ATTENTION SINKS AS KNOWLEDGE-CONFLICT ARBITERS: TRAINING-FREE GATING-PRESERVING ROUTING FOR CONTEXTUAL FAITHFULNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can generate unsupported answers even when the provided context contains sufficient evidence. Improving contextual faithfulness requires distinguishing answer-determining evidence from distracting information, rather than simply increasing attention to context. Attention sinks complicate this distinction: modifying their weights changes the attention budget available to content, while modifying their values changes the information they contribute. Our interventions show that both masking sink attention and replacing sink values with evidence-derived values tend to increase hallucination rather than improve grounding. These observations motivate a mechanistic account: sink gating is a decision variable the model actively modulates under answer-level conflict, and the sink value it weights carries a suppressed, prior-aligned direction. Building on this account, we propose AGPR, a training-free inference-time intervention that reads the model's own gating deviations as a conflict signal, requires layer-level consensus before acting, and applies a bounded correction along the direction the model has already chosen. Every routing correction is then projected onto a gating-neutral subspace via an exact normalization compensation, preserving the arbitration budget. Across four benchmarks, three model families, and multiple scales, AGPR reduces hallucination and improves grounded success while preserving general capability with negligible inference overhead.

1 INTRODUCTION

Hallucinations undermine the reliability of LLMs, particularly in document-grounded applications where users expect answers to be supported by the provided context. Even when sufficient evidence is available, models may generate claims that contradict or lack support from that evidence (Shi et al., 2024; Bi et al., 2025). These failures reflect not merely knowledge or retrieval deficiencies, but a gap between evidence availability and its effective use during generation. We therefore investigate how to reliably ground model outputs in relevant contextual evidence.

A central challenge in steering attention toward evidence is the presence of *attention sinks*: tokens that receive substantial attention despite limited semantic relevance (Gu et al., 2025). Although this mass appears available for redistribution, sinks also regulate non-sink contributions through attention normalization (Qiu et al., 2026; Fu et al., 2026); reducing sink attention therefore alters both routing and the overall content budget. Existing methods either redistribute sink attention to non-sink tokens (Yu et al., 2024) or enrich the BOS representation (Liu et al., 2026). Both modify sink-mediated gating without preserving the model's arbitration between contextual evidence and parametric knowledge. As shown in Section 2, these

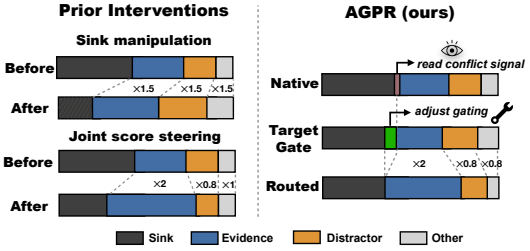

Figure 1: Prior interventions alter sink gating, potentially offsetting routing gains. AGPR instead uses native gating as a conflict signal and exactly preserves its target while reallocating attention only within the non-sink distribution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gating shifts can offset routing gains. Thus, increasing content attention alone does not ensure better discrimination between evidence and distractors.

In a controlled intervention study spanning multiple models, we observe that both masking sink attention and replacing sink Values with evidence-derived Values tend to increase hallucination rather than improve grounding (Appendix B). This counterintuitive result suggests that attention sinks are not passive reservoirs of attention mass, but participate in a functional computation. As shown in Section 2, the model modulates sink gating s_t in response to conflicts between contextual evidence and parametric knowledge, while the associated sink Value encodes a suppressed, prior-aligned direction that gives this modulation semantic effect. Altering either component therefore disrupts the model's native arbitration rather than reclaiming idle attention. This raises a central design question: *how can evidence selection be improved while preserving the model's native sink allocation?*

We address this challenge with **Adaptive Gating-Preserving Routing (**AGPR), a training-free intervention that uses the model's gating response as a conflict signal rather than suppressing it. AGPR detects answer-level conflict from gating deviations that are consistent across heads within a layer, then decouples two operations entangled by the joint softmax: bounded adjustment of the arbitration budget along the model's existing direction and redistribution of content attention within that budget (Figure 1). Each routing correction is projected onto a gating-neutral subspace, exactly preserving the corrected sink allocation and preventing implicit changes to the model's arbitration. Thus, rather than treating attention sinks as surplus mass to redistribute, AGPR preserves their functional role while selectively correcting content routing.

Across multiple knowledge-conflict and hallucination benchmarks, model families, and scales, AGPR consistently reduces hallucination and improves grounded success without compromising instruction-following or reasoning accuracy. It requires only a single forward pass and incurs negligible inference overhead.

The main contributions of this work are summarized as follows:

- **Mechanism.** We identify sink gating as a decision variable actively modulated under answer-level conflict and show that the sink Value encodes a suppressed, prior-aligned direction, revealing a previously unrecognized functional role of attention sinks.
- **Method.** We propose AGPR, which detects conflict from gating deviations, acts only under layer-level consensus, and applies a bounded correction whose routing component is constrained to a gating-neutral subspace through exact normalization compensation.
- **Evaluation.** We evaluate AGPR across knowledge-conflict and hallucination benchmarks, model families, and scales, demonstrating reduced hallucination and improved grounded success without degrading general capabilities, with negligible inference overhead.

2 Coupling of Attention Gating and Content Routing

2.1 The Gating–Routing Decomposition of Attention Sinks

Consider a fixed layer, attention head, and query position t, omitting the layer and head indices. Let $\mathcal{J}_t$ denote the set of visible key positions, $a_{tj} = \mathrm{softmax}_{j\in\mathcal{J}_t}(z_{tj})$, and $\mathbf{v}_j$ the corresponding value vector. Let $S_t \subset \mathcal{J}_t$ be the set of sink tokens, identified by the threshold-based criterion of Yu et al. (2024), which includes the initial token as well as other low-semantic tokens that attract high attention, and $R_t = \mathcal{J}_t \setminus S_t$. For $0 < s_t < 1$, define

$$s_t = \sum_{j\in S_t} a_{tj}, \qquad r_{tj} = \frac{a_{tj}}{1-s_t} \quad (j \in R_t), \qquad \boldsymbol{\mu}_{S,t} = \frac{1}{s_t}\sum_{j\in S_t} a_{tj}\mathbf{v}_j. \tag{1}$$

Writing $\boldsymbol{\mu}_{R,t} = \sum_{j\in R_t} r_{tj}\mathbf{v}_j$, the head output admits an exact decomposition

$$\mathbf{o}_t = s_t\boldsymbol{\mu}_{S,t} + (1-s_t)\boldsymbol{\mu}_{R,t}. \tag{2}$$

This decomposition exposes two control variables: **attention gating**, governed by s_t, sets the balance between sink and non-sink contributions; **content routing**, governed by r_t, determines relative selection among non-sink positions. Sinks have been linked to implicit gating (Fu et al., 2026); our focus is on how score interventions alter this gating indirectly.

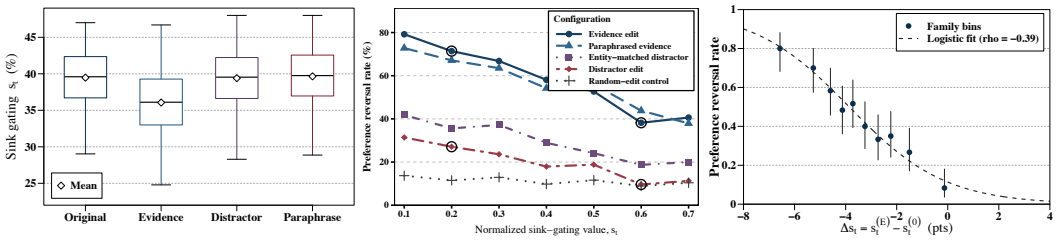

Figure 2: **Left:** s_t under matched context conditions; evidence edits shift s_t lower. **Middle:** Causal interventions on s_t with input fixed. **Right:** Δs_t vs. reversal rate; gating tracks answer-level conflict.

Increased total attention to evidence does not necessarily imply improved relative evidence selection. For a set of evidence positions $E_t \subseteq R_t$,

$$A_{E,t} = \sum_{j \in E_t} a_{tj} = (1 - s_t) \sum_{j \in E_t} r_{tj}. \tag{3}$$

Thus, an increase in $A_{E,t}$ may reflect routing toward evidence or merely an increase in the total non-sink mass $1 - s_t$. Distinguishing these two pathways is essential to explaining why an intervention succeeds or fails.

2.2 Sink Gating as a Knowledge–Conflict Arbiter

Equation 2 identifies s_t as a control variable but does not establish its functional use. To test whether gating responds to prior–context conflict, we construct matched families sharing a question q across four context variants C^v, $v \in \{0, E, D, P\}$: the original context C^0, an answer-changing evidence edit C^E, an answer-preserving distractor edit C^D, and a meaning-preserving paraphrase C^P. Because only C^E changes the context-supported answer, this design isolates answer-level conflict from answer-preserving contextual and surface variation. For each family and variant v, we measure $s_t^{(v)}$ at the final prompt position and average over all heads in layers 12–14, a range identified by a held-out calibration sweep on Llama-3.1-8B as the narrowest contiguous band containing the family-wise peaks in evidence-induced gating deviation; the qualitative pattern remains unchanged over broader bands of layers 10–17 and 8–19 (Appendix C).

Sink gating as an arbiter. The left panel of Figure 2 shows two salient patterns: evidence edits shift s_t downward, whereas distractor and paraphrase edits produce no statistically detectable change relative to the original context. To test whether this modulation is causal, we intervene on s_t while holding the input fixed. We intervene post-softmax by rescaling sink and non-sink weights so that the new sink mass equals a target value while the routing distribution r_t is kept exactly unchanged; the full procedure and a routing-invariance check are in Appendix D. Under evidence edits, restoring s_t to its original-context value reduces the reversal rate from 71.4% to 38.2%; under distractor edits, imposing a matched decrease raises it from 9.6% to 27.1% (middle panel). Together, these directionally consistent interventions demonstrate that s_t causally modulates answer reversal, supporting its role as an arbiter between contextual evidence and parametric recall.

Gating change tracks answer reversal. If s_t mediates knowledge-conflict arbitration, evidence-induced changes in s_t should track shifts in answer preference rather than contextual perturbations per se. For each family, we therefore define the gating decrease as $D_t = s_t^{(0)} - s_t^{(E)} = -\Delta s_t$ and test its association with whether the preference margin $M_i(x) = \ell_\phi(y_i; x) - \ell_\phi(y_i^E; x)$ changes sign between C^0 and C^E. As shown in Figure 2 (right), preference reversals become monotonically more frequent as D_t increases ($\rho = 0.39$, $p < 10^{-3}$ across sites). No corresponding association emerges under distractor or paraphrase controls, which alter token-level statistics while preserving the supported answer, arguing against a purely lexical account.

Interpretation. This pattern suggests that s_t gates the extent to which residual-stream updates favor contextual evidence over the parametric prior in knowledge-conflict QA. Sink gating implements this trade-off: decreasing s_t expands the non-sink budget $1 - s_t$, increasing the head's sensitivity to contextual content, whereas increasing s_t suppresses its contribution and favors the parametric prior. Accordingly, evidence edits signal prior–context conflict and increase the non-sink budget, whereas distractor and paraphrase edits provide no such signal and leave gating unchanged. Because s_t and r_t are coupled through the joint softmax (Section 2.1), any score intervention that alters s_t disrupts this computation rather than introducing a benign perturbation.

2.3 The Inhibitory Geometry of Sink Values

Having established s_t as a decision variable encoding knowledge conflict, we now characterize what it controls. For a reduction in s_t to strengthen contextual influence, the sink component $s_t\boldsymbol{\mu}_{S,t}$ must differ from the contextual-evidence direction in the residual stream. We show that $\boldsymbol{\mu}_{S,t}$ is inhibitory along two dimensions: it has a suppressed magnitude and is directionally aligned with the parametric prior.

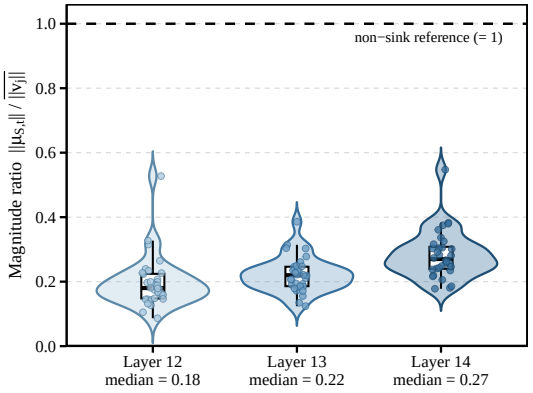

Figure 3: Magnitude ratio $\|\boldsymbol{\mu}_{S,t}\|/\overline{\|\mathbf{v}_j\|}$ for candidate sites in layers 12–14. Violins show distributions, points individual sites, diamonds medians.

Signature I: magnitude suppression. The magnitude $\|\boldsymbol{\mu}_{S,t}\|$ controls how much the sink component contributes to the head output for a given s_t. If this magnitude is small relative to the typical non-sink value norm $\|\mathbf{v}_j\|$, a high s_t does not merely redirect the output toward an alternative direction—it shrinks the head's contribution overall. Figure 3 reports the ratio $\|\boldsymbol{\mu}_{S,t}\|/\overline{\|\mathbf{v}_j\|}$ for candidate sites in layers 12–14. The distribution concentrates well below unity, with medians of 0.18, 0.22, and 0.27 across the three layers. Sink values are therefore not simply another set of content vectors; they are systematically attenuated, so that the sink term $s_t\boldsymbol{\mu}_{S,t}$ remains small even when s_t is large.

Signature II: directional alignment with the parametric prior. Sink values are not merely attenuated in magnitude; they also exhibit a directional bias against contextual evidence. Because $\boldsymbol{\mu}_{S,t}$ lies in head value space, we project it through the head output projection $W_O^{(l,h)}$ before comparison (Appendix D). At each layer, we estimate a prior direction on held-out conflict families as the mean residual-stream difference between prior-following and context-following generations, and compare $\boldsymbol{\mu}_{S,t}$ with this direction and with the mean context-evidence direction $\overline{\mathbf{v}}_{E,t}$ derived from evidence-token Values. Figure 4 plots the resulting per-head cosine similarities. Points lie systematically above the diagonal, indicating stronger alignment with the parametric prior than with contextual evidence, with a median cosine-similarity margin of 0.21 (bootstrap 95% CI $[0.18, 0.24]$). This directional bias is functional: when s_t is large, the head output is pulled toward the parametric prior, suppressing context-driven updates. Sink values therefore carry a competing direction rather than serving as content-agnostic placeholders, providing the physical substrate for the arbitration mechanism.

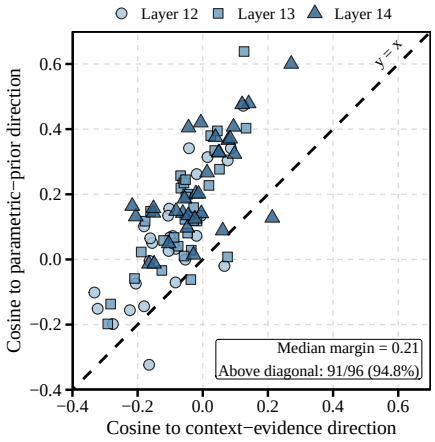

Figure 4: Per-head cosine similarity to the parametric-prior direction versus the context-evidence direction. Points above the dashed diagonal align more with the prior.

Taken together, these signatures indicate that the sink is not content-agnostic; it attenuates the head output while biasing residual contributions toward the parametric prior. Consistent with this role, injecting evidence into sink values increases hallucination, challenging its viability as a means of improving answer accuracy.

3 Method

We propose **Adaptive Gating-Preserving Routing** (AGPR), a training-free inference-time intervention. AGPR detects unusually low sink mass, aggregates these signals within each layer, and

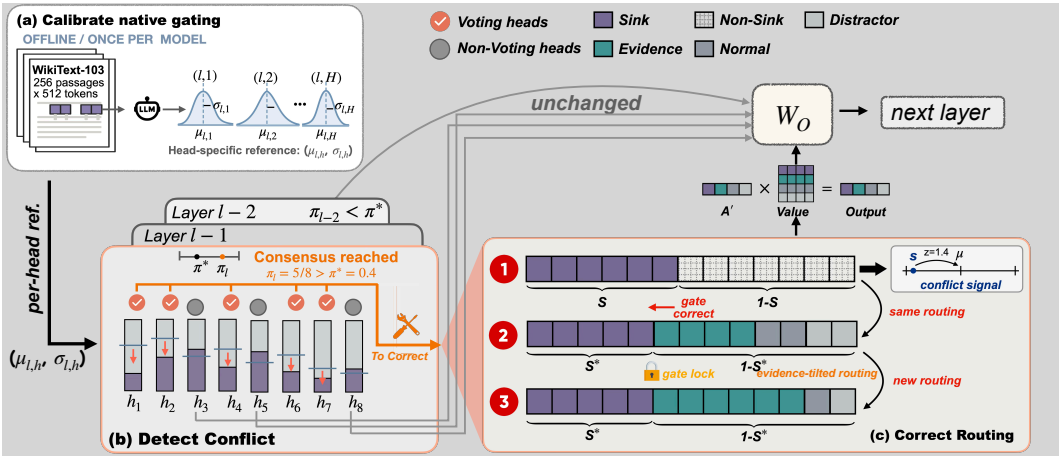

Figure 5: Overview of AGPR. The method detects a conflict signal from the native attention distribution, adjusts and locks the sink gating budget, and then redistributes content attention within this fixed budget. This two-stage design decouples gating from routing.

selectively adjusts gating and routing through an exact normalization correction. The method operates within a single forward pass. Below, we omit the query-position index t for brevity.

3.1 Offline Calibration

Different attention heads exhibit substantially different native levels and variances of sink gating. Therefore, an absolute sink-mass threshold cannot distinguish a conflict-induced decrease from the normal behavior of a head. We perform offline calibration to establish a head-specific reference distribution against which the sink mass observed at inference time can be standardized.

Specifically, we use passages from WikiText-103, a general-domain Wikipedia corpus disjoint from all downstream evaluation benchmarks. We sample 256 passages and truncate each passage to 512 tokens. For every layer l and head h, we collect the native sink mass $s_{l,h}$ over the final 128 valid query positions of each passage and estimate:

$$\mu_{l,h} = \mathbb{E}_{\text{calib}}[s_{l,h}], \qquad \sigma_{l,h} = \sqrt{\mathbb{V}_{\text{calib}}[s_{l,h}]}.$$

For numerical stability, we set $\hat{\sigma}_{l,h} = \max(\sigma_{l,h}, \epsilon_\sigma)$, with $\epsilon_\sigma = 10^{-6}$. The statistics are estimated once during calibration and remain fixed during inference. This parameter-free calibration takes approximately 40 seconds for Llama-3.1-8B-Instruct on a single NVIDIA A100 80GB GPU.

3.2 Conflict Detection

At inference, for each query position t, we compute a normalized deviation from the calibrated expectation:

$$z_{l,h} = \frac{\mu_{l,h} - s_{l,h}}{\hat{\sigma}_{l,h}}. \tag{4}$$

A positive $z_{l,h}$ indicates that the head has lowered its gating below the calibration mean, which Section 2.2 identified as the model's own conflict signal. We convert this signal into a binary head-level vote $v_{l,h} = \mathbb{1}[z_{l,h} > \tau]$, with a default threshold $\tau = 0$.

A single head's vote is noisy; to avoid acting on isolated responses, we compute the within-layer vote fraction, termed *layer-level consensus*:

$$\pi_l = \frac{1}{H}\sum_{h=1}^{H} v_{l,h}, \qquad \text{Consensus}_l = \mathbb{1}[\pi_l > \pi^*], \tag{5}$$

where H is the head count and π^* is the consensus threshold (default 0.4). Only layers satisfying $\text{Consensus}_l = 1$ enter the correction stage. This consensus filter is the operational counterpart of the site-dependence observed in Section 2.2: it retains heads that collectively track a real conflict and discards isolated fluctuations.

3.3 CORRECTION AND COMPENSATION

Bounded arbitration correction. For every layer with $\mathrm{Consensus}_l = 1$, we lower the gating of its voting heads along the direction the model has already chosen:

$$s^*_{l,h} = \begin{cases} s_{l,h} - \lambda\, \hat{\sigma}_{l,h} \max\big(0,\ z_{l,h} - \tau\big), & v_{l,h} = 1, \\ s_{l,h}, & v_{l,h} = 0, \end{cases} \tag{6}$$

with a single strength parameter $\lambda \in [0, 1]$ (default 0.5). Non-voting heads and non-consensus layers are left untouched. Although theoretically possible for very small sink masses, the update never became non-positive in our experiments; we clip it to $[0, 1]$ for safety.

Evidence-tilted routing. Within the non-sink set R_t, we tilt the native routing distribution by an evidence score b_j:

$$\delta_j = \eta\, b_j, \qquad j \in R_t, \tag{7}$$

where $\eta \geq 0$ controls the tilt strength and b_j denotes a token-level evidence score. By default, b_j is the pre-softmax query–key logit used by native attention, so the tilt amplifies the model's existing relevance signal within the corrected budget; alternative scorers are standardized per attention row. Section 4.2 compares this choice with a BM25 retriever score and a lightweight token-level evidence probe. Appendix E gives the exact mapping.

Exact gating compensation. A direct application of δ_j would perturb $s^*_{l,h}$ through the softmax denominator (Section 2.1). We therefore introduce a scalar compensation c and define the corrected non-sink weights as:

$$a'_j = \frac{\exp\big(z_j + \delta_j - c\big)}{Z_S + \exp(-c)\, Z'_R}, \qquad j \in R_t, \tag{8}$$

with $Z_S = \sum_{j \in S_t} e^{z_j}$ and $Z'_R = \sum_{j \in R_t} e^{z_j + \delta_j}$. Imposing $\sum_{j \in S_t} a'_j = s^*_{l,h}$ gives the closed-form solution:

$$\boxed{c = \log Z'_R - \log Z_S + \log \frac{s^*_{l,h}}{1 - s^*_{l,h}}.} \tag{9}$$

Sink logits are left unchanged. The correction satisfies three exact identities: the corrected sink mass equals $s^*_{l,h}$, the sink value statistic $\boldsymbol{\mu}_{S,t}$ is unchanged, and when $s^*_{l,h} = s_{l,h}$ and $\delta_j = 0$ the operator reduces to the identity. Compensation therefore *projects* an arbitrary routing correction onto the gating-neutral subspace, so that routing can be adjusted without disturbing the arbitration budget that Section 2.3 showed to be functionally prior-aligned.

4 EXPERIMENTS

Our evaluation is organized around three questions that mirror the mechanistic claims of Section 2.

- **Q1 (Effectiveness)**: Does AGPR reduce contextual hallucination on natural inputs across model families and scales?
- **Q2 (Attribution)**: Which mechanism-driven components drive the gains—consensus-gated detection, exact gating compensation, or routing tilt alone?
- **Q3 (Transfer and Cost)**: Do the gains transfer across model families and tasks without capability loss, and at what inference overhead relative to the frozen backbone?

Models and evaluation scope. We evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B/14B/32B-Instruct (Qwen Team et al., 2025), and Gemma-2-9B-it (Gemma Team et al., 2024), covering diverse model families and cross-scale comparisons. Evaluation is limited to contextual faithfulness: answers are assessed for consistency with the supplied context, even when they conflict with real-world knowledge, rather than for the context's factual accuracy.

Evaluation datasets. We evaluate contextual faithfulness on ConFiQA (Bi et al., 2025), FaithEval Counterfactual (Ming et al., 2025), HotpotQA (Yang et al., 2018), and RAGTruth (Niu et al., 2024). These benchmarks respectively cover single- and multi-hop knowledge conflicts, counterfactual evidence, distractor-rich multi-hop reasoning, and grounded QA. We further assess robustness

Table 1: **Contextual faithfulness and hallucination evaluation.** CF denotes counterfactual accuracy; EM, exact match; and F1, token-level F1. For RAGTruth-QA, HR measures unsupported or contradictory responses, GSR fully correct and grounded responses, and AR abstentions on answerable inputs. ConflictBank and HaluEval report accuracy.

		ConFiQA	FaithEval		HotpotQA		RAGTruth-QA			ConflictBank	HaluEval
Model	Method	Acc.↑	CF	Acc.↑	EM↑	F1↑	HR↓	GSR↑	AR↓	Acc.↑	Acc.↑
Llama-3.1-8B-it	Native	62.8		69.4	58.3	72.6	25.4	63.1	4.6	54.7	61.2
	CAD	72.6		76.8	59.8	73.1	21.7	66.4	5.3	63.1	68.9
	DAGCD	70.9		75.2	61.2	75.0	19.8	69.0	4.8	61.8	67.5
	Context-DPO	81.5		82.6	60.6	74.4	18.6	70.2	5.8	72.4	76.3
	SinkTrack	74.1		78.3	62.5	75.8	18.9	71.3	**3.7**	66.5	71.8
	AGPR	**85.3**		**86.1**	**63.8**	**77.2**	**16.2**	**74.1**	4.1	**76.8**	**79.4**
Qwen2.5-7B-it	Native	67.2		73.8	60.7	74.5	22.8	67.3	3.8	58.9	64.7
	CAD	75.9		80.4	60.4	74.8	19.5	69.1	5.1	66.8	72.1
	DAGCD	74.6		79.1	63.0	76.4	17.7	72.1	4.0	65.3	70.6
	Context-DPO	84.3		85.6	63.7	77.1	17.2	73.4	4.5	75.9	79.2
	SinkTrack	77.8		80.9	64.2	77.5	16.8	74.1	**3.3**	69.7	74.5
	AGPR	**86.4**		**87.5**	**66.3**	**79.0**	**14.6**	**76.8**	3.7	**79.6**	**82.1**
Gemma2-9B-it	Native	74.6		81.3	65.2	79.8	16.8	74.9	2.9	68.4	73.6
	CAD	80.1		84.7	64.9	79.2	15.1	76.3	3.8	73.8	78.1
	DAGCD	79.4		83.9	66.0	80.4	14.3	78.1	2.6	72.9	77.4
	Context-DPO	**87.6**		87.8	67.1	81.2	13.8	78.6	3.1	79.6	82.7
	SinkTrack	81.3		84.2	67.8	81.7	13.9	78.9	**2.4**	75.8	80.4
	AGPR	87.4		**88.9**	**69.5**	**83.4**	**11.9**	**81.2**	2.7	**82.5**	**84.8**

on ConflictBank (Su et al., 2024) and generalization to dialogue and summarization on HaluEval (Li et al., 2023). Held-out factorial examples diagnose role–fact routing, FaithEval-Unanswerable tests insufficient-evidence behavior, and IFEval (Zhou et al., 2023) and GSM8K (Cobbe et al., 2021) evaluate capability retention. Dataset statistics, preprocessing, and splits are in Appendix F.1.

Baselines and controlled comparisons. We compare AGPR with methods that intervene at complementary stages: CAD (Shi et al., 2024) and DAGCD (Huang et al., 2025) modify decoding distributions, Context-DPO (Bi et al., 2025) aligns toward context-faithful responses, and SinkTrack (Liu et al., 2026) modifies attention or sink-related computation. Internal controls isolate each mechanism-driven component: *Native* (unchanged backbone), *No-Comp.* (routing tilt without gating compensation), *No-Consensus* (per-head trigger instead of layer-level consensus), *Gating-Only* (bounded gating correction without routing tilt), and *Global-Calib* (global instead of head-specific calibration). Implementation and budget-matching details are in Appendix G.

4.1 Contextual Faithfulness Results (Q1)

We evaluate AGPR on knowledge-conflict QA, distractor-rich multi-hop reasoning, and open-ended RAG generation. Table 1 jointly reports task correctness, contextual hallucination, grounded success, and abstention across all evaluated backbones.

Evaluation across model families. AGPR consistently achieves the lowest HR and the highest GSR across all three model families. Relative to the native models, it reduces HR by 29.2%–36.2% and AR by 2.6%–10.9%, while improving GSR by 8.4%–17.4%. Because AR also decreases, the improvement cannot be attributed to increased abstention; instead, it reflects a larger proportion of correct and fully grounded responses. The gains are largest for Llama-3.1-8B, which exhibits the weakest native contextual faithfulness, and smallest for Gemma-2-9B, whose native performance is already the strongest. This pattern suggests diminishing intervention headroom as native contextual faithfulness improves. Consistent trends are observed on ConflictBank and HaluEval, where AGPR yields relative accuracy improvements of 20.6%–40.4% and 15.2%–29.7%, respectively.

Evaluation across model scales. Table 2 shows that native performance improves monotonically across the evaluated model scales: ConFiQA accuracy increases from 67.2% to 78.5%, while RAGTruth HR decreases from 22.8% to 15.2%. Across all three scales, AGPR consistently improves both metrics, increasing ConFiQA accuracy by 19.2, 16.7, and 14.7 percentage points and reducing HR by 8.2, 5.8, and 4.8 percentage points at 7B, 14B, and 32B, respectively. These diminishing absolute gains suggest less intervention headroom as native contextual faithfulness improves, consistent with the cross-family results in Table 1. At 14B, AGPR essentially matches Context-DPO on ConFiQA (90.1% vs. 90.2%) while reducing HR by 1.6 pp (12.6% vs. 14.2%). At 32B, it outperforms Context-DPO on both metrics, indicating that its benefits persist at larger model scales.

Table 2: **Evaluation across Qwen2.5 model scales.**

Scale	Method	ConFiQA Acc.↑	RAGTruth HR↓
7B	Native	67.2	22.8
	Context-DPO	84.3	17.2
	AGPR	**86.4**	**14.6**
14B	Native	73.4	18.4
	Context-DPO	**90.2**	14.2
	AGPR	90.1	**12.6**
32B	Native	78.5	15.2
	Context-DPO	92.8	12.1
	AGPR	**93.2**	**10.4**

4.2 Component Ablation and Sensitivity (Q2)

Component ablation. Table 3 reports results averaged across three backbones. Removing the compensation term causes the largest degradation, increasing HR from 14.2% to 18.6%, consistent with the joint-softmax sink-budget perturbation predicted in Section 2.1. *Gating-Only* produces the second-largest increase in HR (to 17.9%), indicating that gating correction alone is insufficient without evidence-directed routing. *Global-Calib* and *No-Consensus* yield smaller increases (to 16.4% and 16.1%, respectively), supporting head-specific calibration and layer-level consensus. Replacing the default query–key similarity with BM25 or a lightweight evidence probe changes HR by at most 0.8 pp, indicating limited sensitivity to the choice of evidence scorer.

Table 3: **Component-wise ablation (Q2).** Results are averaged across the three backbones.

Variant	ConFiQA Acc.↑	FaithEval CF Acc.↑	RAGTruth HR↓	RAGTruth GSR↑
Native	68.2	74.8	21.7	68.4
No-Comp.	79.2	83.0	18.6	71.4
No-Consensus	84.1	86.3	16.1	74.8
Gating-Only	80.5	84.1	17.9	72.6
Global-Calib.	83.7	86.0	16.4	74.3
BM25	85.1	87.2	15.0	75.8
Probe	85.4	87.4	14.6	76.1
AGPR	**86.4**	**87.5**	**14.2**	**77.4**

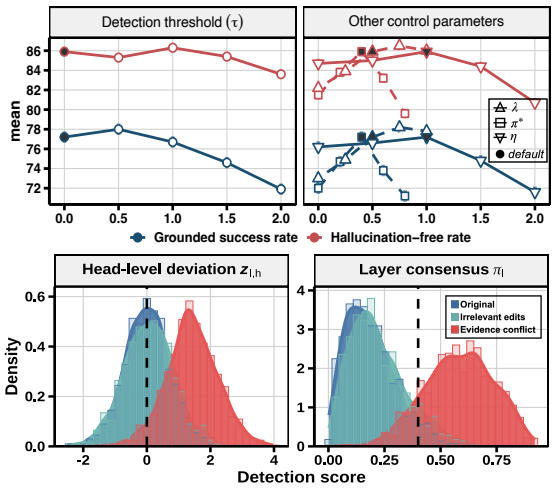

Figure 6: **Top:** Sensitivity to key hyperparameters. **Bottom:** Distributions of $z_{l,h}$ and π_l across original, irrelevant-edit, and evidence-conflict inputs; dashed lines denote intervention thresholds.

Sensitivity and signal separation. Figure 6 (top) reveals that performance remains stable under moderate detection thresholds and routing tilt strengths, but degrades at overly large values. The consensus threshold exhibits a relatively narrow optimum, with performance declining on both sides and more sharply under overly strict consensus. In contrast, the gating correction strength remains effective over a broader range, with both GSR and HFR peaking near $\lambda = 0.75$. The routing tilt remains stable up to approximately $\eta = 1.0$, beyond which excessive perturbation begins to overwhelm the compensation term. Figure 6 (bottom) further shows that both detection signals clearly distinguish *evidence conflict* from *original* and *irrelevant edits*, whose distributions largely overlap. The selected thresholds lie near the boundaries between conflict and non-conflict distributions, retaining high conflict sensitivity while limiting false positives. The heavy right tail of

Table 4: **Capability retention (Q3).** IFEval and GSM8K accuracy across three backbones.

Model	Method	IFEval Acc.↑	GSM8K Acc.↑
Llama-3.1 8B-it	Native	78.4	80.1
	CAD	77.1	78.9
	Context-DPO	76.8	79.4
	AGPR	78.0	80.3
Qwen2.5 7B-it	Native	80.2	85.4
	CAD	79.0	84.1
	Context-DPO	78.6	84.5
	AGPR	80.1	85.5
Gemma2 9B-it	Native	76.5	84.2
	CAD	75.2	82.8
	Context-DPO	74.9	83.3
	AGPR	76.6	84.7

Table 5: **Inference efficiency (Q3).** A100-80GB, batch size 1, and 512-token outputs.

Model	Method	Latency (ms/tok)	Throughput (tok/s)	Memory (GB)
Llama-3.1 8B-it	Native	28.4	35.2	16.8
	Context-DPO	28.7	34.9	17.0
	CAD	54.6	18.3	20.1
	AGPR	29.4	34.0	17.2
Qwen2.5 7B-it	Native	26.1	38.3	15.6
	Context-DPO	26.4	37.9	15.8
	CAD	50.9	19.6	18.7
	AGPR	27.1	36.9	16.1
Gemma2 9B-it	Native	30.2	33.1	18.4
	Context-DPO	30.5	32.8	18.6
	CAD	58.7	17.0	21.4
	AGPR	31.4	31.8	18.9

head-level scores under irrelevant edits indicates occasional spurious head-level activations, further motivating layer-level consensus aggregation. Overall, AGPR exhibits broad hyperparameter robustness and clear conflict-signal separability, with failures emerging primarily under overly aggressive intervention settings, consistent with the mechanism analysis.

4.3 Transfer and Cost (Q3)

We evaluate AGPR along two axes: capability retention on held-out instruction-following and reasoning benchmarks, and inference efficiency relative to the frozen backbone.

Capability retention and inference cost. Table 4 reports IFEval and GSM8K accuracy across the three backbones. AGPR incurs no measurable degradation: the largest drop is 0.4 points on IFEval for Llama-3.1-8B, and all other differences are within ± 0.5 points. In contrast, CAD and Context-DPO reduce both metrics by 1–2 points on average, consistent with the well-known trade-off between aggressive decoding interventions and general capability. Table 5 reports inference cost. AGPR operates within a single forward pass and adds only lightweight per-head statistics and in-attention routing corrections, resulting in 3.5%–4.0% latency overhead across the three backbones, with similarly modest changes in throughput and peak memory. No auxiliary decoding branch, second forward pass, or KV-cache rollback is required, in contrast to CAD, whose context-free branch roughly doubles latency. Although Context-DPO incurs slightly lower inference overhead, it requires backbone-specific post-training, whereas AGPR is training-free. Overall, AGPR provides a favorable trade-off between inference efficiency and deployment cost.

5 Conclusion and Limitations

We have shown that attention sinks are not passive normalization artifacts but functional knowledge–conflict arbiters: the model actively modulates sink gating in response to answer-level conflict, and the sink value carries a suppressed, prior-aligned direction that gives this modulation semantic consequence. Building on this mechanism, AGPR detects conflict through the model's own gating deviations, aggregates them via layer-level consensus, and applies a bounded arbitration correction with an exact compensation that preserves the corrected sink budget during evidence-tilted routing. The method is training-free, operates in a single forward pass, and improves contextual faithfulness across model families and scales while preserving general capability at minimal inference overhead. Our study also has limitations. AGPR assumes that the backbone exhibits measurable sink gating and that conflict signals are distributed across heads; architectures with weak or highly concentrated sinks may require different calibration or detection strategies. The calibration corpus and consensus threshold are chosen once and fixed, which may not transfer to domains far from general text. Finally, we evaluate only on knowledge-conflict and RAG hallucination benchmarks; whether the same arbitration mechanism governs other hallucination types remains an open question.

AI USE STATEMENT

In this work, we used generative AI tools for the following recommended-disclosure tasks: editing the manuscript to improve readability (language polishing, grammar correction, and wording refinement), discovering research topics and identifying gaps, sourcing and searching for information (e.g., literature search assistance), formatting references, and suggesting structural improvements for the paper. We have not used generative AI tools for any required-disclosure tasks, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, or interpreting results.

We have reviewed all AI-assisted work. Specifically, all AI-generated or AI-polished text was manually reviewed and revised by the authors to ensure factual accuracy, technical correctness, and alignment with our original intentions. Any references or literature search results suggested by AI tools were independently verified through manual inspection of the original publications. The authors made all substantive decisions regarding the research content, experimental design, analysis, and conclusions. No AI tool was used to formulate or refine the paper's hypotheses, mechanistic account, mathematical claims, or experimental methodology.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, animal experimentation, or the collection of sensitive personal data. All datasets used in this study are publicly available benchmarks that have been widely adopted in the research community and are used under their respective licenses. The proposed method does not introduce foreseeable risks of misuse, privacy violations, or discriminatory outcomes. We are committed to ensuring fairness, transparency, and reproducibility in our experimentation and reporting.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our results. All datasets, model checkpoints, and evaluation protocols are described in detail in Section 4 and Appendix F.1. Specifically, Appendix F.1 provides dataset statistics, preprocessing steps, and split controls for ConFiQA, FaithEval, HotpotQA, RAGTruth, ConflictBank, HaluEval, IFEval, and GSM8K. The calibration procedure for AGPR, including the choice of WikiText-103 passages, tokenization, and the estimation of per-head statistics, is fully specified in Section 3.1. Hyperparameter defaults and their sensitivity are reported in Section 4.2 and Appendix C. Baseline implementations and budget-matching details are provided in Appendix G. We use three predefined random seeds, $\{42, 43, 44\}$, for training, calibration, and data sampling, as applicable. Given an input and the calibrated statistics, the inference-time intervention is deterministic. We report results on a single NVIDIA A100-80GB GPU with batch size 1 and maximum output length 512, as described in Table 5.

REFERENCES

Bar Alon, Itamar Zimerman, and Lior Wolf. Faithful serum: Mitigating the faithfulness gap in textual explanations of LLM decisions via attribution guidance. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6622–6645, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.300. URL https://aclanthology.org/2026.acl-long.300/.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token?, 2025. URL https://arxiv.org/abs/2504.02732.

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. Context-DPO: Aligning Language Models for Context-Faithfulness. In *Findings of the*

Association for Computational Linguistics: ACL 2025, pp. 10280–10300, 2025. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models, 2026. URL https://arxiv.org/abs/2604.10697.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. Attention sink forges native moe in attention layers: Sink-aware training to address head collapse, 2026. URL https://arxiv.org/abs/2602.01203.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view, 2025. URL https://arxiv.org/abs/2410.10781.

Yanwen Huang, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. Dynamic Attention-Guided Context Decoding for Mitigating Context Faithfulness Hallucinations in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5174–5193, 2025. doi: 10.18653/v1/2025.findings-acl.269. URL https://aclanthology.org/2025.findings-acl.269/.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. URL https://arxiv.org/abs/2305.11747.

Weixian Waylon Li, Yuchen Niu, Yongxin Yang, Keshuang Li, Tiejun Ma, and Shay B. Cohen. Spectral attention steering for prompt highlighting, 2026. URL https://arxiv.org/abs/2603.01281.

Xu Liu, Guikun Chen, and Wenguan Wang. SinkTrack: Attention Sink based Context Anchoring for Large Language Models. *arXiv preprint arXiv:2604.10027*, 2026. URL https://arxiv.org/abs/2604.10027.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows", 2025. URL https://arxiv.org/abs/2410.03727.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. *arXiv preprint arXiv:2505.06708*, 2025. URL https://arxiv.org/abs/2505.06708.

Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, Yang Xu, Haoran Lian, Siqi Zhang, Rui Men, Jianwei Zhang, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training, 2026. URL https://arxiv.org/abs/2601.22966.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=OiC78C68sJ.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen tau Yih. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm, 2024. URL https://arxiv.org/abs/2408.12076.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models, 2024. URL https://arxiv.org/abs/2402.17762.

Chenxu Wang, Chaozhuo Li, Pengbo Wang, Litian Zhang, Songyang Liu, Ji Qi, Jiahui Hu, Yushan Cai, Hao Zhao, and Rui Pu. Lancet: Neural intervention via structural entropy for mitigating faithfulness hallucinations in llms, 2026. URL https://arxiv.org/abs/2601.01401.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57659–57677. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24l.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Related Work

Attention sinks and gating. Attention sinks have been studied as functional components of attention computation, from stabilizing streaming inference (Xiao et al., 2024) to massive activations (Sun et al., 2024), training dynamics (Gu et al., 2025), and the prevention of excessive token mixing (Barbero et al., 2025). More recent work interprets sinks as normalization-induced rescaling factors (Qiu et al., 2026) and implicit gates controlling head contributions (Fu et al., 2026), while

Gated Attention (Qiu et al., 2025) modulates attention outputs with query-dependent gates. A complementary line studies the geometry of sink representations (Ruscio et al., 2025) and uses sink signals for hallucination detection (Binkowski et al., 2026). These accounts describe what sinks are or how they can be read, but not what they do under knowledge conflict. Our analysis shows that sink gating is actively modulated by answer-level conflict and that the sink value carries a prior-aligned direction, motivating an intervention that preserves the native gating budget.

Context-faithful generation and internal interventions. CAD (Shi et al., 2024) contrasts predictions with and without context, DAGCD (Huang et al., 2025) adjusts token probabilities using a context-utilization detector, and Context-DPO (Bi et al., 2025) aligns the model toward context-supported responses. Beyond decoding and alignment, internal interventions aim to correct hallucination-prone computations directly. Faithfulness Serum (Alon et al., 2026) uses attribution guidance to improve the faithfulness of textual explanations, and LANCET (Wang et al., 2026) intervenes on internal pathways via structural entropy to mitigate faithfulness hallucinations. These methods target different failure modes—explanation faithfulness and entropy-driven structural hallucinations—whereas AGPR addresses contextual hallucinations by correcting attention routing at inference time without training or auxiliary decoding.

Training-free attention interventions. ACT (Yu et al., 2024) redistributes sink attention to non-sink tokens, SinkTrack (Liu et al., 2026) enriches the BOS representation, and AdaSEKA (Li et al., 2026) adjusts attention using highlighted evidence spans. All of them alter the sink budget or sink content without preserving the model's native gating allocation. Our mechanism analysis shows that such changes can be offset by induced gating drift under the joint softmax. AGPR instead reads the model's own gating response as a conflict signal and projects every routing correction onto a gating-neutral subspace, preserving the functional role of sinks while correcting content routing.

B Diagnostic Interventions on Attention Sinks

To test whether attention sinks participate in a functional computation rather than merely absorbing unused attention mass, we perform two targeted interventions on the candidate layers (layers 12–14) across all evaluated models and the four core contextual-faithfulness datasets: ConFiQA, FaithEval-Counterfactual, HotpotQA, and RAGTruth. The first, *sink removal*, sets the attention weights of sink tokens to zero and renormalizes the remaining non-sink weights, thereby increasing the total content budget. The second, *sink filling*, replaces the sink value statistics $\boldsymbol{\mu}_{S,t}$ with evidence-derived values, i.e., the mean value of evidence tokens, while keeping the attention weights unchanged. We compare both interventions against two randomized controls: *random removal*, which removes the same number of randomly selected non-sink tokens (mass-matched), and *random filling*, which replaces the sink value with a content-matched random value drawn from non-evidence tokens.

Figure 7 reports the hallucination rate (HR) before (horizontal axis) and after (vertical axis) each intervention. Points above the dashed diagonal indicate an increase in hallucination. Across all tested model–dataset pairs, sink removal increases HR in 75% of cases, and sink filling increases it in 68.8% of cases. In contrast, the randomized controls remain close to the diagonal, with no systematic increase. This asymmetry indicates that the effect is specific to altering the sink component, not to any perturbation of attention mass or value statistics.

These results provide direct empirical support for the mechanism analysis in Section 2: altering either the gating budget (removal) or the value direction (filling) of attention sinks disrupts the model's native arbitration between contextual evidence and parametric knowledge, leading to more hallucination rather than better grounding. The randomized controls confirm that the effect is not explained by a generic reduction of attention mass or by value-space perturbations alone.

C Layer Selection for Mechanism Analysis

Protocol. The mechanism analysis in Section 2 focuses on layers 12–14 and all attention heads in those layers. This range is the output of a held-out calibration sweep. Two distinct data sources are used. The per-head statistics $(\mu_{l,h}, \sigma_{l,h})$ are estimated on 256 passages from WikiText-103, a general-domain corpus disjoint from all evaluation benchmarks. The layer-selection sweep, which computes $D_\ell = \mathbb{E}[|z_{\ell,h}|]$, is performed on a separate set of matched evidence-edit families con-

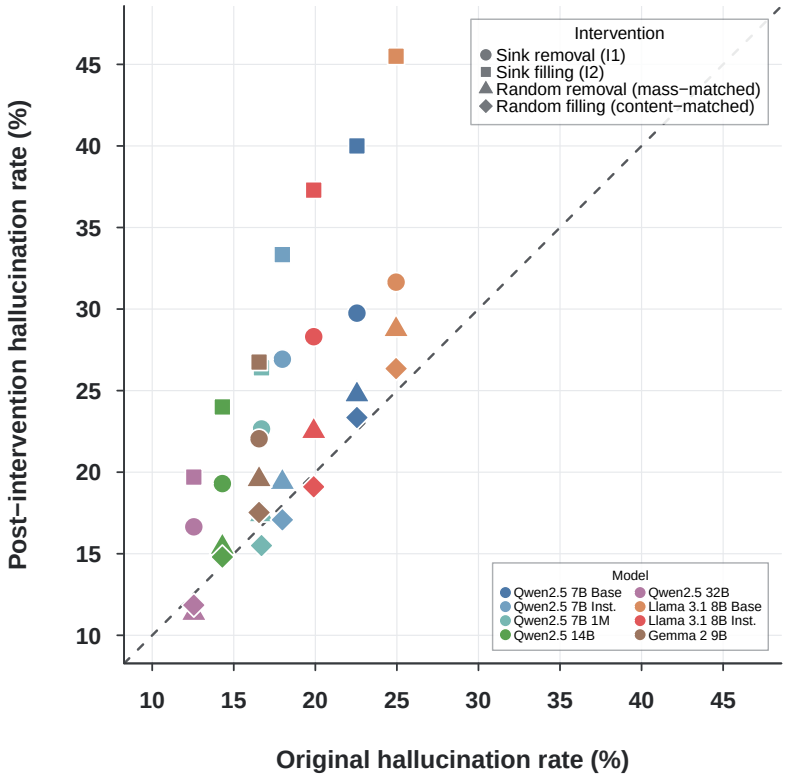

Figure 7: Diagnostic interventions on attention sinks. Each marker represents one model under one intervention, with both the native hallucination rate (horizontal axis) and the post-intervention hallucination rate (vertical axis) averaged over the four datasets (ConFiQA, FaithEval-Counterfactual, HotpotQA, and RAGTruth). The percentages reported in the main text are computed over the 32 underlying model–dataset pairs for each intervention. The dashed line denotes $y = x$. Sink removal and sink filling move most model-level averages above the diagonal, whereas the randomized controls remain close to it.

structed from the training portion of ConFiQA. This set is disjoint from all evaluation data and from the families used in Figure 2. For each family and each context variant C^E, we compute $z_{\ell,h}$ and average its absolute value over heads and families.

$$D_\ell = \mathbb{E}\big[\,|z_{\ell,h}|\,\big], \qquad z_{\ell,h} = \frac{\mu_{\ell,h} - s_{\ell,h}}{\hat{\sigma}_{\ell,h}},$$

where the expectation is taken over all heads and all matched families under the evidence-edit condition C^E. The statistics $\mu_{\ell,h}$ and $\hat{\sigma}_{\ell,h}$ are the same calibration statistics used by AGPR and are estimated on the same conflict-free corpus.

Results. Table 6 reports D_ℓ for layers 8–18. In both backbones the deviation peaks at layer 13 and decreases monotonically toward the shallow and deep ends. Layers 12–14 form the smallest contiguous range that covers the peak and its immediate neighbors on both models, and we adopt this range for all mechanism experiments. The same range is retained for Gemma2-9B in the main results to keep the intervention identical across backbones; because Gemma2-9B has 42 layers, this range corresponds to an earlier relative depth. A full sweep on Gemma2-9B is left to future work.

Robustness to the selected range. To confirm that the mechanism conclusions do not depend on the exact boundary of the candidate range, we repeat the main analysis of Section 2.2 with two expanded ranges, layers 10–17 and layers 8–19. In both cases the qualitative pattern is unchanged: the evidence edit shifts s_t toward smaller values, distractor and paraphrase edits leave s_t statistically

Table 6: Mean absolute gating deviation D_ℓ under evidence edits, for layers 8–18. The peak occurs at layer 13 on both backbones, and layers 12–14 form the smallest contiguous range covering the peak and its immediate neighbors. Values are averaged over all heads and all matched families.

Layer	Llama-3.1-8B	Qwen2.5-7B
8	0.42	0.38
9	0.51	0.46
10	0.63	0.58
11	0.78	0.71
12	0.89	0.82
13	0.94	0.87
14	0.91	0.84
15	0.85	0.79
16	0.76	0.71
17	0.65	0.60
18	0.54	0.50

indistinguishable from the original, and the causal interventions on s_t produce the same directional effects on the reversal rate. The effect size is slightly attenuated when the range is expanded, consistent with the sweep showing that the conflict signal is concentrated in the middle layers. We therefore report the tighter 12–14 range in the main text for clarity, while noting that the mechanism is not sensitive to this choice.

D Causal Intervention and Directional Projection

Intervention algorithm. Given a native attention row (s_t, r_t), we intervene post-softmax by rescaling the sink and non-sink weights so that the new sink mass equals a target s^{target} while r_t is preserved:

$$a'_j = \frac{s^{\text{target}}}{s_t} a_j \ (j \in S_t), \qquad a'_j = \frac{1 - s^{\text{target}}}{1 - s_t} a_j \ (j \in R_t).$$

Under evidence edits, $s^{\text{target}} = s_t^{(0)}$; under distractor edits, $s^{\text{target}} = s_t^{(E)}$. No sink logit, non-sink logit, or value vector is modified. This is the only operation that changes s_t without altering r_t.

Routing invariance. At every modified site we compute the total variation distance between the native r_t and the rescaled routing. Across all sites and evaluation inputs, the distance is below 10^{-6}, confirming that only s_t is changed.

Projection. $\boldsymbol{\mu}_{S,t}$ has dimension d_{head}, while the prior direction lives in the residual stream of dimension d_{model}. We project $\boldsymbol{\mu}_{S,t}$ through the head output projection $W_O^{(l,h)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$, giving $\tilde{\boldsymbol{\mu}}_{S,t} = W_O^{(l,h)} \boldsymbol{\mu}_{S,t}$. The same projection is applied to the mean context-evidence direction $\overline{\mathbf{v}}_{E,t}$.

Prior direction estimation. The prior direction is estimated on a calibration set disjoint from the conflict families used for Figure 4. For each family, we generate one prior-following and one context-following continuation, and record the residual-stream state at the first generated token. The prior direction is the mean difference between the two groups, averaged over families. The same procedure is repeated per layer.

E Evidence Scorer Mapping

This appendix specifies how the token-level evidence score b_j is computed for each scorer used in Section 3.3. All scorers are mapped to the same scale before the tilt $\delta_j = \eta b_j$ is applied.

Default: native query–key logit. For every non-sink position $j \in R_t$, we set

$$b_j = z_j = \frac{\mathbf{q}_t^\top \mathbf{k}_j}{\sqrt{d_h}},$$

the same pre-softmax logit used by the native attention, excluding any positional bias term. This choice requires no extra computation and sharpens the model's existing relevance signal within the corrected budget. Because b_j is already defined on the logit scale, no further normalization is applied.

BM25. We compute a BM25 score between the question q and the sentence containing each non-sink token j. All tokens in the same sentence receive the same raw score. Within each attention row, the raw scores are standardized to zero mean and unit variance over R_t:

$$\tilde{b}_j = \frac{\mathrm{BM25}_j - \mu_{\mathrm{BM25}}}{\sigma_{\mathrm{BM25}}},$$

where μ_{BM25} and σ_{BM25} are the mean and standard deviation of the raw BM25 scores over R_t. The standardized score $\tilde{b}_j$ is then used in place of b_j. This per-row standardization makes the tilt strength comparable across heads and layers.

Learned probe. We train a lightweight logistic regression probe on the final-layer hidden states of non-sink tokens to predict whether a token belongs to an evidence span. The probe is trained on the same held-out conflict families used for calibration, with evidence spans derived from the answer-changing context C^E. At inference, the probe outputs a probability $p_j \in [0, 1]$ for each non-sink token. As with BM25, we standardize these probabilities within each attention row:

$$\tilde{b}_j = \frac{p_j - \mu_p}{\sigma_p},$$

where μ_p and σ_p are the mean and standard deviation of $\{p_j\}_{j \in R_t}$. The standardized value replaces b_j.

Implementation. For all scorers, the tilt is applied only to non-sink positions. The compensation term c in Eq. 9 is recomputed after substituting the chosen b_j, so the target sink mass $s^*_{l,h}$ is preserved exactly regardless of the scorer. The default scorer requires no additional forward pass. BM25 and the probe are computed once per input and reused across all layers and heads. Code for all three mappings is available in the supplementary material.

F Experiment Details

F.1 Dataset Statistics and Preprocessing

ConFiQA. ConFiQA (Bi et al., 2025) contains 18,000 examples, split evenly across its QA, MR, and MC subsets (6,000 each). We use the official test split for evaluation. Each example provides a question, a supporting context, and a context that conflicts with the model's parametric knowledge. No preprocessing beyond the dataset's released tokenization is applied.

FaithEval. FaithEval-Counterfactual and FaithEval-Unanswerable (Ming et al., 2025) contain 1,000 and 2,492 examples, respectively. We use the full sets as released. For FaithEval-Counterfactual, each example provides a context that contradicts world knowledge; for FaithEval-Unanswerable, the context is insufficient to answer the question. We do not filter or resample either set.

HotpotQA. We use all 7,405 examples from the HotpotQA distractor development set (Yang et al., 2018). Every method receives the same ten candidate paragraphs per question, with supporting-fact labels withheld. This ensures that all methods face identical distractor conditions and cannot use gold evidence annotations at inference time.

RAGTruth. We use the official test split of RAGTruth (Niu et al., 2024). We deduplicate responses by `source_id`, retain only answerable QA sources that do not appear among the training responses, and generate one new response for each retained source. Answerability is determined before generation by the dataset's original annotation, and annotations attached to the original responses are not transferred to the regenerated outputs. This procedure yields $n = 720$ evaluation examples. All methods use identical prompts, decoding parameters, and stopping criteria for regeneration.

ConflictBank. We use the ConflictBank benchmark (Su et al., 2024). We evaluate on the QA subset and use the official test split, which contains $n = 10{,}000$ QA examples spanning misinformation, temporal discrepancies, and semantic divergences. No additional filtering is applied.

HaluEval. We use the HaluEval benchmark (Li et al., 2023), which contains approximately $30{,}000$ samples. We evaluate on all three subsets (question answering, dialogue, and summarization) using the official splits. Each subset contains $10{,}000$ test examples. This benchmark tests whether the method generalizes beyond question answering to dialogue and summarization tasks.

Role–fact routing diagnostic. The role–fact routing diagnostic is constructed from the training portion of ConFiQA. Each held-out factorial example contains a shared question with two queried roles and two counterbalanced fact assignments, yielding $n = 720$ diagnostic examples. These examples are disjoint from all evaluation data and from the matched families used in Figure 2. They are used only for the mechanism analysis in Section 2, not for method evaluation.

Capability retention. Instruction-following capability is evaluated on the released IFEval set (Zhou et al., 2023) (541 examples) and mathematical reasoning on the GSM8K test set (Cobbe et al., 2021) ($1{,}319$ examples). Both sets are used in their released form without modification.

Split controls. All examples derived from the same source document are assigned to a single split. No final evaluation data are used for training, calibration, or model selection. The calibration corpus for AGPR is drawn from WikiText-103 and is disjoint from all evaluation benchmarks. The layer-selection sweep is performed on a separate held-out subset of ConFiQA training families, as described in Appendix C. All methods use the same evaluation examples, prompts, and decoding configuration for a given backbone.

G Baseline Implementations and Comparison Protocol

This appendix describes the external baselines and internal controls compared in Section 4.1 and Section 4.2, and specifies the protocol used to ensure fair comparison across methods.

G.1 External Baselines

CAD. Context-aware Decoding (CAD) (Shi et al., 2024) contrasts the output distribution obtained with the supplied context against the distribution obtained without context, amplifying the difference between the two. We use the authors' original implementation[1] and follow its default configuration. CAD introduces a contrastive weight $(1 + \alpha)$ on the context-aware branch and $-\alpha$ on the context-free branch. We tune α on the validation set over $\{0.5, 1.0, 1.5, 2.0\}$ and report the best value per backbone. The context-free decoding branch is retained during evaluation and its computation is included in the reported inference cost (Table 5).

DAGCD. Dynamic Attention-Guided Context Decoding (DAGCD) (Huang et al., 2025) uses token-level uncertainty and attention-derived context-utilization signals to adjust token probabilities during decoding. Following the authors' implementation,[2] the context-utilization detector is trained separately for each backbone on 100 samples constructed from HotpotQA. The scaling factor is set to 2 for pretrained models and 4 for instruction-tuned models. The logistic regression classifier and utilization distribution are computed using the top-10 attention heads, with adjustments restricted to the top-10 ranked tokens. Detector training and inference-time computation are included in the corresponding resource measurements.

Context-DPO. Context-DPO (Bi et al., 2025) is a post-training alignment method that uses direct preference optimization to align LLMs toward context-faithful responses. We re-train Context-DPO separately for each evaluated backbone, including Llama-3.1-8B-Instruct, Qwen2.5-7B/14B/32B-Instruct, and Gemma-2-9B-it, using the authors' released implementation rather than their released

[1] `https://github.com/xhan77/context-aware-decoding`

[2] https://github.com/uestc-huangyw/DAGCD.

checkpoints. For all backbones, preference pairs are constructed exclusively from the ConFiQA training split. Examples derived from the same source document or conflict family are assigned to the same split, ensuring that the training and validation data are disjoint from the final evaluation set. Each model is initialized from the same native backbone used by the other methods, and a separate LoRA adapter is trained under the same data-splitting and validation protocol. No evaluation examples are used for training, hyperparameter selection, or checkpoint selection.

SinkTrack. SinkTrack (Liu et al., 2026) treats the BOS token as an information anchor and injects key contextual features into its representation via a dual-track cross-attention mechanism during the prefill stage. We follow the authors' released implementation[3] and apply its context-anchoring procedure at the designated injection layers. All method-specific settings (injection layers, cross-attention configuration, and any scaling coefficients) are selected on the validation set. The method is training-free and plug-and-play; its inference-time overhead is included in the efficiency measurements.

G.2 Internal Controls

Native. Native denotes the frozen backbone without any attention adapter, decoding intervention, or alignment training. It establishes the model's original contextual-faithfulness, capability, and efficiency levels, and serves as the reference point for all reported gains.

No-Comp. This variant applies the same routing tilt $\delta_j = \eta b_j$ as AGPR but removes the gating compensation term c. Consequently, the corrected non-sink weights do not preserve the target sink mass $s^*_{l,h}$; the induced gating drift is left unconstrained. This control isolates the contribution of the compensation term and directly tests the coupling analysis of Section 2.1.

No-Consensus. This variant replaces the layer-level consensus filter with a per-head trigger: any head with $z_{l,h} > \tau$ enters the correction stage regardless of whether its layer reaches consensus. This control tests whether layer-level aggregation provides information beyond individual head deviations, and directly addresses the site-dependence observed in Section 2.2.

Gating-Only. This variant applies the bounded gating correction (Eq. 6) but sets $\eta = 0$, so no routing tilt is applied. The non-sink distribution remains native within the corrected budget. This control tests whether gating correction alone is sufficient for improving contextual faithfulness, or whether evidence-directed routing is also necessary.

Global-Calib. This variant replaces the head-specific calibration statistics $(\mu_{l,h}, \sigma_{l,h})$ with a single global mean and variance computed across all layers and heads. All other components of AGPR remain unchanged. This control tests whether per-head calibration is necessary, or whether a global reference suffices.

[3] https://github.com/67L1/SinkTrack