CoDA: Restoring Contextual Dominance via Copy-Encouraged Attention Intervention for Mitigating RAG Hallucinations

Jinwei Shi*, **Qizhuo Xie***, **Qianzi Hou, Zhipeng Wang, Wanting Su, Jianhua Zhao, Tao Zheng, Tieke He***
State Key Laboratory for Novel Software Technology, Nanjing University, China
hetieke@gmail.com

Abstract

Retrieval-augmented generation reduces hallucination by grounding model outputs in external evidence, yet hallucinations can still occur even when the retrieved context is accurate and sufficient. From the perspective of information routing in the residual stream, this reflects an imbalance where internal parametric knowledge overwhelms external context during generation. We present an attention-centric analysis of RAG hallucination under valid evidence, showing that hallucinated and factual tokens diverge in mid-to-late Transformer layers as context-selective attention routing weakens, allowing parametric influence to dominate the residual stream. Motivated by prior studies showing that some attention heads—often referred to as copying heads—exhibit stronger information transport capacity, we aim to extend similar evidence-carrying behavior to a broader set of attention heads. To this end, we introduce CoDA, a lightweight inference-time attention intervention that amplifies evidence-aligned value states, enabling more attention heads to transport reliable external evidence in a copy-encouraged manner. Experiments demonstrate that CoDA improves contextual faithfulness, reduces hallucination, and remains robust under long and noisy contexts with modest and stable inference overhead.

1 Introduction

Large language models perform well on knowledge-intensive tasks but remain vulnerable to hallucination, producing fluent yet factually incorrect content (Xu et al., 2025a; Chuang et al., 2024; Niu et al., 2024a). Retrieval-augmented generation mitigates this issue by grounding outputs in external evidence (Ge et al., 2025; Chen et al., 2025). However, even with accurate retrieval, RAG systems can still contradict the provided context, a phenomenon known as RAG hallucination (Sun et al., 2025; Chen et al., 2024; Matys et al., 2025; Yeh et al., 2025; Tan et al., 2025). This indicates that hallucination arises not only from retrieval errors, but also from how Transformer architectures integrate parametric knowledge with external context during generation (Long et al., 2025; Goyal et al., 2025).

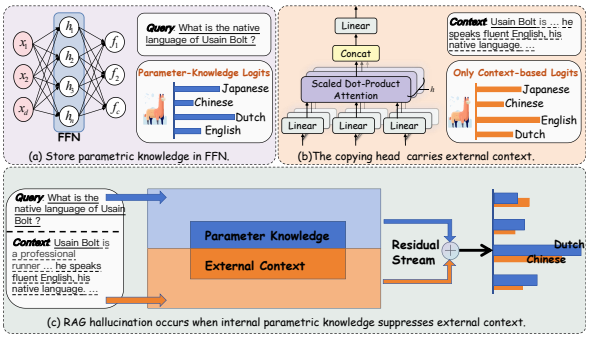

Figure 1: Knowledge contribution to the residual stream via copying heads and FFNs in a one-layer Transformer.

A substantial body of work reduces hallucination by improving retrieval quality or verification, such as re-ranking, adaptive context selection, and external checking pipelines (Asai et al., 2023; Bergeron et al., 2025; Datta et al., 2022; Magesh et al., 2024; Friel and Sanyal, 2023). While effective when evidence is missing or noisy, these approaches do not directly address a more challenging case frequently observed in practice: the retrieved context already contains valid evidence, yet the model under-utilizes it and falls back on parametric associations. For example, SELF-RAG (Asai et al., 2023) and HalluGuard (Bergeron et al., 2025) regulate retrieval timing or consistency, but leave unresolved how internal and external knowledge compete once context enters the network.

Recent mechanistic analyses provide insight into this issue by examining how different components contribute to the residual stream. Re-

*Corresponding author: Tieke He (hetieke@gmail.com).
*Jinwei Shi, Qizhuo Xie contributed equally to this work.

Findings of the Association for Computational Linguistics: ACL 2026, pages 11879–11892
July 2-7, 2026

DeEP (Sun et al., 2025) shows that some attention heads act as copying modules that transport external context, while FFNs inject strong parametric signals (Fig. 1). Complementary studies such as LUMINA (Yeh et al., 2025), InterpDetect (Tan et al., 2025), and Context–Parametric Inversion (Goyal et al., 2025) further indicate that hallucination often emerges when internal activations dominate evidence-aligned representations. However, these analyses are largely descriptive and offer limited guidance on where hallucination-related divergence forms and how to intervene during inference.

Inspired by the locate–edit paradigm in model editing (Santurkar et al., 2021; Xu et al., 2023; Meng et al., 2023a,b; Li et al., 2023b, 2024b), we propose CoDA, an attention-centric locate-and-intervene framework for RAG hallucination. Our key insight is that copying is not a behavior exclusive to a small set of specialized heads, but can be extended to non-copying heads when evidence is semantically aligned. Accordingly, CoDA first localizes hallucination-prone layers where hallucinated and factual tokens begin to diverge. Then it applies a lightweight, plug-in attention modulation that amplifies evidence-aligned value states, promoting copy-encouraged routing before downstream parametric amplification.

Our design complements mechanistic findings on parametric circuits and knowledge neurons (Dai et al., 2022a; Yao et al., 2025) and aligns with fine-grained analyses of harmful generation (Wang et al., 2024b). Extensive experiments show that CoDA consistently improves contextual faithfulness and reduces hallucination under both standard and long/noisy-context settings, while incurring only modest inference overhead.

We summarize our contributions as follows:

1. **Attention-centric view of RAG hallucination.** We reinterpret hallucination under valid evidence as a failure of context-selective attention routing in the residual stream (Sun et al., 2025; Long et al., 2025; Goyal et al., 2025).

2. **Fine-grained localization of hallucination-prone layers.** We develop a layer-wise framework that pinpoints where hallucinated and factual tokens diverge, enabling targeted inference-time intervention.

3. **Plug-in attention modulation for copy-encouraged routing.** We propose a lightweight method that amplifies evidence-aligned value states and promotes copy-encouraged attention routing without modifying model parameters.

2 Mechanistic Insights into Knowledge in Transformer Layers

Mechanistic interpretability studies have revealed how different components of Transformer architectures contribute to knowledge representation and utilization (Geva et al., 2021; Meng et al., 2023b; Wang et al., 2024a; Yao et al., 2025). We focus on decoder-only Transformers (Grattafiori et al., 2024; OpenAI et al., 2024), where the residual stream acts as the primary communication, through which both parametric knowledge and retrieved context are propagated (Geva et al., 2021; Li et al., 2024a; Sui et al., 2025; Bi et al., 2025).

Within this architecture, attention heads and FFNs play complementary roles: attention routes and aggregates contextual information across tokens, while FFNs inject latent associations memorized during pretraining. Their interaction jointly determines how internal and external knowledge influence model predictions.

2.1 Copying Heads

Prior work has identified a subset of attention heads that propagate token representations with minimal transformation, effectively copying information from source tokens to target positions (Dai et al., 2022b; Elhage et al., 2021). These copying heads (Sun et al., 2025) are particularly effective at preserving high-fidelity contextual signals, such as entity mentions or factual spans retrieved from external documents.

As illustrated in Figure 1, in a RAG setting where the context explicitly states that *Usain Bolt speaks fluent English*, copying heads can directly route this evidence to downstream layers, yielding context-consistent logits that favor the correct answer. In contrast, non-copying heads often perform more abstract or associative transformations, which may dilute or override such evidence when internal parametric associations are activated.

2.2 Feed-forward Neural Networks

Feed-forward neural networks, which comprise approximately two-thirds of model parameters, have been shown to function as key–value memories that store and recall factual associations acquired

during pretraining (Mitchell et al., 2022; Dai et al., 2022a; Meng et al., 2023a; Geva et al., 2021, 2022; Sun et al., 2025). These modules exert strong influences on factual predictions, often amplifying parametric knowledge even in the presence of external evidence (Niu et al., 2024b; Goyal et al., 2025).

3 Methodology

This section presents a unified framework for *localizing* and *mitigating* hallucination in retrieval-augmented generation models from an attention-centric perspective. Our approach builds upon the residual-stream decomposition framework introduced in ReDeEP (Sun et al., 2025), which distinguishes internal parametric knowledge from externally retrieved evidence. While prior work primarily attributes hallucination to excessive parametric activation, we reframe the problem as a failure of context-selective attention routing: even when correct and sufficient evidence is retrieved, attention may fail to copy valuable information, and thus fail to adequately prioritize context-aligned representations, allowing downstream parametric amplification to dominate the residual stream.

Based on this view, we develop a layer-wise analysis framework that quantitatively characterizes how contextual evidence and parametric signals interact across depth, identifies hallucination-prone regions, and applies attention-level intervention during inference. Importantly, our framework does not modify model parameters nor explicitly edit feed-forward networks; instead, it reshapes how contextual information is routed and encourages more copy-encouraged behaviors through attention into the residual stream.

3.1 Quantification of Knowledge Contributions

Following prior mechanistic analyses (Sun et al., 2025; Long et al., 2025), we decompose each layer's residual stream into two primary sources influencing a generated token r_i: (1) external contextual evidence propagated via attention mechanisms, quantified by the External Context Score (ES), and (2) internal parametric influence injected through feed-forward transformations, quantified by the Parametric Knowledge Score (PS). This decomposition reflects the complementary roles of attention and FFNs in Transformer models, while enabling fine-grained tracking of their relative dominance during generation.

External Context Score (ES). Let the retrieved context be $C = \{c_1, \ldots, c_m\}$ and the generated response be $R = \{r_1, \ldots, r_n\}$. At layer l, let $A^l = [a^l_{i,j}]$ denote the attention matrix from response tokens to context tokens. For each response token r_i, we define the indices of the top-$k\%$ attended context tokens as:

$$I^l_i = \{\, j \mid a^l_{i,j} \text{ is in the top-}k\% \text{ of } a^l_{i,1:m} \,\}. \quad (1)$$

Let x^L_t denote the hidden state of token t at the final layer L. We then define the external context score as:

$$\mathrm{ES}^l_i = \cos\left(x^L_{r_i}, \frac{1}{|I^l_i|} \sum_{j \in I^l_i} x^L_{c_j} \right), \quad (2)$$

which measures the semantic alignment between the generated token and its most attended contextual evidence. A higher ES^l_i indicates that attention at layer l effectively routes context-consistent information toward the final representation.

Parametric Knowledge Score (PS). To quantify the parametric influence associated with feed-forward transformations at layer l, we follow established mechanistic analyses showing that FFNs function as key–value memories encoding factual associations acquired during pretraining (Dai et al., 2022a; Mitchell et al., 2022; Meng et al., 2023a; Geva et al., 2021, 2022). These parametric activations are known to exert strong influence on factual predictions, and can dominate generation even in the presence of external evidence (Niu et al., 2024b; Goyal et al., 2025).

Let $x^{\mathrm{mid},l}_{r_i}$ denote the residual-stream state of token r_i immediately before the FFN at layer l, and $x^l_{r_i}$ the state after the FFN. We map these states to vocabulary distributions via:

$$q(x) = \mathrm{softmax}(\mathrm{LogitLens}(x)). \quad (3)$$

The parametric knowledge score is defined as:

$$\mathrm{PS}^l_i = \mathrm{JSD}\left(q\left(x^{\mathrm{mid},l}_{r_i}\right) \,\|\, q\left(x^l_{r_i}\right) \right), \quad (4)$$

where $\mathrm{JSD}(\cdot\|\cdot)$ denotes the Jensen–Shannon divergence. A higher PS^l_i reflects a stronger FFN-induced shift in the token distribution, corresponding to increased parametric influence injected into the residual stream.

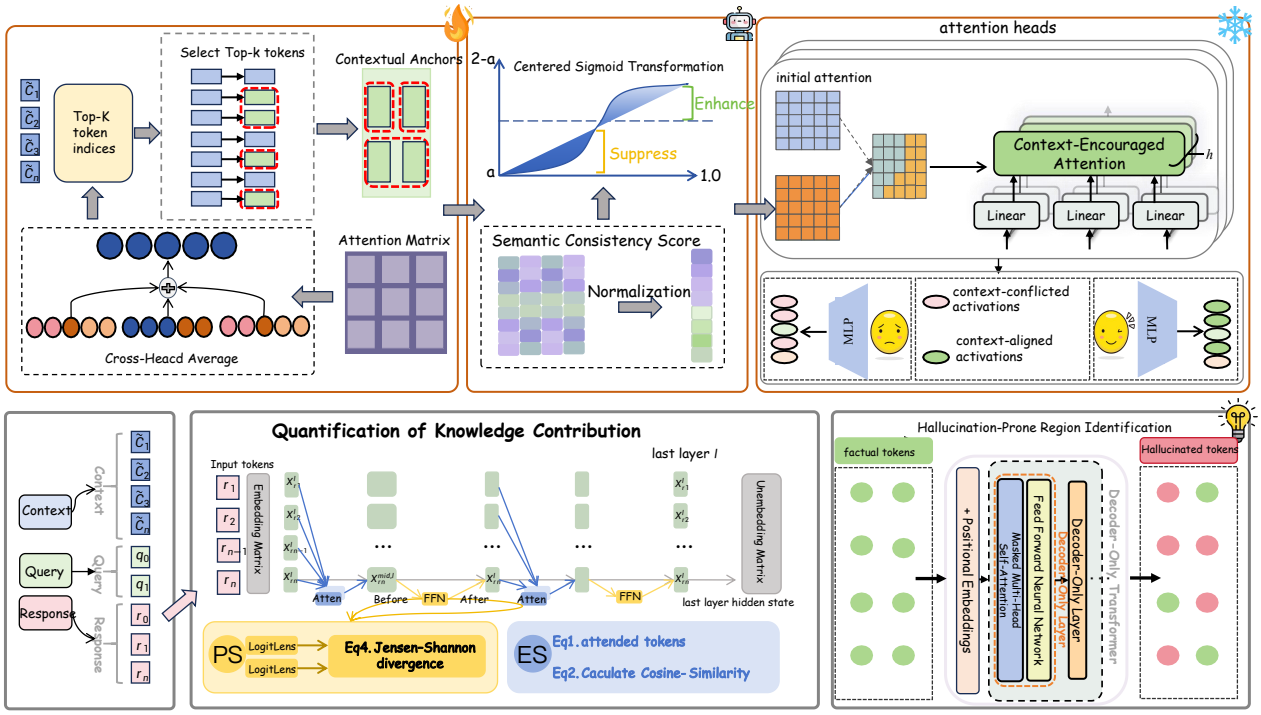

Figure 2: Overview of CoDA. **Bottom (Localization):** CoDA diagnoses where hallucination emerges by contrasting layer-wise behaviors of factual versus hallucinated tokens, and selects a small set of hallucination-prone layers R for intervention. **Top (Intervention):** On layers in R, CoDA performs inference-time attention modulation: it selects contextual anchors, measures value-level semantic consistency, and reweights attention to upweight evidence-consistent key–value pairs, thereby inducing context-encouraged routing without updating model parameters.

Cumulative Ratio. To capture how parametric dominance accumulates across depth, we define the cumulative ratio up to layer l as:

$$\mathrm{CR}_i^l = \frac{\sum_{k=1}^{l} \widetilde{\mathrm{PS}}_i^k}{\sum_{k=1}^{l} \left(\widetilde{\mathrm{PS}}_i^k + \widetilde{\mathrm{ES}}_i^k \right)}, \tag{5}$$

where $\widetilde{\mathrm{PS}}_i^k$ and $\widetilde{\mathrm{ES}}_i^k$ denote normalized scores. This metric reflects the extent to which internal parametric signals dominate the formation of token r_i's representation as generation progresses.

3.2 Identification of Hallucination-Prone Layers

Based on the cumulative ratio CR_i^l, we introduce dataset-level indicators to identify layers where attention-context misalignment and parametric dominance jointly emerge.

Activation Strength. We define the average residual activation at layer l as:

$$\mathcal{A}^l = \mathbb{E}_i \left[\left| \mathrm{CR}_i^l \right| \right], \tag{6}$$

which summarizes the overall strength of internal influence at that layer.

Divergence Strength. To measure representational divergence between hallucinated and factual tokens, we define:

$$\Delta^l = \left| \mathbb{E} \left[\mathrm{CR}_i^l \mid y_i{=}1 \right] - \mathbb{E} \left[\mathrm{CR}_i^l \mid y_i{=}0 \right] \right|, \tag{7}$$

where $y_i{=}1$ denotes hallucinated tokens and $y_i{=}0$ factual tokens.

Hallucination Sensitivity. We combine these indicators to obtain:

$$\mathcal{H}^l = \mathcal{A}^l \times \Delta^l, \tag{8}$$

and extract the hallucination-prone layer set:

$$\mathcal{R} = \text{Top-}k_l \left(\mathcal{H}^l \right), \tag{9}$$

which identifies the subset of layers where attention fails to sufficiently privilege context-aligned representations, resulting in amplified parametric dominance.

3.3 Copy-encouraged Attention Modulation

Having localized hallucination-prone layers, We introduce CoDA, an inference-time, layer-wise attention modulation mechanism that leverages copy-encouraged attention to strengthen attention heads'

ability to transport and route evidence from the retrieved context into the residual stream, applied only to the hallucination-prone layers $l \in \mathcal{R}$. Unlike prior approaches that directly regulate FFN activations via fixed scaling (Sun et al., 2025; Huang et al., 2025), CoDA does not modify FFNs or model parameters. Instead, it dynamically reweights attention to promote copy-encouraged behavior in non-copying heads when contextual evidence is semantically consistent.

Let the input to layer l consist of attention weights $\mathcal{A}$ and hidden states X, where $\mathcal{A} \in \mathbb{R}^{B\times H\times L\times L}$ and $X \in \mathbb{R}^{B\times L\times d}$. We first compute the mean attention map $A \in \mathbb{R}^{L\times L}$ by averaging over heads:

$$\bar{A}_{i,j} = \frac{1}{H}\sum_{h=1}^{H} \mathcal{A}^h_{i,j}. \tag{10}$$

We then estimate token-level attention intensity $s_j = \frac{1}{L}\sum_{i=1}^{L} \bar{A}_{i,j}$ and select the top-K most attended context tokens $\mathcal{T} = \mathrm{TopK}(s)$.

For each head h, we designate the corresponding value vectors $\{v^h_k \mid k \in \mathcal{T}\}$ as contextual anchors. The semantic consistency of each token j is computed as:

$$S^h_j = \frac{1}{K}\sum_{k\in\mathcal{T}} \frac{\langle v^h_j, v^h_k\rangle}{\|v^h_j\|_2 \cdot \|v^h_k\|_2}. \tag{11}$$

After min–max normalization to $\tilde{S}^h_j \in [0,1]$, we derive the adaptive dosing factor:

$$w^h_j = \alpha + (2-2\alpha)\cdot\sigma\left(\tau\cdot(\tilde{S}^h_j - 0.5)\right), \tag{12}$$

where τ controls the sharpness of modulation and α specifies the minimum weight.

Finally, attention weights are modulated as:

$$\tilde{A}^h_{i,j} = A^h_{i,j}\cdot w^h_j, \tag{13}$$

which scales the contribution of each key–value pair according to its contextual relevance. By amplifying context-aligned value states prior to aggregation, CoDA effectively encourages copy-encouraged behavior in attention heads that would otherwise dilute external evidence, thereby restoring contextual dominance in the residual stream.

4 Experiments

4.1 Settings

Datasets. We evaluate our method on two categories of datasets: (1) hallucination benchmarks, including *RAGTruth* (Niu et al., 2024a) and *Dolly* (Hu et al., 2024), which assess factual consistency under retrieved evidence; and (2) contextual faithfulness benchmarks derived from *CoFaithfulQA* (Huang et al., 2025), comprising six subsets—HotpotQA, NewsQA, NQ, SearchQA, SQuAD, and TriviaQA. For each dataset, we report the number of faithful examples used for evaluation, as summarized in Appendix B.

Baselines. We compare CoDA with representative baselines spanning prompt-based, decoding-based, fine-tuning, and alignment-based paradigms, including *AttrPrompt* and *OIPrompt* (Zhou et al., 2023), *COIECD* (Yuan et al., 2024), *SFT* and *KAFT* (Wei et al., 2022; Li et al., 2023a), as well as *C-DPO* (Bi et al., 2024), *DDR* (Li et al., 2025), and *ParamMute* (Huang et al., 2025). All baselines are evaluated using Context Recall (ConR) and Memory Recall (MemR), as defined in Appendix C. In addition, we directly compare CoDA with ReDeEP (Sun et al., 2025) on RAGTruth and Dolly (AC).

4.2 Main Results and Analysis

4.2.1 Hallucination-Prone Region Localization

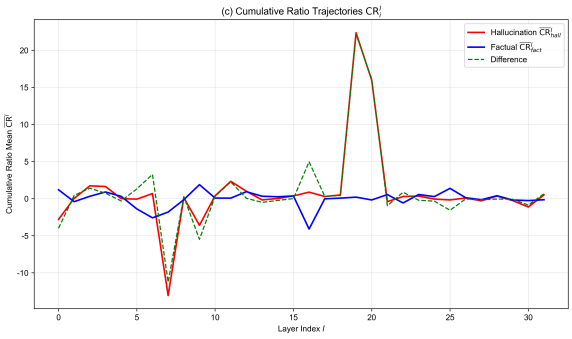

Figure 3: Layer-wise cumulative ratio trajectories CR_i^l (Eq. 5) comparing hallucinated (red) and factual (blue) samples. The green dashed curve shows their difference.

We localize hallucination-prone regions by analyzing where factual and hallucinated tokens begin to diverge across model depth. Figure 4 summarizes two layer-wise indicators. As shown in Figure 4(b), the divergence strength Δ^l stays near zero in early layers and exhibits a sharp rise in the mid-to-late layers, indicating clear separation points where hallucinated and factual tokens start to differ. In contrast, the activation strength $\mathcal{A}^l$ in Figure 4(a) varies substantially across layers, but elevated activation alone does not reliably distin-

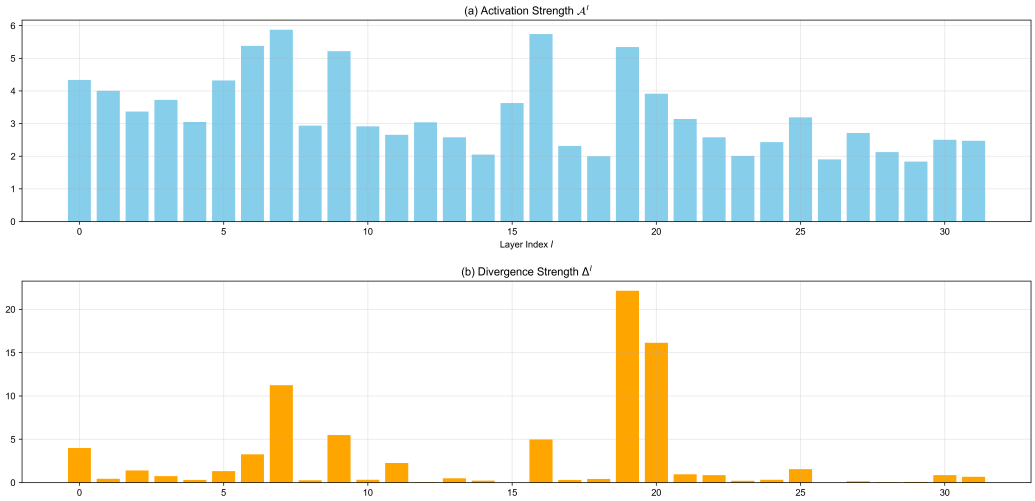

Figure 4: Layer-wise indicators for hallucination-prone region identification: (a) Activation Strength $\mathcal{A}^l$ (Eq. 6) and (b) Divergence Strength Δ^l (Eq. 7) measuring the separation between factual and hallucinated behaviors across layers. High-score layers are treated as hallucination-sensitive candidates.

guish hallucinated from factual behaviors, suggesting that "being active" is not sufficient for hallucination formation.

To provide a more intuitive view of this separation process, Figure 3 visualizes the cumulative ratio trajectories CR_i^l. The red and blue curves largely overlap in early layers, while their gap rapidly expands at the identified mid-to-late layers, matching the peak in Δ^l. Beyond these layers, the difference stabilizes, implying that the representational separation has already been established. Taken together, these observations suggest that hallucination formation is localized to a small subset of mid-to-late layers, where insufficient context-selective attention routing allows parametric influence to dominate the residual stream. Accordingly, we treat these mid-to-late layers as hallucination-sensitive regions and select them as our intervention set $\mathcal{R}$.

4.2.2 Evaluation of Contextual Faithfulness

Across all six CoFaithfulQA subsets, CoDA consistently achieves the highest ConR, indicating stronger grounding in retrieved evidence. As shown in Table 1, CoDA improves performance on HotpotQA, SearchQA, SQuAD, and TriviaQA by 1.5–3.0 points over ParamMute, with comparable gains on NQ and NewsQA. These consistent improvements demonstrate robustness across datasets with diverse domains, context lengths, and reasoning requirements. Importantly, these gains align with the attention-centric design of CoDA. Rather than globally suppressing FFN activations, CoDA dynamically reshapes attention routing by amplifying context-consistent value states and encouraging copy-encouraged behavior in non-copying heads when retrieved evidence is semantically aligned. This mechanism allows contextual signals to be more effectively propagated into the residual stream, which downstream FFNs then amplify in a context-consistent manner. Consequently, CoDA strengthens reliance on faithful external evidence without eliminating useful parametric knowledge. This leads to a more favorable trade-off between evidence grounding and representational flexibility compared to prior suppression-based approaches.

Moreover, Figure 6 illustrates that CoDA consistently maintains low MemR across datasets, confirming that improved contextual grounding does not increase reliance on parametric knowledge .

Models	**HotPotQA**	**NQ**	**NewsQA**	**SearchQA**	**SQuAD**	**TriviaQA**
	ConR ↑	ConR ↑	ConR ↑	ConR ↑	ConR ↑	ConR ↑
Vanilla-RAG	60.34	53.09	60.27	66.76	77.93	61.80
Attr_prompt	58.93	55.36	58.80	62.53	77.35	59.97
OI_prompt	47.79	49.25	52.03	52.26	76.81	55.41
COIECD	62.51	56.21	51.81	69.74	73.12	63.62
SFT	70.92	59.76	61.96	75.29	79.19	59.60
KAFT	69.52	60.89	65.09	77.38	80.04	62.32
C-DPO	67.20	62.24	61.40	64.12	80.08	58.67
DDR	68.66	63.29	64.74	78.07	81.36	60.71
ReDeEP	68.40	60.98	65.12	71.62	81.57	61.15
ParamMute	71.06	60.68	65.24	78.76	80.58	60.89
CoDA (Ours)	**73.09**	**63.09**	**68.62**	**80.92**	**83.06**	**63.10**

Table 1: Contextual faithfulness on CoFaithfulQA. We report ConR (↑) on six subsets under the same RAG setting. CoDA achieves the best ConR across all subsets, benefiting from copy-encouraged attention that strengthens evidence-grounded routing; higher ConR translates to more effective hallucination mitigation.

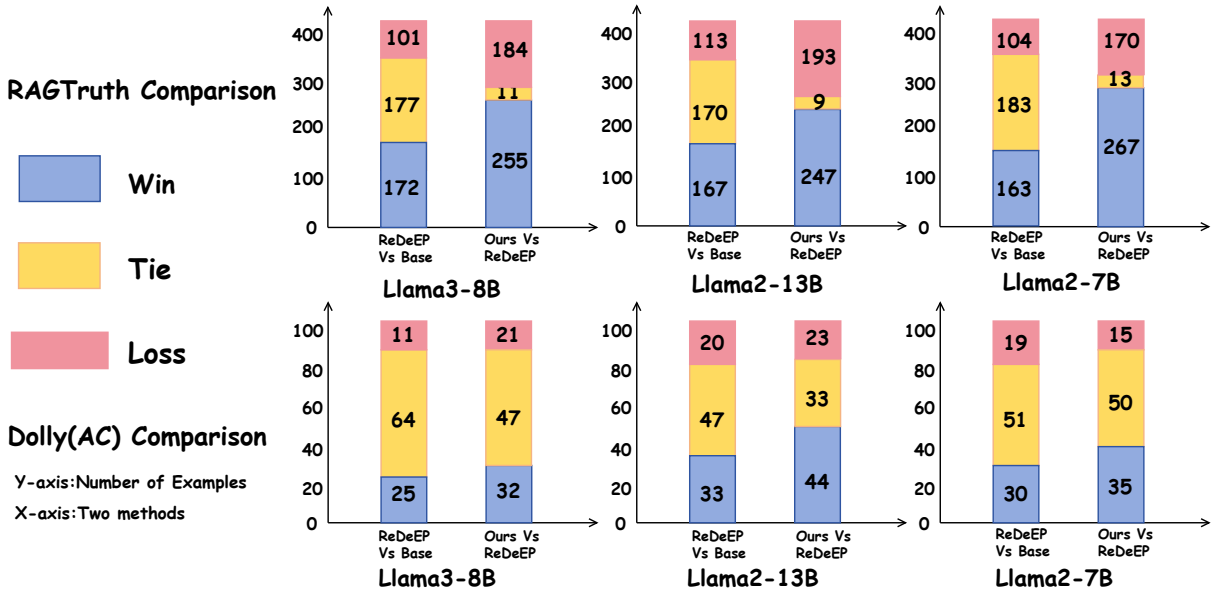

Figure 5: Pairwise hallucination mitigation comparison on RAGTruth (top) and Dolly (AC) (bottom) judged by GPT-4o. Left bar: ReDeEP vs Base; right bar: CoDA vs ReDeEP. Blue/yellow/red denote win/tie/loss for the former method (more faithful = fewer hallucinations); numbers are counts.

4.2.3 Hallucination Mitigation Comparison

Figure 5 presents pairwise comparisons between CoDA and ReDeEP, evaluated by GPT-4o. For each backbone, the left bars compare ReDeEP against the base model, while the right bars compare CoDA against ReDeEP; blue, yellow, and red denote better, comparable, and worse faithfulness (i.e., lower hallucination) for the former method, respectively. On RAGTruth, CoDA surpasses ReDeEP on 57.5%, 56.1%, and 61.1% of examples for Llama3-8B, Llama2-13B, and Llama2-7B, respectively. On Dolly (AC), CoDA's win rates further rise to the 60–70% range, suggesting consistent gains across datasets and model scales. Overall, these results indicate that CoDA delivers more effective hallucination mitigation than ReDeEP.

Crucially, the improvement is not driven by stronger suppression of FFN activations, but by more accurate attention-level routing of contextual evidence. By adaptively promoting copy-encouraged behavior in attention heads when external evidence is relevant, CoDA ensures that context-aligned representations dominate the residual stream before parametric amplification takes effect, thereby reducing the chance of evidence-contradicting generations. Representative comparison cases and detailed evaluation outcomes are provided in Appendix D.

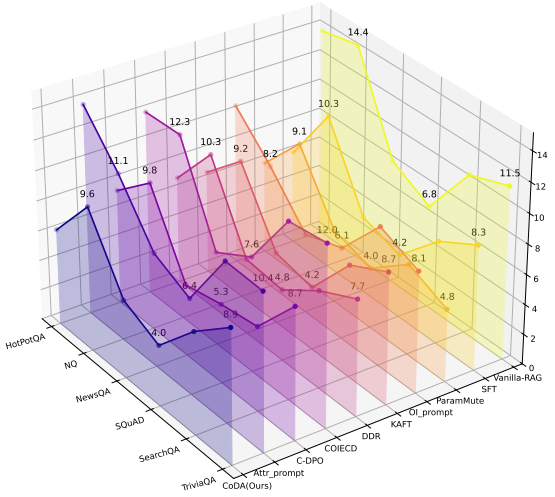

Figure 6: MemR comparison across six CoFaithfulQA subsets. CoDA achieves competitive memory reliance on all datasets, indicating reduced dependence on parametric memory during generation.

5 Efficiency Analysis

5.1 Noise and Length-Robust Efficiency under Valid Context

In realistic RAG settings, retrieved contexts are often long and noisy, even when gold-supporting evidence is present. Under such conditions, hallucination mitigation methods should remain robust to irrelevant context while introducing limited inference overhead. To evaluate CoDA in this

regime, we design a controlled experiment where the retrieved context always contains the gold evidence, while varying only the context length and the amount of injected noise.

Specifically, on CoFaithfulQA (NQ / TriviaQA), we construct clean contexts containing only the evidence span, and noisy contexts by appending unrelated passages sampled from other instances; the total context length is controlled at different scales to assess scalability and noise robustness. We report end-to-end latency together with Context Recall (ConR) and $\Delta\text{ConR}_{\text{drop}}$, defined as the ConR difference between clean and noisy contexts.

As the context becomes longer and noisier, the base model suffers substantially larger ConR degradation, indicating heightened sensitivity to irrelevant information, whereas CoDA consistently exhibits smaller drops across context scales. As shown in Figure 7, CoDA achieves the smallest $\Delta\text{ConR}_{\text{drop}}$ across all six CoFaithfulQA subsets, outperforming both Vanilla-RAG and ReDeEP under noisy contexts.

This robustness stems from CoDA's attention-centric intervention. Rather than operating on the full context or globally modifying activations, CoDA selectively reshapes attention routing using a small set of context-aligned anchors and hallucination-prone layers. By amplifying context-consistent value states, CoDA limits the influence of noisy tokens on downstream representations, remaining stable under long and noisy inputs with modest and stable latency overhead.

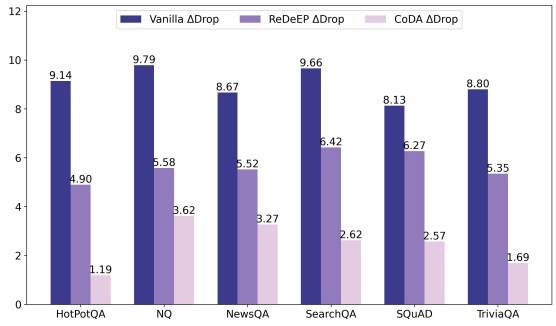

Figure 7: Comparison of ConR degradation under noisy contexts. CoDA consistently exhibits smaller performance drops than Vanilla-RAG and ReDeEP across all six subsets.

Inference Efficiency Evaluation

To evaluate the inference overhead of CoDA, we compare end-to-end single-sample latency with and without CoDA on RAGTruth and Dolly (AC), under identical retrieval results, model configurations, and hardware settings. As shown in Table 2, CoDA introduces a modest and stable latency overhead of approximately 10%–20% across model sizes. This overhead stems from two lightweight inference-time operations: (1) attention-based Top-K context token selection and (2) semantic similarity computation with differential dosing applied to a small set of hallucination-prone layers. Notably, hallucination localization is performed offline and incurs no online cost. Since Transformer inference is dominated by attention and FFN matrix multiplications, these additional operations constitute only a minor fraction of the forward pass. Furthermore, as CoDA depends on a fixed number of anchors and selected layers rather than full context length, its relative overhead remains stable as context length increases.

Dataset	Model	Baseline (s)	CoDA (s)	Δ(s)	Overhead (%)
RAGTruth					
	Llama3-8B	7.82	9.13	1.31	16.74%
	Llama2-13B	10.50	12.08	1.58	15.00%
	Llama2-7B	6.50	7.54	1.04	16.00%
Dolly (AC)					
	Llama3-8B	7.10	8.28	1.18	16.62%
	Llama2-13B	9.70	11.20	1.50	15.46%
	Llama2-7B	5.95	6.93	0.98	16.47%

Table 2: Inference latency with and without CoDA. Latency overhead remains stable (15–17%) across datasets and model sizes.

CoDA achieves substantial improvements in hallucination mitigation and contextual faithfulness while introducing limited and predictable inference overhead. These results demonstrate that attention-level routing interventions can offer a practical and scalable alternative to heavier model-editing or retraining-based approaches in RAG applications.

6 Conclusion

Motivated by the observation that certain attention heads function as copying heads with strong information transport capacity, we ask whether similar copy-encouraged behavior can be induced in other heads to preserve external context better. Based on this insight, we propose CoDA, a lightweight inference-time attention intervention that amplifies evidence-aligned value states and promotes copy-encouraged routing in non-copying heads, restoring context dominance in the residual stream without modifying model parameters or requiring retraining.

Limitations

Our evaluation focuses on Llama-family backbones; extending to a wider range of model architectures is left to future work. CoDA adds lightweight inference-time computation for anchor selection and attention modulation, and further system optimizations may benefit strict latency settings. We adopt simple, practical default choices for anchor selection and dosing strength, and more adaptive variants could improve robustness under diverse retrieval conditions.

Ethical Considerations

CoDA aims to improve evidence use in RAG, but it does not guarantee correctness, especially when retrieved sources are incomplete or misleading; high-stakes use should include standard safeguards (e.g., provenance and human review). Because CoDA is inference-time and training-free, it introduces no new data collection, but responsible deployment still depends on privacy/compliance controls in the retrieval pipeline. Automated evaluation can be imperfect, and targeted human checks can complement automatic metrics.

Acknowledgments

This work is partially supported by National Science and Technology Major Project (2026ZD1611200), and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Loris Bergeron, Ioana Buhnila, Jérôme François, and Radu State. 2025. Halluguard: Evidence-grounded small reasoning models to mitigate hallucinations in retrieval-augmented generation. *Preprint*, arXiv:2510.00880.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2024. Context-dpo: Aligning language models for context-faithfulness. *Preprint*, arXiv:2412.15280.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *Preprint*, arXiv:2503.15888.

Baiyu Chen, Wilson Wongso, Xiaoqian Hu, Yue Tan, and Flora Salim. 2025. Multi-stage verification-centric framework for mitigating hallucination in multi-modal rag. *Preprint*, arXiv:2507.20136.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *Preprint*, arXiv:2402.03744.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. *Preprint*, arXiv:2309.03883.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. *Preprint*, arXiv:2104.08696.

Damai Dai, Wenbin Jiang, Qingxiu Dong, Yajuan Lyu, Qiaoqiao She, and Zhifang Sui. 2022b. Neural knowledge bank for pretrained transformers. *Preprint*, arXiv:2208.00399.

Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, Ricardo Shih, and Zifan Wang. 2022. Exploring conceptual soundness with trulens. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 302–307. PMLR.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *Preprint*, arXiv:2310.18344.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Danying Ge, Jianhua Gao, Yixue Yang, and Weixing Ji. 2025. Ha-rag: Hotness-aware rag acceleration via mixed precision and data placement. *Preprint*, arXiv:2510.20878.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *Preprint*, arXiv:2203.14680.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. 2025. Context-parametric inversion: Why instruction finetuning can worsen context reliance. *Preprint*, arXiv:2410.10796.

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. *Preprint*, arXiv:2407.21783.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *Preprint*, arXiv:2405.14486.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. 2025. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. *Preprint*, arXiv:2502.15543.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Jiahang Li, Taoyu Chen, and Yuanli Wang. 2023b. Trace and edit relation associations in gpt. *Preprint*, arXiv:2401.02976.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024b. Pmet: Precise model editing in a transformer. *Preprint*, arXiv:2308.08742.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, Maosong Sun, and Chenyan Xiong. 2025. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. *Preprint*, arXiv:2410.13509.

Yongchao Long, Xian Wu, Yingying Zhang, Xianbin Wen, Yuxi Zhou, and Shenda Hong. 2025. Copy-paste to mitigate large language model hallucinations. *Preprint*, arXiv:2510.00508.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Preprint*, arXiv:2405.20362.

Piotr Matys, Jan Eliasz, Konrad Kiełczyński, Mikołaj Langner, Teddy Ferdinan, Jan Kocoń, and Przemysław Kazienko. 2025. *AggTruth: Contextual Hallucination Detection Using Aggregated Attention Scores in LLMs*, page 227–243. Springer Nature Switzerland.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. Mass-editing memory in a transformer. *Preprint*, arXiv:2210.07229.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. *Preprint*, arXiv:2110.11309.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024a. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024b. What does the knowledge neuron thesis have to do with knowledge? *Preprint*, arXiv:2405.02421.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *Preprint*, arXiv:2112.01488.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. 2021. Editing a classifier by rewriting its prediction rules. *Preprint*, arXiv:2112.01008.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *Preprint*, arXiv:2104.07567.

Yi Sui, Chaozhuo Li, Chen Zhang, Dawei song, and Qiuchi Li. 2025. Bridging external and parametric knowledge: Mitigating hallucination of llms with shared-private semantic synergy in dual-stream knowledge. *Preprint*, arXiv:2506.06240.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *Preprint*, arXiv:2410.11414.

Likun Tan, Kuan-Wei Huang, Joy Shi, and Kevin Wu. 2025. Interpdetect: Interpretable signals for detecting hallucinations in retrieval-augmented generation. *Preprint*, arXiv:2510.21538.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *Preprint*, arXiv:2307.03987.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024a. Knowledge mechanisms in large language models: A survey and perspective. *Preprint*, arXiv:2407.15017.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024b. Detoxifying large language models via knowledge editing. *Preprint*, arXiv:2403.14472.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 5554–5569. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025a. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025b. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2025. Knowledge circuits in pretrained transformers. *Preprint*, arXiv:2405.17969.

Samuel Yeh, Sharon Li, and Tanwi Mallick. 2025. Lumina: Detecting hallucinations in rag system with context-knowledge signals. *Preprint*, arXiv:2509.21875.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models. *Preprint*, arXiv:2501.13958.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Related Work

A.1 RAG Hallucination

LLM hallucination refers to the phenomenon where large language models generate plausible yet factually incorrect outputs, undermining their reliability in real-world applications (Kaddour et al., 2023; Kalai et al., 2025; Xu et al., 2025b). Retrieval-augmented generation (RAG) alleviates this issue by retrieving up-to-date and relevant external evidence and conditioning generation on the retrieved context (Gao et al., 2024; Fan et al., 2024; Zhang et al., 2025; Guu et al., 2020; Lewis et al., 2021; Santhanam et al., 2022; Varshney et al., 2023; Mishra et al., 2024).

However, even when the retrieved evidence is accurate and contextually appropriate, RAG systems may still produce unsupported or contradictory statements—a phenomenon commonly referred to as *RAG hallucination* (Shuster et al., 2021; Niu et al., 2024a; Magesh et al., 2024; Sun et al., 2025). This observation suggests that hallucination in RAG is not merely a retrieval failure, but can arise from how internal parametric knowledge and external context are integrated during inference.

Following the experimental setting of ReDeEP (Sun et al., 2025), our work focuses on hallucinations that *persist despite sufficient and relevant retrieved evidence*, and aims to understand and mitigate this failure mode from a mechanistic perspective.

B Datasets

B.1 Dataset Overview and Usage Protocol

We evaluate contextual faithfulness and hallucination mitigation using two complementary dataset families. **(i) Hallucination benchmarks** test whether a model can remain faithful under retrieved evidence, where human or automatic annotations explicitly characterize unsupported or contradictory generations. **(ii) Contextual-faithfulness benchmarks** test whether the model grounds generated facts in the provided context when the evidence is present and sufficient.

For clarity and to keep the main paper concise, we summarize here the exact evaluation splits used in our experiments. Table 3 reports the *number of faithful samples* used for evaluation in each dataset or subset. These counts correspond to the effective evaluation pool after applying the dataset's official filters/constraints and our experimental protocol (e.g., keeping only instances with valid retrieved evidence and well-formed annotation fields). In particular, for CoFaithfulQA-derived subsets, we follow the standard subset partitioning and evaluate under the same RAG setting across subsets, so the reported counts are directly comparable across HotpotQA, NewsQA, NQ, SearchQA, SQuAD, and TriviaQA.

Why report "#Faithful Samples"? Our primary objective is to quantify *contextual grounding* when the context is available and informative. Therefore, we adopt the dataset-provided notion of "faithful samples" as the evaluation substrate: each sample provides a retrieved context and a reference for checking whether generated facts are supported by that context. This design aligns with our focus on hallucinations that occur *even under valid evidence*, and complements case-based analysis in Appendix D.

Dataset	#Faithful Samples
Hallucination	
RAGTruth (Niu et al., 2024a)	450
Dolly (AC) (Hu et al., 2024)	100
CoFaithfulQA Subsets	
HotpotQA	1,546
NewsQA	374
NQ	3,010
SearchQA	10,692
SQuAD	2,799
TriviaQA	5,887

Table 3: Statistics of datasets used in our experiments (moved from the main-body table to Appendix B.2 for readability).

B.2 Additional Notes on Evaluation Consistency

To ensure a consistent comparison across methods, we keep the evaluation interface fixed: (1) each method receives the same query and retrieved context (when applicable), (2) metrics are computed on the generated response under the same fact-extraction and support-checking protocol defined by the benchmark, and (3) we aggregate results at the dataset/subset level using the same averaging scheme across all baselines.

Hallucination benchmarks. RAGTruth and Dolly (AC) provide representative contexts where hallucinations manifest as statements that are unsupported by or contradictory to the evidence. These benchmarks are used to measure whether an intervention reduces such misaligned generations and to support the qualitative, case-level comparisons reported later.

Contextual-faithfulness benchmarks. The CoFaithfulQA-derived subsets cover diverse QA sources and context characteristics. Evaluating on multiple subsets helps verify that gains are not confined to a single domain, context length distribution, or question style.

C Evaluation Metrics for Contextual Faithfulness

The following metrics originate from CoFaithfulQA (Zhou et al., 2023; Huang et al., 2025)

and quantify how well a model balances external-evidence grounding and reliance on parametric memory.

Context Recall (ConR, Higher is Better)

Meaning. ConR measures the proportion of generated facts that are explicitly supported by the retrieved context. Higher values indicate stronger grounding in external evidence.

Computation. For each answer, let

$$F_{\text{ctx}} : \text{ context-supported}, \quad F_{\text{all}} : \text{ total}.$$

Then,

$$\text{ConR} = \frac{F_{\text{ctx}}}{F_{\text{all}}}.$$

Memory Recall (MemR, Lower is Better)

Meaning. MemR measures the proportion of generated facts that originate from parametric memory and are not grounded in retrieved evidence.

Computation. To avoid long text overflowing the column, we define

$$F_{\text{mem}} = \text{facts not supported by context but. present in parametric memory}$$

Then,

$$\text{MemR} = \frac{F_{\text{mem}}}{F_{\text{all}}}.$$

Misalignment Rate (MR, Lower is Better)

Meaning. MR quantifies the proportion of generated facts that are either unsupported by or contradictory to the retrieved evidence—i.e., hallucinations or conflicting statements.

Computation. Let

$$F_{\text{mis}} = \text{number of unsupported or contradicted. generated facts}$$

Then,

$$\text{MR} = \frac{F_{\text{mis}}}{F_{\text{all}}}.$$

Summary of the Metrics

- **ConR (Higher)** measures grounding in retrieved context.
- **MemR (Lower)** measures reliance on internal parametric memory.
- **MR (Lower)** measures the fraction of unsupported/contradictory facts.

D Case Evaluation Analysis

We present representative case-level comparisons between **Ours (CoDA)** and the baseline **ReDeEP**. Each example corresponds to a human evaluation entry in the RAGTruth dataset, indexed by `ours_index`. We include the evaluator comments (translated) and our interpretive analysis to highlight common patterns: (1) cases where improvements mainly come from clearer adherence to the instruction "strictly based on the given content", (2) cases where the baseline hallucinates unsupported details, and (3) marginal cases where both outputs are factually equivalent and differences are stylistic.

Case 1: ours_index = 21 (Finland and NATO)

Original Comment (translated). Both answers correctly state that Finland established a formal relationship with NATO in 1994. However, **Ours** adds "Therefore, the answer is 1994.", making the conclusion more explicit and decisive.

Analysis. Both outputs are factually equivalent; the preference is primarily stylistic.

Conclusion. Reasonable but based on stylistic preference.

Case 2: ours_index = 30 (Nickname of Stefano Magaddino)

Original Comment (translated). Both models correctly list the nicknames "Don Stefano" and "The Undertaker." **Ours** explicitly adds "Based on the provided content," reinforcing grounding.

Analysis. Both answers are content-equivalent; **Ours** better matches the instruction constraint (grounding emphasis).

Conclusion. Fair and well-justified judgment.

Case 3: ours_index = 41 (Cardiovascular Risk Factors)

Original Comment (translated). Ours lists only risk factors with explicit numerical contributions (hypertension, smoking, diabetes, physical inactivity, obesity), and notes missing numbers for others. The baseline misinterprets "53% dietary risk" and fabricates proportions.

Analysis. The evaluator correctly identified baseline hallucinations. **Ours** follows the rule "report only numerically quantified factors" and avoids making up unsupported ratios.

Conclusion. Completely correct evaluation.

Prompt Length (tokens)	Baseline (s)	CoDA (s)
100	26.7219	33.4681
259	28.3332	32.9018
341	29.3058	37.4260
1142	33.5325	37.8841
2600	39.6130	44.7178

Table 4: Inference latency under different prompt lengths.

Case 4: ours_index = 52 (Tesla Bot Alias)

Original Comment (translated). Ours uses quotation marks for "Optimus," which the evaluator interprets as more precise. Baseline content is identical but lacks quotes.

Analysis. Both answers are factually identical; quotation marks are stylistic.

Conclusion. Marginal justification; factual equivalence holds.

Case 5: ours_index = 56 (Flex Computer System)

Original Comment (translated). Both correctly describe the architecture. The baseline more explicitly mentions "strong type checking and memory safety," aligning better with the source.

Analysis. Baseline legitimately infers "memory safety guarantees" from "safe implementation." **Ours** emphasizes organization/execution and slightly under-specifies the technical focus.

Conclusion. Evaluation correct; baseline superiority justified.

E Inference Latency

We evaluate the inference latency of CoDA under different prompt lengths. Our method applies attention modulation only to a small set of critical layers and operates on top-K attention elements, which limits the additional computation.

As the prompt length increases, the relative overhead decreases, since the baseline attention cost grows with sequence length while CoDA only processes a sparse subset.