

Factual Decoding via Head Coordination Graphs in Large Language Models

Shike Li¹, Xiaokai Wang¹, Xin Tong², Xiaofeng Liu^{3*}, Hu Zhang^{1*}

¹School of Computer and Information Technology, Shanxi University, Taiyuan, China

²School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China

³College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China

{lsk,wangxiaokai,zhanghu}@sxu.edu.cn, tongxin@bupt.edu.cn, liuxiaofeng03@tyut.edu.cn

Abstract

Hallucinations remain challenging for large language models (LLMs), especially in factual generation. Existing mitigation methods often rely on retrieval, supervision, representation editing, or output-level decoding, while paying limited attention to how factual support is internally coordinated across attention heads. In this work, we study factual decoding from a head-level coordination perspective and find that unreliable generations are often associated with fragmented and unstable use of factual anchors across heads. To improve such coordination, we propose Head Coordination Graphs (HCG), a plug-and-play inference-time framework that models attention heads as interacting nodes. HCG uses explicit factual anchors as observable support signals to construct sparse head graphs and applies risk-gated propagation to adaptively regulate cross-head influence without parameter updates, additional training, or external knowledge. Experiments across multiple factuality and hallucination benchmarks demonstrate broad improvements in factual consistency across open-source LLMs, while analyses and ablations demonstrate the importance of head-level coordination, sparse graph construction, and risk-gated propagation.

1 Introduction

Large language models (LLMs) have achieved strong performance across diverse natural language processing tasks (Li et al., 2025; Yu et al., 2025). Nevertheless, they still frequently generate fluent but factually incorrect responses, known as hallucinations, which undermine reliability in high-stakes applications (Zhang et al., 2025b). Existing mitigation methods include retrieval- and knowledge-augmented generation (Song et al., 2024; Hu et al., 2025; Yan et al., 2025c), fact-level revision (Yan et al., 2025a,b), alignment fine-tuning (Deng et al.,

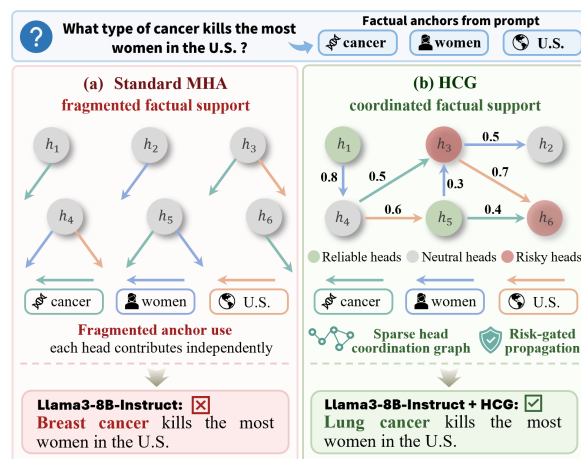


Figure 1: **Illustration of HCG for factual decoding.** Standard MHA computes head outputs in parallel and combines them without explicitly coordinating their use of factual anchors. HCG instead constructs a sparse graph over heads and applies risk-gated propagation to coordinate factual support during decoding.

2025), representation editing (Li et al., 2023b; Zhang et al., 2024), and contrastive decoding (Chuang et al., 2024; Jiang et al., 2025). Despite these advances across different intervention levels, the coordination of factual support across attention heads within multi-head attention (MHA) remains underexplored (Vaswani et al., 2017).

MHA consists of multiple parallel attention heads that capture diverse semantic information, such as entity relations, contextual dependencies, reasoning cues, and long-range associations (Kim et al., 2019; Gong et al., 2021; Michel et al., 2019; Voita et al., 2019). This diversity is central to Transformer expressiveness and provides a natural unit for analyzing internal information processing. However, within each MHA block, head outputs are computed in parallel and combined through a fixed output projection, without an explicit mechanism for adapting their relative influence according

*Corresponding authors: liuxiaofeng03@tyut.edu.cn and zhanghu@sxu.edu.cn

to the factual support they provide. Factual errors may therefore also arise from fragmented or insufficiently coordinated support across attention heads.

To make this internal coordination process observable, we use explicit factual anchors, such as entities, dates, numbers, and other fact-related tokens, as interpretable support signals. Our analysis shows that hallucinated and non-hallucinated responses do not necessarily differ substantially in their average attention to factual anchors; instead, hallucinated responses tend to exhibit stronger cross-head dispersion and more unstable anchor utilization. This suggests that factual errors may depend not only on whether relevant factual cues are attended to, but also on whether attention heads coordinate their use of these cues.

Motivated by this observation, we propose head coordination graph (HCG), an inference-time framework for factual decoding. HCG models attention heads as interacting nodes and uses anchors to construct sparse head graphs that capture complementary factual support among heads. It then applies risk-gated propagation to softly regulate cross-head influence, allowing heads with weak support or higher risk to integrate information from well-supported and complementary neighboring heads. HCG requires no parameter updates, training, or external knowledge. Rather than forcing all heads to behave uniformly, it performs sparse, selective, and risk-aware coordination while preserving head diversity.

We evaluate HCG on multiple open-source LLMs and factuality benchmarks, including TruthfulQA, HaluEval-QA, TriviaQA, Natural Questions, CQA2, and QASC. Results show broad factuality improvements across different models and task settings. Further analyses provide an interpretable view of internal support coordination during factual generation.

Summary of Contributions. (1) We study factual hallucination from the perspective of head-level support coordination and characterize the relationship between unreliable generation and unstable factual support across attention heads. (2) We introduce Head Coordination Graphs (HCG), an inference-time framework that uses factual anchors as observable support signals for sparse and risk-aware cross-head propagation. (3) We evaluate HCG across multiple models and benchmarks, and provide mechanistic analyses and ablation studies that characterize the roles of head coordination, graph structure, and risk-gated propagation.

2 Related Work

Inference-time hallucination mitigation. Activation intervention has become an important direction for mitigating hallucinations in LLMs (Chen et al., 2024a; Sriramanan et al., 2024; Orgad et al., 2025; Huang et al., 2025). State-editing methods such as ITI (Li et al., 2023b), TruthX (Zhang et al., 2024), TrFr (Chen et al., 2024b), and TruthFlow (Wang et al., 2025) guide generation by modifying activation states or representation flows, with some approaches relying on auxiliary supervision. Contrastive decoding methods, including DoLa (Chuang et al., 2024) and ICD (Zhang et al., 2025a), improve factuality by comparing representations or logits across layers during decoding. Unlike these representation-level or logit-level interventions, HCG targets the coordination of attention heads during inference, regulating cross-head influence through sparse graph propagation.

Attention-head analysis and coordination. Prior studies have documented diverse and sometimes redundant behaviors among individual attention heads, motivating research on head redundancy and pruning (Cordonnier et al., 2020; Michel et al., 2019; Voita et al., 2019), diversity regularization (Li et al., 2018; Cui et al., 2019), and inter-head interaction modeling (Wang et al., 2022; Peng et al., 2020; Peters et al., 2019). These studies show that heads can specialize in different linguistic or semantic functions and that modeling their interactions can improve efficiency or representation quality. However, existing head-level methods primarily target training-time optimization, architectural modification, or model compression rather than factuality-oriented intervention during decoding. In contrast, HCG treats attention heads as graph nodes, uses factual anchors as observable support signals, and performs sparse, risk-gated propagation to coordinate head-level support without updating model parameters.

3 The Proposed HCG Method

In this section, we formulate factual decoding from a head-level support coordination perspective (§ 3.1) and present HCG for sparse, risk-gated coordination among attention heads (§ 3.2).

3.1 Problem Formulation

Let $x = (x_1, \dots, x_n)$ denote an input sequence, and consider a pretrained decoder-only language model that generates the next token y_t conditioned

Overview of Head Coordination Graphs (HCG)

HCG coordinates head-level factual support during decoding using factual anchors from the prompt

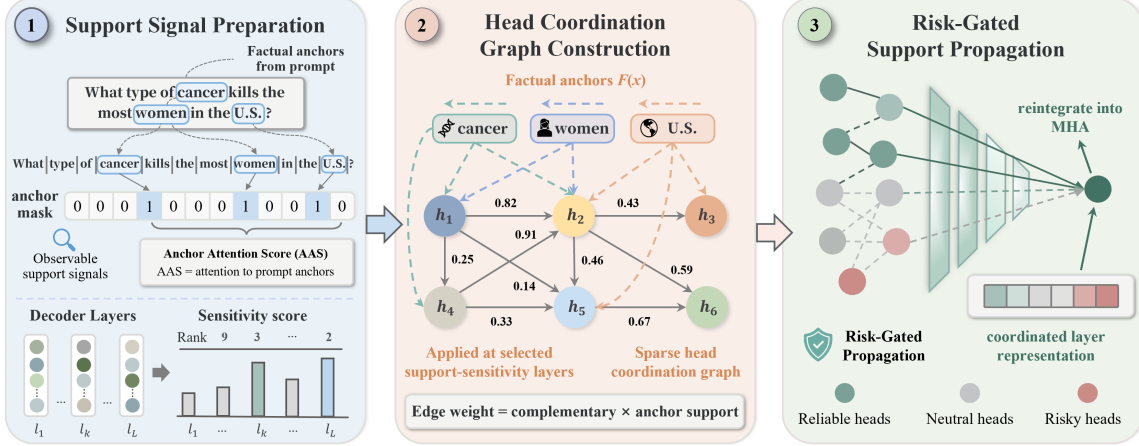


Figure 2: **Overview of the HCG framework.** HCG first extracts factual anchors from the prompt as observable support signals, then constructs a sparse coordination graph over attention heads based on complementary anchor use, and finally applies risk-gated propagation to regulate cross-head influence.

on x and the previously generated tokens $y_{<t}$:

$$p_{\theta}(\cdot | x, y_{<t}) = \text{Softmax}(f_{\theta}(x, y_{<t})), \quad (1)$$

where f_{θ} produces the next-token logits.

At decoding step t , the multi-head self-attention module at layer l computes an attention distribution $A_t^{(l,h)} \in \mathbb{R}^{1 \times (n+t-1)}$ for each head $h \in \{1, \dots, H\}$ over all key positions. Each head produces an output representation:

$$z_t^{(l,h)} = A_t^{(l,h)} V_t^{(l,h)} \in \mathbb{R}^{d_h}, \quad (2)$$

which contributes to the layer output through concatenation and linear projection.

We study factual decoding from the perspective of head-level support coordination. During generation, different heads may provide complementary but uneven support for the next-token prediction. To make such support observable, we define a set of factual-anchor positions $\mathcal{F}(x) \subseteq \{1, \dots, n\}$, corresponding to tokens that provide explicit factual cues, such as named entities, dates, or numerical expressions. We define the anchor indicator as $m_u = \mathbb{I}(u \in \mathcal{F}(x))$ for $u \leq n$, and $m_u = 0$ for $u > n$. The anchor attention score (AAS) for head (l, h) at decoding step t is defined as:

$$\text{AAS}_t^{(l,h)} = \sum_{u=1}^n A_t^{(l,h)}[u] m_u, \quad (3)$$

AAS quantifies how strongly a head attends to observable factual anchors and serves as an operational measure of head-level factual support.

Within standard MHA, head outputs are computed in parallel and combined through a fixed output projection, without explicitly adapting their relative contributions according to the factual support they provide. Consequently, heads with weak or unstable factual support may influence the layer output without coordination with better-supported or complementary heads. Our objective is therefore to transform the set of head representations:

$$\{z_t^{(l,1)}, \dots, z_t^{(l,H)}\} \rightarrow \{\tilde{z}_t^{(l,1)}, \dots, \tilde{z}_t^{(l,H)}\}, \quad (4)$$

so as to coordinate head-level support through adaptive cross-head influence while preserving the functional diversity of attention heads. The coordinated multi-head representation before output projection is then formed as:

$$\tilde{Z}_t^{(l)} = \text{Concat}(\tilde{z}_t^{(l,1)}, \dots, \tilde{z}_t^{(l,H)}). \quad (5)$$

3.2 Head Coordination Graph Execution

HCG centers on constructing sparse coordination graphs over attention heads and performing risk-gated support propagation along the resulting graphs. Figure 2 provides an overview of the HCG workflow. At each decoding step, HCG first derives anchor-based support signals and restricts coordination to support-sensitive layers. It then constructs a sparse head coordination graph whose edges capture complementary factual support among heads. Finally, HCG applies risk-gated propagation to softly regulate cross-head influence

and reintegrates the updated head representations into the original attention module, leaving subsequent computations unchanged. We describe these components below.

Support Signal Preparation. Given an input sequence x , HCG constructs a binary anchor mask $m \in \{0, 1\}^n$ from the factual-anchor positions $\mathcal{F}(x)$, where $m_u = 1$ for selected anchors such as named entities, dates, or numerical expressions. These anchors can be extracted using lightweight rules or off-the-shelf token-level taggers. Rather than constraining generation, the mask provides an observable support signal for computing each head’s anchor attention score in Eq. (3).

Since not all layers are equally affected by factual anchors, HCG further restricts coordination to highly support-sensitive layers. For each layer $l \in \{1, \dots, L\}$, we compute a support sensitivity score S_l as the expected change in its hidden representation when factual anchors are perturbed:

$$S_l = \mathbb{E}_t \left[\left\| h_t^l(x, y_{<t}) - h_t^l(x_{\setminus \mathcal{F}(x)}, y_{<t}) \right\|_2 \right], \quad (6)$$

where $h_t^l(\cdot)$ denotes the layer- l hidden state at decoding step t , and $x_{\setminus \mathcal{F}(x)}$ is obtained by replacing factual anchors with neutral placeholders while keeping the generated prefix $y_{<t}$ unchanged. A larger S_l indicates that layer l is more sensitive to factual support signals. We select the K_l layers with the highest support sensitivity scores:

$$\mathcal{L} = \text{TopK}(\{S_l\}_{l=1}^L, K_l). \quad (7)$$

HCG is applied only to layers in $\mathcal{L}$. This focused intervention reduces unnecessary computation and avoids indiscriminate coordination in layers less related to factual support.

Head Coordination Graph Construction. For each selected layer l , HCG models H attention heads as graph nodes $\mathcal{H}^l = \{1, \dots, H\}$. To characterize how each head distributes its attention over factual anchors, we define an *anchor coverage pattern* (ACP) by masking and renormalizing its attention distribution over anchor tokens. For each position $u \in \{1, \dots, n\}$, the ACP is defined as:

$$\text{ACP}_{t,h}^l(u) = \frac{A_t^{(l,h)}[u] \cdot m_u}{\sum_{j=1}^n A_t^{(l,h)}[j] \cdot m_j + \varepsilon}, \quad (8)$$

where $\varepsilon > 0$ is a small constant for numerical stability. If a head assigns zero attention mass to

all anchors, its incident edge weights are set to zero. We measure the complementarity between two heads h and h' as:

$$\text{comp}_{t,h,h'}^l = 1 - \cos(\text{ACP}_{t,h}^l, \text{ACP}_{t,h'}^l), \quad (9)$$

which captures whether the two heads attend to different subsets of factual anchors. To combine complementarity with the anchor support of both heads, their edge weight is defined as:

$$w_{t,h,h'}^l = \text{comp}_{t,h,h'}^l \cdot \text{AAS}_t^{(l,h)} \cdot \text{AAS}_t^{(l,h')}. \quad (10)$$

Finally, for each head, we L_2 -normalize its outgoing edge weights, retain edges whose normalized weights exceed a threshold τ , and row-normalize the retained weights to obtain a sparse propagation matrix $\mathbf{P}_t^l \in \mathbb{R}^{H \times H}$. Rows with no surviving edges are set to zero. This defines the head coordination graph $G_t^l = (\mathcal{H}^l, \mathbf{P}_t^l)$.

Risk-Gated Support Propagation. Given the head coordination graph G_t^l , HCG performs a single round of risk-gated propagation. For each head h , we estimate its risk from the relative attention assigned to generated tokens and factual anchors:

$$g_t^{(l,h)} = \alpha \cdot \frac{A_{\text{gen},t}^{(l,h)}}{A_{\text{anch},t}^{(l,h)} + A_{\text{gen},t}^{(l,h)} + \varepsilon}, \quad (11)$$

where $A_{\text{anch},t}^{(l,h)} = \sum_{u=1}^n A_t^{(l,h)}[u] m_u$ and $A_{\text{gen},t}^{(l,h)} = \sum_{u=n+1}^{n+t-1} A_t^{(l,h)}[u]$ denote the attention mass assigned to factual anchors and previously generated tokens, respectively, and α controls the propagation strength. The neighboring representations are aggregated as:

$$\tilde{z}_t^{(l,h)} = \sum_{h'=1}^H \mathbf{P}_t^l[h, h'] z_t^{(l,h')}, \quad (12)$$

The head representation is then updated by interpolation:

$$\tilde{z}_t^{(l,h)} = z_t^{(l,h)} + g_t^{(l,h)} \left(\tilde{z}_t^{(l,h)} - z_t^{(l,h)} \right). \quad (13)$$

For heads with no surviving neighbors, we set $g_t^{(l,h)} = 0$ so that their representations remain unchanged during propagation. Intuitively, low-risk heads largely preserve their original representations, whereas higher-risk heads integrate signals from anchor-supported and complementary neighbors. The updated representations replace the original head outputs before concatenation, while subsequent computations remain unchanged.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate HCG on benchmarks covering factual generation, hallucination QA, open-domain factual QA, and knowledge-intensive reasoning under diverse task formats. Our primary evaluation is conducted on TruthfulQA (Lin et al., 2021), where we report truthfulness (True), informativeness (Info), and their composite score (T*I). To assess generalization, we further evaluate HCG on HaluEval-QA (Li et al., 2023a), TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), CommonsenseQA 2.0 (CQA2) (Talmor et al., 2021), and QASC (Khot et al., 2020).

Models. We evaluate HCG on open-source instruction-tuned LLMs across architectures and scales: Llama2-7B-Chat, Llama2-13B-Chat (Touvron et al., 2023), Llama3-8B-Instruct, Llama3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct v0.2 and v0.3 (Jiang et al., 2023), and Vicuna-7B-v1.5 (Chiang et al., 2023).

Baselines. We compare HCG with representative inference-time methods. Contrastive decoding includes DoLa (Chuang et al., 2024), which contrasts decoding behaviors across layers to refine token selection. Representation editing includes ITI (Li et al., 2023b), TruthX (Zhang et al., 2024), and TruthFlow (Wang et al., 2025), which modify internal representation flows, often using auxiliary supervision or intervention directions. For open-ended generation, we follow Wang et al. (2025) and use GPT-4 to evaluate True, Info, and T*I.

Implementation Details. All HCG experiments use zero-shot prompting on a single NVIDIA A800 GPU (80 GB). We extract factual anchors with the spaCy `en_core_web_lg` model from named entities and tokens tagged as PROP, NOUN, NUM, VERB, or ADJ, mapping their character spans to model subwords by offset overlap. The resulting mask is applied only to prompt positions, with zero anchor weight assigned to previously generated tokens. HCG performs one propagation round per decoding step with no cross-step state. Unless varied explicitly, we use $K_l = 3$, $\alpha = 0.5$, and $\tau = 0.4$ across datasets and backbones. HCG_{avg} uses fixed-weight propagation and HCG_{rg} uses the head-specific risk gate; otherwise, they are identical. HCG denotes HCG_{rg} in cross-dataset experiments.

4.2 Main Results

Hallucination Mitigation. Table 1 reports the main results on TruthfulQA across four instruction-tuned language models, evaluated by True, Info, and their combined score T*I. Compared with the base models, both HCG variants improve T*I across all four backbones while maintaining or improving informativeness. Contrastive decoding yields smaller and less consistent gains, often trading off between True and Info. Representation editing methods such as ITI, TruthX, and TruthFlow achieve stronger results in some settings, but their performance varies across models, and they often require auxiliary supervision or learned intervention directions. HCG_{avg} improves T*I on every evaluated backbone, indicating the benefit of head-level coordination. HCG_{rg} further achieves the highest T*I score on all four models, showing the additional benefit of risk-gated propagation. Overall, these results demonstrate that HCG improves factuality across the evaluated backbones while preserving informativeness.

Cross-Dataset Generalization. As shown in Table 2, HCG improves performance across the majority of evaluated datasets and model backbones. On HaluEval-QA, HCG yields gains across all Llama variants and most other backbones, although it produces a slight decline on Mistral-v0.3. On the open-domain factual QA benchmarks TriviaQA and NQ, HCG improves every evaluated model, with particularly pronounced gains on NQ for several backbones. These results indicate that anchor-based head coordination transfers beyond TruthfulQA to broader factual question answering settings. For knowledge-intensive reasoning, HCG also improves performance in most settings on CQA2 and QASC. The larger gains on QASC for several models suggest that coordinating complementary factual support can benefit questions involving multiple facts. Overall, HCG provides broad cross-dataset improvements, although the effect sizes remain model- and task-dependent.

5 Analyses

5.1 Factual Anchor Perturbation Analysis

Since HCG relies on factual anchors as observable support signals, we first examine whether they provide informative support for factual decoding. We conduct controlled perturbation experiments on TruthfulQA with Llama3-8B-Instruct using three

Method	Llama2-7B-Chat			Llama2-13B-Chat			Llama3-8B-Instruct			Mistral-7B-Instruct-v0.2		
	True ↑	Info ↑	T*I ↑	True ↑	Info ↑	T*I ↑	True ↑	Info ↑	T*I ↑	True ↑	Info ↑	T*I ↑
LM	49.57	90.24	43.08	57.64	91.39	50.77	61.52	89.14	52.00	70.41	87.99	61.36
<i>Contrastive Decoding</i>												
DoLa	48.35	91.25	44.41	56.69	91.12	50.81	63.42	91.36	54.52	69.49	89.12	60.73
<i>Representation Editing</i>												
ITI	47.97	87.36	42.63	57.20	90.47	51.98	60.72	90.39	51.33	70.98	88.17	61.05
TruthX	55.20	88.24	50.89	-	-	-	-	-	-	<u>74.43</u>	90.11	<u>65.28</u>
TruthFlow	<u>57.41</u>	<u>92.91</u>	<u>52.32</u>	<u>58.82</u>	91.76	<u>52.51</u>	<u>65.17</u>	92.09	<u>61.81</u>	72.28	89.45	63.14
<i>Head Coordination</i>												
HCG _{avg}	52.18	93.00	48.86	57.95	<u>91.83</u>	52.19	63.42	<u>92.20</u>	60.12	70.51	94.51	63.38
HCG _{rg}	59.71	91.88	54.51	61.49	93.20	56.35	68.74	94.77	62.91	75.57	<u>92.78</u>	67.91

Table 1: Results on TruthfulQA under open-ended generation, evaluated by True (%), Info (%), and their combined score T*I (%). All methods are inference-time approaches evaluated under identical decoding settings without prompt tuning. Best results are shown in bold, and second-best results are underlined.

Backbone	Method	Hallucination QA	Open-domain Factual QA		Knowledge-intensive Reasoning	
		HaluEval-QA	TriviaQA	NQ	CQA2	QASC
Llama2-7B-Chat	LM	52.85	48.25	31.05	43.13	32.29
	+ HCG	58.12 ↑5.27	54.88 ↑6.63	39.09 ↑8.04	46.34 ↑3.21	48.28 ↑15.99
Llama2-13B-Chat	LM	48.27	59.22	42.22	57.81	43.60
	+ HCG	52.46 ↑4.19	62.34 ↑3.12	49.95 ↑7.73	59.60 ↑1.79	51.49 ↑7.89
Llama3-8B-Instruct	LM	43.78	60.28	42.02	56.12	59.50
	+ HCG	52.08 ↑8.30	62.11 ↑1.83	50.73 ↑8.71	60.43 ↑4.31	71.38 ↑11.88
Llama3.1-8B-Instruct	LM	63.01	50.69	42.47	62.42	66.95
	+ HCG	66.57 ↑3.56	60.14 ↑9.45	52.95 ↑10.48	65.82 ↑3.40	72.60 ↑5.65
Vicuna-7B-v1.5	LM	55.43	39.20	25.28	50.53	43.95
	+ HCG	56.00 ↑0.57	54.10 ↑14.90	38.17 ↑12.89	50.53	50.65 ↑6.70
Mistral-7B-v0.2	LM	40.63	56.45	38.56	63.99	64.66
	+ HCG	41.22 ↑0.59	57.70 ↑1.25	47.40 ↑8.84	64.92 ↑0.93	66.09 ↑1.43
Mistral-7B-v0.3	LM	54.31	61.30	44.32	65.75	65.12
	+ HCG	54.13	62.99 ↑1.69	49.58 ↑5.26	67.37 ↑1.62	66.77 ↑1.65

Table 2: Cross-dataset evaluation of HCG across hallucination QA, open-domain factual QA, and knowledge-intensive reasoning benchmarks. Results are reported in accuracy (%) for multiple-choice datasets and factual consistency scores for open-ended generation tasks. Improvements over the base model are shown with ↑.

complementary settings: (i) perturbing factual anchors, non-anchor tokens, or random tokens in the input prompt; (ii) replacing the HCG support mask with random, non-anchor, or all-token masks; and (iii) introducing mask noise by dropping anchors, adding false positives, or shifting entity spans.

Table 3 reports the results. At the input level, perturbing factual anchors produces the largest decline in T*I. Although True increases, the concurrent sharp drop in Info suggests that this likely reflects uninformative or abstention-style responses after key prompt tokens are masked. The larger performance degradation relative to the evaluated non-anchor and random-token perturbations confirms the importance of factual anchors for informative generation. Among the mask-source controls, the clean anchor mask achieves the highest T*I, indicat-

ing that HCG benefits from meaningful anchor selection rather than arbitrary token marking. Noisy masks yield lower T*I than the clean mask but remain above the base LM, suggesting that moderate anchor-detection errors reduce, rather than eliminate, the gains from HCG.

Overall, these results support factual anchors as operational signals for head coordination and further demonstrate the practical importance of accurate and selective anchor masks.

5.2 Head-Level Coordination Dynamics

We examine how HCG changes the coordination of factual support across attention heads during decoding. The analysis considers two aspects: cross-head dispersion in anchor attention and the magnitude of risk-gated head corrections.

Setting	True $\uparrow$	Info $\uparrow$	T*I $\uparrow$
A. Input perturbation on the base LM			
LM w/ original prompt	61.52	89.14	52.00
LM w/ anchor-perturbed prompt	83.72	18.36	10.04
LM w/ non-anchor-perturbed prompt	73.19	52.19	27.91
LM w/ random-token-perturbed prompt	77.23	35.01	18.97
B. Mask-source controls for HCG			
HCG w/ anchor mask	68.74	94.77	62.91
HCG w/ random mask	66.07	86.29	52.73
HCG w/ non-anchor mask	66.56	86.16	53.22
HCG w/ all-token mask	65.58	85.80	52.48
C. Noisy anchor masks for HCG			
HCG w/ Drop20 anchors	58.16	92.63	54.73
HCG w/ Add5 false positives	58.40	95.04	55.30
HCG w/ Shift spans +1	57.67	95.01	54.44

Table 3: Factual anchor perturbation analysis on TruthfulQA with Llama3-8B-Instruct. Group A perturbs prompt tokens, Group B changes the HCG support mask, and Group C drops 20% of anchors, adds five false-positive tokens, or shifts spans by one token.

Anchor Attention Dispersion. To quantify how unevenly heads attend to factual anchors, we introduce *anchor attention dispersion* (AAD). At decoding step t in layer l , let $a_{t,h}^l = \text{AAS}_t^{(l,h)}$ denote the anchor attention mass of head h , and let $\bar{a}_t^l = H^{-1} \sum_{h=1}^H a_{t,h}^l$ denote its cross-head mean. We then define AAD as:

$$\text{AAD}_t^l = \left[\frac{1}{H} \sum_{h=1}^H \left(a_{t,h}^l - \bar{a}_t^l \right)^2 \right]^{1/2}. \quad (14)$$

Higher AAD indicates a less balanced distribution of anchor attention across heads. For each response, we average AAD over the first ω generated tokens, with $\omega = 20$ fixed across all responses. This aggregation captures persistent dispersion while reducing sensitivity to individual decoding steps. We analyze 817 matched pairs of hallucinated and non-hallucinated TruthfulQA generations under identical decoding settings.

Figure 3 compares windowed AAD under standard MHA and HCG. Under standard MHA, hallucinated generations exhibit a descriptively heavier upper tail than non-hallucinated generations, consistent with more uneven use of factual anchors across heads. After applying HCG, the overall distributions remain broadly comparable to those under standard MHA. This suggests that HCG does not globally homogenize anchor-attention patterns but largely preserves the diversity of head behavior.

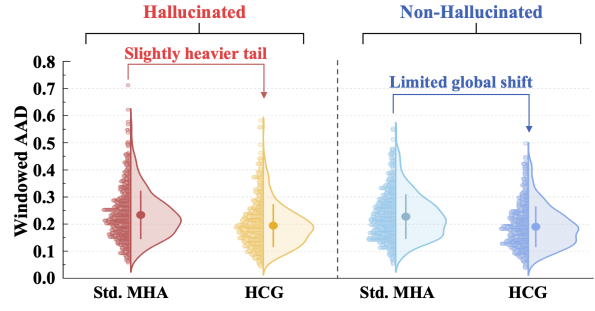


Figure 3: Windowed AAD across attention heads. Under standard MHA, hallucinated generations show slightly heavier-tailed distributions, indicating more fragmented anchor use. HCG largely preserves the overall AAD distribution, suggesting selective coordination rather than global reshaping of anchor attention.

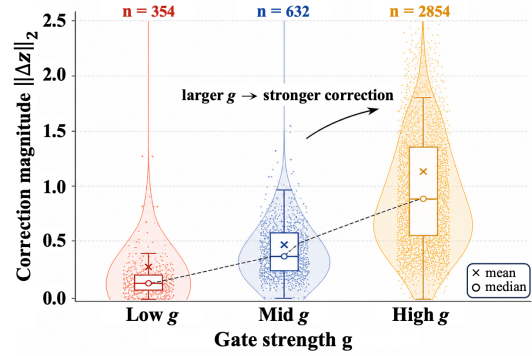


Figure 4: Gate strength g versus head correction magnitude Δ . Larger gate values yield stronger head-level corrections, indicating that HCG adaptively regulates cross-head influence by estimated risk.

Risk-Gated Head Correction. We further examine how risk-gated propagation modulates head representations. For each head, we measure its gate strength g and correction magnitude $\Delta = \|\tilde{z} - z\|_2$. As shown in Figure 4, correction magnitudes increase from low- to high-gate groups. This pattern is consistent with the intended role of the risk gate in controlling the strength of cross-head updates.

Table 4 reports problem-level statistics for hallucinated and non-hallucinated responses. Gate strengths and correction magnitudes are first aggregated over decoding steps and attention heads within each response. Hallucinated responses exhibit descriptively higher gate activation and larger corrections, suggesting stronger coordination when factual support is less stable. Together with the AAD analysis, these observations are consistent with HCG operating through localized, risk-aware propagation rather than globally reshaping anchor-attention distributions.

Statistic	Hallucination	Non-hallucination
Mean g	0.319	0.308
Median g	0.354	0.331
Mean Δ	1.074	0.984
Median Δ	0.862	0.758
Std(Δ)	0.924	0.907

Table 4: Risk-gated coordination statistics for hallucinated and non-hallucinated responses. Statistics are aggregated over decoding steps and attention heads within each problem. g denotes the gate strength, and $\Delta = \|\tilde{z} - z\|_2$ denotes the head correction magnitude.

Setting	True $\uparrow$	Info $\uparrow$	T*I $\uparrow$
<i>Layer selection strategy</i>			
HCG@All Layers	60.83	92.18	55.48
HCG@Random- K_l	62.05	93.16	59.09
HCG@Top- K_l (Ours)	68.74	94.77	62.91
<i>Number of selected layers</i>			
$K_l = 3$	68.74	94.77	62.91
$K_l = 6$	64.79	93.53	58.35
$K_l = 9$	64.40	93.16	57.67
$K_l = 12$	62.60	93.06	56.83

Table 5: Layer selection and sensitivity analysis on TruthfulQA with Llama3-8B-Instruct. The upper block compares layer selection strategies for controlling HCG coordination scope, while the lower block varies the number of selected support-sensitive layers.

5.3 Layer Selection and Sensitivity

Layer Selection. Factual-anchor perturbations affect layers unevenly, and applying HCG to all layers may introduce unnecessary interference. We therefore restrict coordination to the Top- K_l layers ranked by the support sensitivity score in Eq. (6).

Table 5 compares layer selection strategies on TruthfulQA with Llama3-8B-Instruct. Top- K_l selection achieves the best performance across all metrics. Coordinating all layers causes a clear performance drop, while random selection also underperforms, supporting the use of sensitivity-based coordination. We further vary the number of selected layers K_l . Performance is highest with a small layer set and gradually declines as K_l increases, indicating that focused intervention is preferable to broad coordination.

Gate Scaling Sensitivity. We analyze the sensitivity of HCG to the gate scaling coefficient α on TruthfulQA with Llama3-8B-Instruct. As shown in Table 6, setting $\alpha = 0$ disables risk-gated propagation. Performance improves as α increases from 0.1 to 0.5, reaches its highest level at $\alpha = 0.5$,

α	True (%) $\uparrow$	Info (%) $\uparrow$	T*I (%) $\uparrow$
0.0	62.58	87.51	51.34
0.1	62.33	90.47	55.26
0.3	64.05	92.28	58.19
0.5	68.74	94.77	62.91
0.7	60.20	86.23	52.64

Table 6: Sensitivity to gate scaling coefficient α on TruthfulQA with Llama3-8B-Instruct. The default value used in all experiments is $\alpha = 0.5$.

Variant	True $\uparrow$	Info $\uparrow$	T*I $\uparrow$
w/o Graph	60.82	87.76	53.21
w/o Propagation	62.58	87.51	51.34
w/o AAS Modulation	59.58	93.63	55.69
w/o Anchor Complementarity	56.56	97.79	55.69
Dense Graph	55.56	97.55	54.83
HCG (Full)	68.74	94.77	62.91

Table 7: Ablation study of HCG components on TruthfulQA with Llama3-8B-Instruct. AAS denotes anchor attention score for edge modulation.

Backbone	Peak Memory (MB)		Latency (s/sample)	
	Base	HCG	Base	HCG
Llama3-8B	15,915	16,931	1.45	2.19/2.24/2.29
Mistral-7B	14,449	14,999	1.49	2.42/2.47/2.54

Table 8: Peak GPU memory and average inference latency. The three HCG latency values correspond to $K_l = 1/2/3$, respectively.

and declines at $\alpha = 0.7$. This trend suggests that weak propagation limits cross-head coordination, whereas overly strong propagation may introduce excessive interference. Accordingly, we use $\alpha = 0.5$ as the default setting.

5.4 Ablation Analysis

Table 7 reports the component ablations of HCG. Removing the head graph or support propagation substantially reduces T*I, showing that both graph structure and message passing contribute to HCG. Removing AAS-based edge modulation or anchor complementarity also degrades performance, indicating that graph construction benefits from both anchor-support strength and complementary anchor coverage. The dense graph further underperforms its sparse counterpart, consistent with unnecessary connections introducing noisy cross-head influence. Overall, the full HCG achieves the best performance, demonstrating the joint contribution of sparse graph construction, anchor-aware edge weighting, and risk-gated propagation.

5.5 Efficiency Analysis

We evaluate the computational overhead of HCG on Llama3-8B-Instruct and Mistral-7B under the same hardware and decoding settings as their base models. We measure peak GPU memory and average inference latency per sample, with the results reported in Table 8.

HCG increases peak memory by 1,016 MB for Llama3-8B and 550 MB for Mistral-7B. Depending on the number of coordinated layers, inference latency increases by approximately 51–58% and 62–70%, respectively. The additional memory is mainly required for graph construction and intermediate head representations, while the latency overhead arises from per-token graph construction and propagation. Although HCG avoids training and parameter updates, its computational cost is incurred throughout decoding. These results characterize a clear trade-off between improved factuality and inference efficiency.

6 Conclusion

In this work, we study factual decoding in large language models from a head-level coordination perspective. Our analyses suggest an association between unreliable generation and uneven use of observable factual anchors across attention heads. We propose Head Coordination Graphs (HCG), a plug-and-play inference-time framework that constructs sparse head graphs from anchor-based support signals and applies risk-gated propagation to regulate cross-head influence during autoregressive decoding. Across multiple LLMs and factuality benchmarks, HCG delivers broad improvements in factual consistency without parameter updates, additional training, or external knowledge. Further analyses support the contributions of factual anchors, support-sensitive layer selection, sparse graph construction, and risk-gated coordination.

Limitations

HCG is a training-free inference-time method, but its effectiveness depends on the availability and quality of prompt-level factual anchors. Our perturbation analysis shows that inaccurate masks reduce performance, although moderate noise retains part of the gains. When anchors are absent, sparse, or ambiguous, or when relevant evidence is implicit, the resulting support signals may be less informative. Since HCG coordinates existing internal support rather than introducing new knowledge,

it may be less effective for errors caused by missing parametric knowledge, underspecified prompts, or reasoning failures not reflected in the anchors. Our evaluation focuses on factual tasks, leaving the preservation of general non-factual capabilities unverified. HCG also incurs additional memory and latency costs during decoding. Finally, anchor identification, layer selection, graph sparsification, and risk-gated propagation rely on interpretable heuristics that may not be optimal across all tasks or model families. Future work could explore adaptive anchor discovery, learned graph construction, and integration with external knowledge.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62476161 and the Major Program of the National Social Science Fund of China under Grant No. 24&ZD227.

References

- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: LLMs’ internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.
- Zhongzhi Chen, Xingwu Sun, Xianfeng Jiao, Fengzong Lian, Zhanhui Kang, Di Wang, and Chengzhong Xu. 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, pages 20967–20974.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS Org blog.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.
- Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*.
- Hongyi Cui, Shohei Iida, Po-Hsuan Hung, Takehito Utsuro, and Masaaki Nagata. 2019. Mixed multi-head

- self-attention for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 206–214.
- Jianfeng Deng, Qingfeng Chen, Debo Cheng, Jiuyong Li, and Lin Liu. 2025. Logit space constrained fine-tuning for mitigating hallucinations in LLM-based recommender systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29311–29324.
- Hongyu Gong, Yun Tang, Juan Pino, and Xian Li. 2021. Pay better attention to attention: Head selection in multilingual and multi-domain sequence modeling. *Advances in Neural Information Processing Systems*, 34:2668–2681.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7B*. 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*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.
- Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.
- Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2897–2903.
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.
- Kenneth Li, Oam Patel, Fernanda Vi  gas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *ICLR*.
- Hao Peng, Roy Schwartz, Dianqi Li, and Noah A Smith. 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*, pages 6566–6577.
- Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 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)*, pages 43–54.
- Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558.
- Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.
- Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. *Advances in Neural Information Processing Systems*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.
- Hanyu Wang, Bochuan Cao, Yuanpu Cao, and Jinghui Chen. 2025. TruthFlow: Truthful LLM generation via representation flow correction. *arXiv preprint arXiv:2502.04556*.
- Huadong Wang, Xin Shen, Mei Tu, Yimeng Zhuang, and Zhiyuan Liu. 2022. Improved transformer with multi-head dense collaboration. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2754–2767.
- Zhichao Yan, Jiaoyan Chen, Jiapu Wang, Xiaoli Li, Ru Li, and Jeff Z Pan. 2025a. Decomposing and revising what language models generate. *arXiv preprint arXiv:2509.00765*.
- Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Jiye Liang, Ru Li, and Jeff Z. Pan. 2025b. Atomic fact decomposition helps attributed question answering. *IEEE Transactions on Knowledge and Data Engineering*, 37(12):6959–6972.
- Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Yanyan Wang, Hongye Tan, Jiye Liang, Xiaoli Li, Ru Li, and Jeff Z Pan. 2025c. Prompting large language models with partial knowledge for answering questions with unseen entities. *arXiv preprint arXiv:2508.01290*.
- Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. 2025. A survey on trustworthy LLM agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6216–6226.
- Shaolei Zhang, Tian Yu, and Yang Feng. 2024. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.
- Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2025a. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8233–8247.
- Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjuan Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. 2025b. LLM hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):481–503.

A Qualitative Case Study

We present a qualitative case study illustrating how HCG dynamically coordinates attention heads during autoregressive decoding. Figure 5 shows the evolution of head-level coordination for the factual question “What is the spiciest part of a chili pepper?”. HCG-guided decoding produces the response “The spiciest part of a chili pepper is the placenta.”, allowing us to examine how the coordination graph changes as the answer is generated.

For visualization, the generation process is divided into three descriptive stages. **Lexical formation** covers function words and intermediate subword pieces; **entity grounding** includes content words such as “chili” and “pepper”; and **factual resolution** corresponds to the answer-bearing subwords “plac” and “enta”. For each stage, we visualize representative head graphs from selected decoding steps in the final transformer layer.

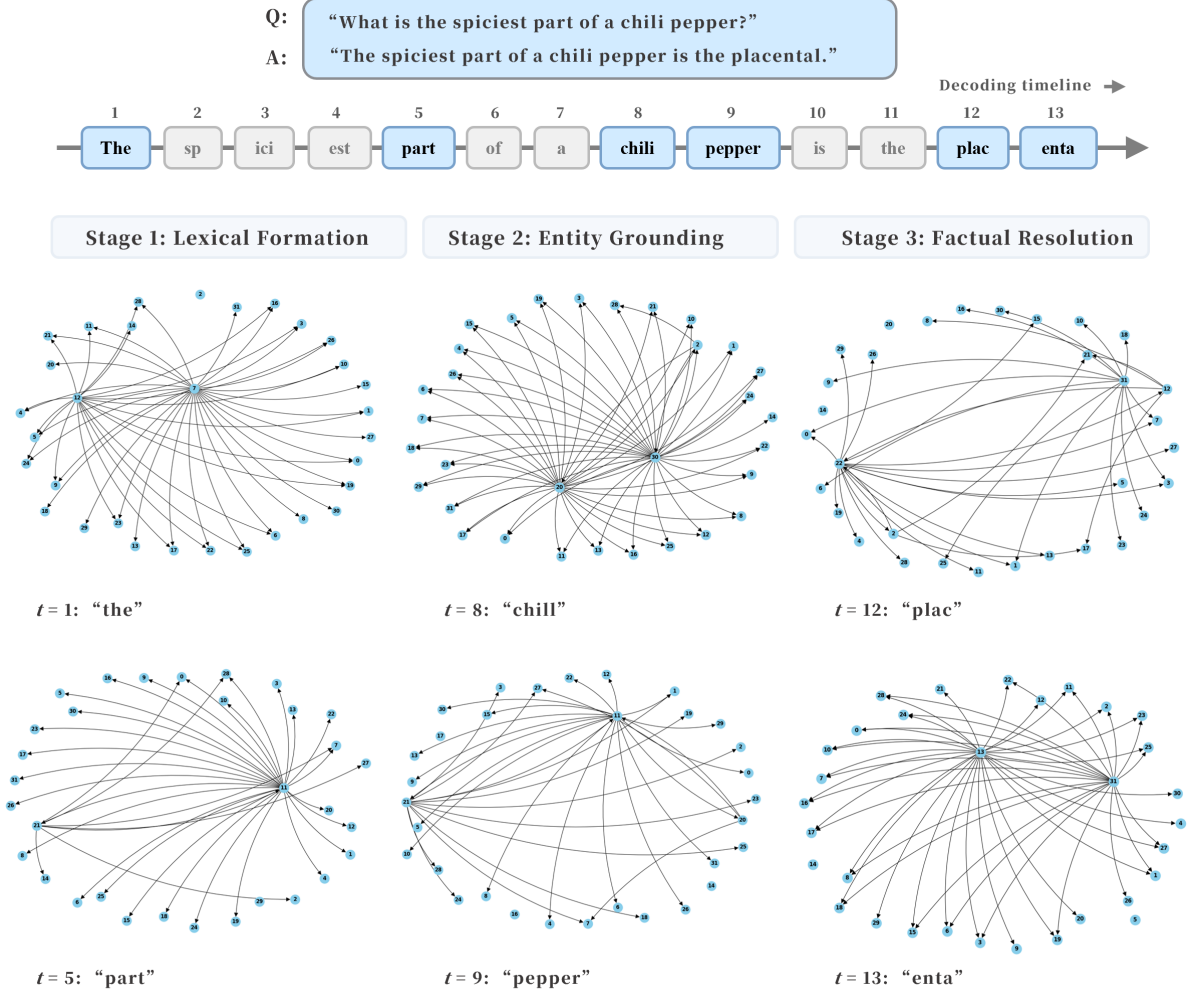


Figure 5: Qualitative example of head-level coordination under HCG in the final transformer layer. The top shows generated tokens along the decoding timeline, while the bottom presents risk-gated head graphs at representative steps from lexical formation, entity grounding, and factual resolution. Nodes denote attention heads, and directed edges indicate active cross-head coordination.

The graph topology and prominent heads vary across decoding steps. During lexical formation, coordination is organized around a small subset of heads. As entity-related tokens are generated, the active coordination pattern shifts to different heads, followed by another reconfiguration during the generation of the final answer. This example illustrates that HCG does not impose a fixed or uniform interaction pattern across heads; instead, it produces decoding-step-dependent coordination patterns as generation progresses.

B Layer-wise Support Sensitivity

Figure 6 visualizes the support sensitivity score S_l defined in Eq. (6) across layers and model backbones. Sensitivity is non-uniform across the network, with larger representation shifts generally

appearing in middle-to-upper layers and relatively small shifts in lower layers. The locations of the most sensitive layers also vary across backbones, supporting the use of sensitivity-based selection rather than a fixed layer range. Together with the layer selection results in Section 5.3, this analysis shows that HCG focuses coordination on layers whose representations are most responsive to factual-anchor perturbation.

C Detailed Efficiency Analysis

To complement the overall efficiency results in Section 5.5, we provide detailed GPU memory measurements and inference latency at different generation lengths. All measurements use the same hardware and decoding settings for baseline decoding (BL) and HCG.

Metric	Llama3-BL	Llama3-HCG	Mistral-BL	Mistral-HCG
Memory before forward (MB)	15,737	15,737	14,245	14,245
Peak memory (MB)	15,915	16,931	14,449	14,999
Peak increase over BL (MB)	–	1,016	–	550
Relative peak increase	–	6.38%	–	3.81%

Table 9: Detailed GPU memory usage for baseline decoding (BL) and HCG. Relative increases are calculated against the peak memory of the corresponding baseline. Memory measurements obtained with nvidia-smi include CUDA caching allocator overhead.

Generation Length	Llama3-BL	Llama3-HCG	Mistral-BL	Mistral-HCG
10 tokens	0.32	1.48	0.34	1.52
30 tokens	1.13	2.40	1.12	2.42
50 tokens	1.45	3.09	1.49	2.94

Table 10: Inference latency in seconds at different generation lengths. HCG uses the default setting $K_l = 3$ for the length-specific measurements.

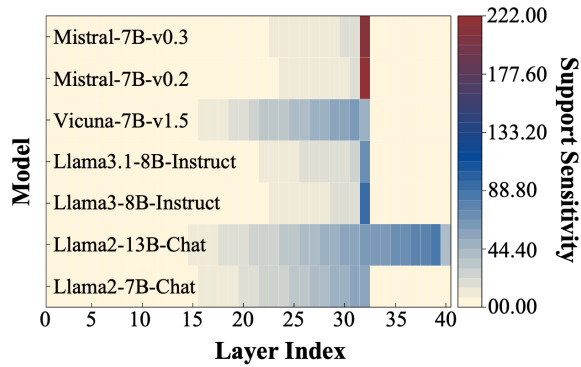


Figure 6: Layer-wise support sensitivity across model backbones. Each row represents a model, each column represents a layer, and the color scale indicates the representation shift under factual-anchor perturbation. Layer indices beyond the depth of a model are omitted.

GPU Memory. As shown in Table 9, HCG does not change memory allocation before the forward pass but increases peak memory during decoding. The additional memory is mainly associated with head-graph construction and the intermediate representations used for cross-head propagation.

Inference Latency. Table 10 shows that the latency overhead increases with generation length because HCG constructs coordination graphs and propagates head representations at each decoding step. Together with the results in Section 5.5, these measurements characterize the inference-time trade-off between factuality improvement and computational cost.