual value of each head with respect to the current token. To translate utility scores into operational behavior, we compute a softmax-based activation distribution:

$$\tilde{\alpha}_i^{(l,t)} = \frac{\exp(\lambda U^{(l,t)}(h_i))}{\sum_{j=1}^{m} \exp(\lambda U^{(l,t)}(h_j))}, \tag{7}$$

where λ is a temperature parameter that controls distribution sharpness. This softmax transformation normalizes utilities and emphasizes differences between high- and low-utility heads. It instantiates the abstract activation profile $\alpha_i^{(t)}$ from Problem Setup, now realized as layer-specific weights $\tilde{\alpha}_i^{(l,t)}$ based on utility. Semantically informative and non-redundant heads receive greater influence, while less useful ones are down-weighted. This enables differentiable, bounded redistribution without hard pruning. The final output of layer l at step t is computed via weighted aggregation:

$$O_t^{\text{CoFact},(l)} = \sum_{i=1}^{m} \tilde{\alpha}_i^{(l,t)} \cdot O_i^{(l,t)}. \tag{8}$$

This mechanism enables CoFact to dynamically reallocate attention based on per-token context, promoting diverse and specialized behaviors across heads. Hyperparameter choices (α, β, and λ) are detailed in Appendix B.

Experiments

We evaluate CoFact on multiple language models and standard factuality benchmarks to assess its effectiveness in improving factual consistency. This section outlines the evaluation setup and presents core experimental results.

Experiment Setup

Dataset and Evaluation Metrics. We evaluate our method on three factuality benchmarks: TruthfulQA (Lin, Hilton, and Evans 2021), TriviaQA (Joshi et al. 2017), and Natural Questions (NQ) (Kwiatkowski et al. 2019). TruthfulQA includes both multiple-choice and open-ended formats; for the former, we report accuracy, and for the latter, we follow its dual-metric protocol by reporting truthfulness (True), informativeness (Info), and their composite score (True*Info) as the main metric. We also include BLEURT (Sellam, Das, and Parikh 2020) scores to assess surface-level similarity with gold references. Following recent practice, we fine-tune a GPT-4o-Mini-based evaluator using publicly available annotations to replace the deprecated GPT-3 classifier. To assess generalization, we adopt TriviaQA and NQ, using 3,610 question subsets from each dataset as constructed in Li et al. 2023, and follow the factuality evaluation method proposed in Wang et al. 2025.

Models. We evaluate CoFact on a diverse set of widely-used open-source LLMs, including LLaMA2-7B-Chat and LLaMA2-13B-Chat (Touvron et al. 2023), LLaMA3-8B-Instruct (Dubey et al. 2024), Mistral-7B-Instruct-v0.2/v0.3 (Jiang et al. 2023), and Gemma2-9B-it (Team et al. 2024). These models span different sizes, training recipes, and architectural generations, enabling a comprehensive evaluation of CoFact's effectiveness and compatibility.

Baselines. We compare CoFact against several representative inference-time baselines. Base refers to the unmodified LLM without any intervention. ITI (Li et al. 2023) and its non-linear variant NL_ITI (Hoscilowicz et al. 2024) steer internal activations toward truthful semantics via directional edits. DoLa (Chuang et al. 2023) aligns hidden states through contrastive decoding. AD (Chen et al. 2024a) adjusts token probabilities based on the sharpness of in-context activations, while TruthFlow (Wang et al. 2025) detects hallucinations by tracing disruptions in hidden state propagation across layers.

Main Results

TruthfulQA Results Across Models. Table 1 reports the factuality evaluation results of CoFact and several baselines across six base models on the TruthfulQA benchmark. In the open-ended generation setting, CoFact consistently achieves the highest factual accuracy, with improvements over the base model ranging from 3.46% to 18.44% in the True score, and reaching up to 18.54% in the composite True*Info metric. It also increases BLEURT scores across all base models, indicating better semantic alignment with reference answers. Compared to inference-time methods such as ITI and DoLa, CoFact demonstrates superior truthfulness while maintaining strong informativeness. Although DoLa yields slightly higher informativeness in some cases, it often underperforms in truthfulness, reflecting a tendency toward verbose or less accurate outputs. In contrast, CoFact strikes a better balance by enhancing factual accuracy without compromising information richness or semantic quality, as reflected in both Info and BLEURT scores. This suggests that its coordination mechanism effectively promotes precise and relevant attention. In the multiple-choice setting, CoFact achieves competitive accuracy across models and outperforms the base method in five out of six cases. While the absolute gains in this format are moderate, the substantial improvements in open-ended tasks, where hallucination is more prevalent, highlight CoFact's robustness in enhancing factual reliability during generation.

Cross-Dataset Generalization. To evaluate generalization beyond TruthfulQA, we test CoFact on TriviaQA and NQ, using LLaMA3-8B-Instruct as the base model. As shown in

Model	Method	Open-ended Generation				MC Acc.
		BLEURT (%)	True (%)	Info (%)	True*Info (%)	(%)
LLaMA2-7B-Chat	Base	47.68	49.39	90.22	44.56	32.03
	+ DoLa	49.39	49.63	92.18	45.75	24.94
	+ AD	49.39	50.37	91.44	46.06	30.32
	+ ITI	48.90	48.17	89.49	43.11	30.81
	+ NL_ITI	45.48	42.79	89.49	38.29	31.30
	+ TruthFlow	**57.95**	59.41	92.42	54.91	34.47
	+ CoFact	55.63$^{+7.95}$	**67.83**$^{+18.44}$	**93.68**$^{+3.46}$	**59.86**$^{+15.30}$	**35.51**$^{+3.48}$
LLaMA2-13B-Chat	Base	56.23	56.23	93.89	52.79	28.12
	+ DoLa	53.55	55.10	92.42	50.84	25.92
	+ AD	53.55	55.26	91.93	50.80	28.36
	+ ITI	50.12	51.29	91.93	47.43	27.14
	+ NL_ITI	54.03	57.46	92.18	52.97	28.61
	+ TruthFlow	57.46	58.68	92.18	54.09	34.23
	+ CoFact	**62.04**$^{+5.81}$	**62.15**$^{+5.92}$	**94.79**$^{+0.90}$	**58.84**$^{+6.05}$	**36.70**$^{+8.58}$
LLaMA3-8B-Instruct	Base	51.34	52.32	91.69	47.97	32.76
	+ DoLa	52.08	55.50	91.69	50.89	25.18
	+ AD	46.70	46.21	81.66	37.74	28.36
	+ ITI	51.83	54.52	90.46	49.32	35.45
	+ NL_ITI	55.26	54.52	90.71	49.46	36.19
	+ TruthFlow	62.59	64.79	94.38	61.15	41.08
	+ CoFact	**63.93**$^{+12.59}$	**65.13**$^{+12.81}$	**95.68**$^{+3.99}$	**62.56**$^{+14.59}$	**44.41**$^{+11.65}$
Mistral-7B-Instruct-v0.2	Base	65.04	75.31	98.78	74.39	47.43
	+ DoLa	62.35	73.84	98.53	72.75	36.19
	+ AD	65.28	76.28	99.02	75.53	44.74
	+ ITI	65.77	72.13	98.53	71.07	46.70
	+ NL_ITI	64.55	72.37	98.04	70.95	44.74
	+ TruthFlow	67.24	78.48	97.80	76.75	49.39
	+ CoFact	**69.14**$^{+4.10}$	**79.07**$^{+3.76}$	**98.82**$^{+0.04}$	**77.26**$^{+2.87}$	**49.66**$^{+2.23}$
Mistral-7B-Instruct-v0.3	Base	61.86	71.39	**98.04**	69.99	**47.43**
	+ DoLa	63.81	72.37	**98.04**	70.95	37.41
	+ AD	62.35	75.79	97.07	73.57	42.54
	+ ITI	60.88	67.48	95.35	64.34	43.52
	+ NL_ITI	60.88	66.99	97.07	65.03	43.77
	+ TruthFlow	67.48	77.26	96.82	74.80	46.70
	+ CoFact	**69.91**$^{+8.05}$	**77.34**$^{+5.95}$	97.31$^{-0.73}$	**75.18**$^{+5.19}$	46.95$^{-0.48}$
Gemma2-9B-it	Base	62.35	64.30	90.71	58.33	35.21
	+ DoLa	61.61	66.26	92.42	61.24	30.56
	+ AD	62.35	66.01	89.00	58.75	32.76
	+ ITI	63.81	66.50	92.42	61.46	36.43
	+ NL_ITI	56.23	57.70	84.84	48.95	34.47
	+ TruthFlow	68.95	76.53	95.84	73.35	44.01
	+ CoFact	**74.28**$^{+11.93}$	**80.19**$^{+15.89}$	**96.53**$^{+5.82}$	**76.87**$^{+18.54}$	**45.58**$^{+10.37}$

Table 1: Factuality results on TruthfulQA across different base models. For open-ended generation, we report BLEURT, True, Info, and their composite (True*Info) as main metrics. MC Acc. refers to multiple-choice accuracy. Superscripts indicate the relative improvement of CoFact over the Base method. Results for some baselines are obtained from Wang et al. 2025.

Table 2, CoFact achieves the highest factuality scores across both datasets, reaching 65.31% on TriviaQA and 59.91% on NQ, which are 1.29% and 2.13% higher than the base model, respectively. For the composite True*Info metric, CoFact also surpasses the base model by 1.35% on TriviaQA and 1.50% on NQ, consistently outperforming all other baselines. Compared to ITI, which exhibits noticeable degradation across both datasets, CoFact demonstrates stronger robustness to distributional shift. While TruthFlow performs competitively, its gains remain smaller and less stable. These results confirm that CoFact generalizes well across QA formats and maintains its effectiveness in promoting factual reliability across different question types and knowledge domains. Notably, all improvements are achieved without model fine-tuning or external supervision, underscoring CoFact's plug-and-play flexibility. This suggests its potential for broader deployment in real-world LLM applications facing open-domain and dynamic query distributions.

Method	Metric	TriviaQA	NQ
Base	True	64.02	57.78
	True*Info	55.43	50.36
ITI	True	58.56	49.22
	True*Info	46.85	38.07
TruthFlow	True	64.90	58.01
	True*Info	56.49	51.21
CoFact	True	**65.31**$^{+1.29}$	**59.91**$^{+2.13}$
	True*Info	**56.78**$^{+1.35}$	**51.86**$^{+1.50}$

Table 2: Generalization to TriviaQA and NQ with LLaMA3. Superscripts indicate the relative improvement of CoFact over the Base method.

Analyses

To deepen our understanding of CoFact's effectiveness, we analyze its influence on attention behavior. Specifically, we examine: (1) The impact of activation modulation on factuality. (2) The structural changes induced in attention patterns across tokens and layers. (3) The robustness and necessity of its core components. These analyses offer further insight into CoFact's mechanism and design.

Effect of Head Modulation on Factuality

We evaluate CoFact against three activation strategies on TruthfulQA using LLaMA2-7B-Chat: the original model, Uniform weighting (equal contribution from all heads), and Top-k selection (activating the top 16 heads per layer).

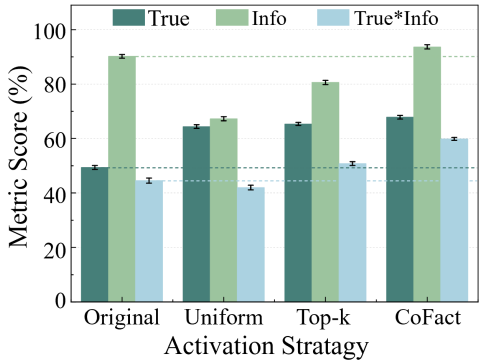
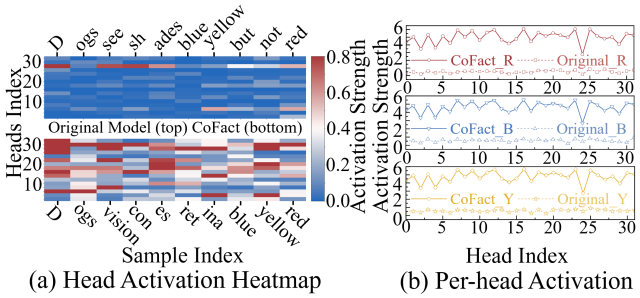


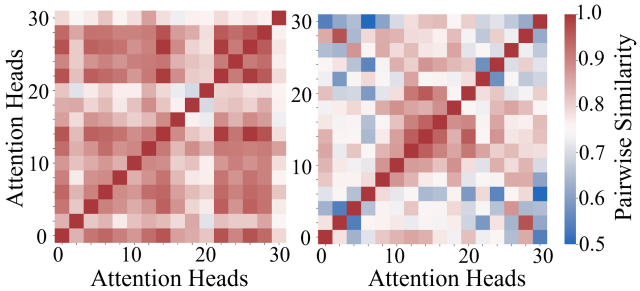
Figure 3: TruthfulQA factuality results under different attention activation strategies using LLaMA2-7B-Chat. CoFact achieves the best balance of True and Info.

As shown in Figure 3, CoFact achieves the best overall performance, with a True*Info score of 59.86%, compared to 44.56% for the original model, 42.01% for Uniform, and 50.82% for Top-k. While Uniform slightly improves truthfulness to 64.39%, it significantly reduces informativeness to 67.28% due to indiscriminate aggregation. Top-k reduces informativeness to 80.62%, and also has a lower truthfulness score due to lack of adaptive flexibility. In contrast, CoFact achieves both high truthfulness at 67.83% and informativeness at 93.68% by dynamically emphasizing informative heads and suppressing redundant ones. These results confirm that context-aware coordination outperforms static or uniform activation schemes in mitigating hallucinations.



(a) Head Activation Heatmap (b) Per-head Activation

Figure 4: Token-level attention behavior under "What colors do dogs see?".



(a) Original Model (b) CoFact

Figure 5: Pairwise similarity of attention heads in an intermediate layer during generation.

Token-Level and Inter-Head Attention Analysis

To understand how CoFact reshapes attention behaviors during generation, we conduct a fine-grained analysis on a representative prompt *"What colors do dogs see?"*, using the LLaMA2-7B-Chat model with and without CoFact. We examine both token-level activation dynamics and intra-layer head interactions. Full generation outputs for this example are provided in Appendix C (Example 2).

Token-Level Activation Behavior. Figure 4 visualizes head activation dynamics in the final layer across generated tokens. Figure 4(a) shows heatmaps of head activation strengths for both the original model and CoFact, while Figure 4(b) plots per-head activation trends for three key tokens: "red," "blue," and "yellow." In Figure 4(a), the original model displays uniformly low and homogeneous activations, with most heads focusing weakly and similarly across tokens, reflecting redundancy and lack of specialization. In contrast, CoFact produces more concentrated, token-specific activations, where different heads respond distinctly to semantically informative tokens such as "blue" and "yellow". Figure 4(b) provides a closer look at activation strength distributions across heads for three color-related tokens. CoFact amplifies the response of high-utility heads while suppressing inactive or redundant ones, resulting in greater head-level differentiation and clearer semantic attribution.

Setting	%True	%Info	%T*I
CoFact + Base	67.83	93.68	59.86
CoFact + Noisy Prefix	60.34	87.79	47.41
CoFact + Token Noise	65.44	89.67	54.52
CoFact + Gaussian Noise	66.07	90.63	55.08

Table 3: Robustness of CoFact under input and attention perturbations (TruthfulQA, LLaMA2-7B-Chat, 50 samples).

Intra-Layer Head Redundancy. Figure 5 presents pairwise cosine similarities between attention heads in an intermediate layer when generating a representative token. The visualizations illustrate how attention is distributed across heads within a layer, revealing critical insights into redundancy. In the original model, head activations are highly similar, with most pairwise cosine similarities exceeding 0.85. This redundancy suggests limited functional diversity and excessive focus overlap across heads, contributing to semantic inefficiency and hallucination risk. CoFact exhibits significantly reduced inter-head similarity. Many attention head pairs show moderate to low similarity, with values frequently below 0.75. This pattern reflects CoFact's ability to encourage specialization, promoting complementary attention flows that support more nuanced and accurate generation.

Robustness to Attention Perturbation

We evaluate CoFact's performance under three types of inference-time perturbations, using the first 50 prompts from TruthfulQA and LLaMA2-7B-Chat as the base model. Across all scenarios, CoFact shows graceful degradation and consistently outperforms the base model, demonstrating resilience against both input- and attention-level perturbations. Table 3 summarizes the results.

Noisy Prompt Prefix. We prepend a misleading clause *"Some people believe the wrong answer to this question."* to each input prompt. This setup introduces semantic ambiguity and tests a model's resistance to input-induced hallucination. CoFact remains highly robust under this setting, achieving a True score of 60.34% and an Info score of 87.79%, comparable to the unperturbed baseline.

Random Token Replacement. We randomly replace each token in the prompt with a probability of 0.1, simulating corrupted input semantics. Despite the noise, CoFact achieves 65.44% True and 89.67% Info, with a True*Info score of 54.52%, indicating its ability to preserve core factual alignment under partial input corruption.

Gaussian Noise on Attention. We inject Gaussian noise into the attention values to perturb the attention distribution. CoFact demonstrates graceful degradation, maintaining 66.07% True and 90.63% Info. This suggests that CoFact's utility-aware redistribution strategy provides inherent robustness by down-weighting unstable or noisy attention heads.

Ablation Study

We conduct an ablation study by removing either the semantic valuator or the conflict arbiter, two core components in CoFact. The former estimates token-wise head utility, while the latter penalizes inter-head redundancy. Table 4 presents the results across six base models. Removing either module consistently leads to performance degradation, particularly in the composite Truth*Info metric. For example, on LLaMA2-7B-Chat, disabling utility scoring or redundancy penalization reduces the Truth*Info score from 59.86% to 46.23% and 45.11%, respectively. Similar trends hold across larger models: on LLaMA3-8B-Instruct, the composite score drops by over 8% under either ablation, and on Mistral-v0.2, removing any single module results in a 10% reduction. Even in models where degradation is less pronounced (e.g., Mistral-v0.3 and Gemma2-9B-it), the full CoFact configuration consistently outperforms both ablated variants. The ablation confirms that factual gains in CoFact arise from the synergy between semantic discrimination and structured attention diversity.

Model	Variant	%True	%Info	%T*I
LLaMA2-7B-Chat	w/o Re.	58.10	86.53	45.11
	w/o U.	58.74	85.93	46.23
	CoFact	**67.83**	**93.68**	**59.86**
LLaMA2-13B-Chat	w/o Re.	55.65	92.60	52.11
	w/o U.	54.42	**95.71**	51.73
	CoFact	**62.15**	94.79	**58.84**
LLaMA3-8B-Instruct	w/o Re.	60.45	91.21	54.02
	w/o U.	61.20	92.97	52.76
	CoFact	**65.13**	**95.68**	**62.56**
Mistral-v0.2	w/o Re.	72.22	94.82	66.51
	w/o U.	71.59	95.31	67.04
	CoFact	**79.07**	**98.82**	**77.26**
Mistral-v0.3	w/o Re.	69.05	90.56	65.59
	w/o U.	73.81	92.19	67.10
	CoFact	**77.34**	**97.31**	**75.18**
Gemma2-9B-it	w/o Re.	74.05	92.56	69.59
	w/o U.	73.81	93.19	68.10
	CoFact	**80.19**	**96.53**	**76.87**

Table 4: Ablation results of CoFact with core modules removed. "w/o Re." removes redundancy penalization; "w/o U." removes semantic utility scoring.

Conclusion

In this work, we present CoFact, a plug-and-play inference-time mechanism that enhances factual consistency in large language models by dynamically coordinating attention head behaviors. Grounded in cooperative game theory, CoFact evaluates each head's semantic utility and redundancy, and reallocates attention without modifying model weights or requiring additional training. Extensive experiments across diverse model families and factuality benchmarks demonstrate that CoFact consistently enhances factual accuracy while maintaining fluency. Fine-grained analyses further reveal that CoFact improves token-level selectivity and reduces redundant head activations, validating the effectiveness of its utility-aware coordination strategy. To improve factual grounding, we will extend CoFact by integrating retrieval-based signals into the coordination mechanism.

Acknowledgments

We thank all the anonymous reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Foundation of China (62476161).

References

Agrawal, G.; Kumarage, T.; Alghamdi, Z.; and Liu, H. 2023. Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv preprint arXiv:2311.07914*.

Andriopoulos, K.; and Pouwelse, J. 2023. Augmenting LLMs with Knowledge: A survey on hallucination prevention. *arXiv preprint arXiv:2309.16459*.

Arteaga, G. Y.; Schön, T. B.; and Pielawski, N. 2024. Hallucination detection in llms: Fast and memory-efficient finetuned models. *arXiv preprint arXiv:2409.02976*.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Bian, Y.; Huang, J.; Cai, X.; Yuan, J.; and Church, K. 2021. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, 930–945.

Chen, S.; Xiong, M.; Liu, J.; Wu, Z.; Xiao, T.; Gao, S.; and He, J. 2024a. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*.

Chen, Z.; Sun, X.; Jiao, X.; Lian, F.; Kang, Z.; Wang, D.; and Xu, C. 2024b. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20967–20974.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Cordonnier, J.-B.; Loukas, A.; and Jaggi, M. 2020. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*.

Cui, H.; Iida, S.; Hung, P.-H.; Utsuro, T.; and Nagata, M. 2019. Mixed Multi-Head Self-Attention for Neural Machine Translation. *EMNLP-IJCNLP 2019*, 206.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186.

Duan, J.; Kong, F.; Cheng, H.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Zhu, X.; Shi, X.; and Xu, K. 2025. TruthPrInt: Mitigating LVLM Object Hallucination Via Latent Truthful-Guided Pre-Intervention. *arXiv preprint arXiv:2503.10602*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Guan, X.; Liu, Y.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18126–18134.

He, J.; Gong, Y.; Lin, Z.; Wei, C.; Zhao, Y.; and Chen, K. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, 10218–10230.

Hoscilowicz, J.; Wiacek, A.; Chojnacki, J.; Cieslak, A.; Michon, L.; Urbanevych, V.; and Janicki, A. 2024. Nl-iti: Optimizing probing and intervention for improvement of iti method. *Preprint at arXiv. https://doi. org/10.48550/arXiv*, 2403.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, G.; Chen, Y.; and Tong, H. 2025. Taming Knowledge Conflicts in Language Models. *arXiv preprint arXiv:2503.10996*.

Li, J.; Consul, S.; Zhou, E.; Wong, J.; Farooqui, N.; Ye, Y.; Manohar, N.; Wei, Z.; Wu, T.; Echols, B.; et al. 2024. Banishing LLM hallucinations requires rethinking generalization. *arXiv preprint arXiv:2406.17642*.

Li, J.; Tu, Z.; Yang, B.; Lyu, M. R.; and Zhang, T. 2018. Multi-Head Attention with Disagreement Regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2897–2903.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

McGrath, T.; Rahtz, M.; Kramar, J.; Mikulik, V.; and Legg, S. 2023. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*.

Mentzelopoulou, A. 2024. *Reducing LLM Hallucinations with Retrieval Prompt Engineering*. Ph.D. thesis, Delft University of Technology.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Owen, G. 2013. *Game theory*. Emerald Group Publishing.

Peng, H.; Schwartz, R.; Li, D.; and Smith, N. A. 2020. A Mixture of h-1 Heads is Better than h Heads. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6566–6577.

Peters, M. E.; Neumann, M.; Logan, R.; Schwartz, R.; Joshi, V.; Singh, S.; and Smith, N. A. 2019. Knowledge Enhanced Contextual Word Representations. 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)*, 43–54.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Shazeer, N.; Lan, Z.; Cheng, Y.; Ding, N.; and Hou, L. 2020. Talking-heads attention. *arXiv preprint arXiv:2003.02436*.

Sriramanan, G.; Bharti, S.; Sadasivan, V. S.; Saha, S.; Kattakinda, P.; and Feizi, S. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37: 34188–34216.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Sessa, P. G.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.-C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J.-B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; Mao-Jones, J.; Lee, K.; Yu, K.; Millican, K.; Sjoesund, L. L.; Lee, L.; Dixon, L.; Reid, M.; Mikuła, M.; Wirth, M.; Sharman, M.; Chinaev, N.; Thain, N.; Bachem, O.; Chang, O.; Wahltinez, O.; Bailey, P.; Michel, P.; Yotov, P.; Chaabouni, R.; Comanescu, R.; Jana, R.; Anil, R.; McIlroy, R.; Liu, R.; Mullins, R.; Smith, S. L.; Borgeaud, S.; Girgin, S.; Douglas, S.; Pandya, S.; Shakeri, S.; De, S.; Klimenko, T.; Hennigan, T.; Feinberg, V.; Stokowiec, W.; hui Chen, Y.; Ahmed, Z.; Gong, Z.; Warkentin, T.; Peran, L.; Giang, M.; Farabet, C.; Vinyals, O.; Dean, J.; Kavukcuoglu, K.; Hassabis, D.; Ghahramani, Z.; Eck, D.; Barral, J.; Pereira, F.; Collins, E.; Joulin, A.; Fiedel, N.; Senter, E.; Andreev, A.; and Kenealy, K. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

Tjandra, A.; Liu, C.; Zhang, F.; Zhang, X.; Wang, Y.; Synnaeve, G.; Nakamura, S.; and Zweig, G. 2020. Deja-vu: Double feature presentation and iterated loss in deep transformer networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6899–6903. IEEE.

Tonmoy, S.; Zaman, S.; Jain, V.; Rani, A.; Rawte, V.; Chadha, A.; and Das, A. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Voita, E.; Talbot, D.; Moiseev, F.; Titov, I.; and Sennrich, R. 2020. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5797–5808.

Wang, H.; Cao, B.; Cao, Y.; and Chen, J. 2025. TruthFlow: Truthful LLM Generation via Representation Flow Correction. In *Forty-second International Conference on Machine Learning*.

Wang, H.; Shen, X.; Tu, M.; Zhuang, Y.; and Liu, Z. 2022. Improved transformer with multi-head dense collaboration. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 2754–2767.

Xu, Y.; Huang, H.; Feng, C.; and Hu, Y. 2021. A supervised multi-head self-attention network for nested named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14185–14193.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Young, H. P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2): 65–72.

Yu, C.-E. J.; Jalaian, B.; and Bastian, N. D. 2024. Mitigating Large Vision-Language Model Hallucination at Post-hoc via Multi-agent System. In *Proceedings of the AAAI Symposium Series*, volume 4, 110–113.

Zhang, S.; Pan, L.; Zhao, J.; and Wang, W. Y. 2024. The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 2025–2038.

Zhang, S.; Yu, T.; and Feng, Y. 2024. TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8908–8949.

Zhang, Y.; Cui, L.; Bi, W.; and Shi, S. 2023. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Zhao, X.; Zhang, H.; Pan, X.; Yao, W.; Yu, D.; Wu, T.; and Chen, J. 2024. Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 8702–8718.

Zhou, G.; Yan, Y.; Zou, X.; Wang, K.; Liu, A.; and Hu, X. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.