Attention Sinks as Agreement Relays: Selective Collapse of Morphosyntactic Dependencies under Sink Token Removal

Ji Eun Lee
Yonsei University
jieuny@yonsei.ac.kr

Abstract

Attention sinks—initial tokens that absorb disproportionate attention mass—are widely regarded as softmax artifacts. We show they carry a second function: relaying morphosyntactic features for agreement computation. Masking sink positions across 67 BLiMP subtasks in five models (three architectures) produces a selective dissociation: agreement collapses ($\Delta = -0.50$) while binding, NPI licensing, and island constraints degrade moderately ($\Delta = -0.17$; $d = -1.89$). Activation patching isolates a write–propagate–read circuit: L0 attention writes subject number to sinks, intermediate MLPs maintain it, and cliff-layer heads consume it—early-layer restoration recovers 77% of the effect at 1B. The read phase is sparse at 1B (3 head indices) and distributed at 2.8B (32 heads), yet agreement-selective at both scales. The relay emerges at training steps 512–2000, coinciding with sink formation, and attenuates beyond $\sim$25 tokens as direct content-to-content pathways become available. German replicates both behavioral and causal signatures; Korean replicates the behavioral pattern. These results establish sinks as positional relays: emergent fixed-address loci through which transformers route grammatical features across structurally intervening material.

1 Introduction

Attention sinks—initial tokens that absorb disproportionate attention mass regardless of content (Xiao et al., 2024)—emerge early in training (Gu et al., 2025) and persist across architectures. The prevailing account treats them as numerical byproducts of softmax normalization.

This account makes a testable prediction: removing sinks should degrade all linguistic tasks uniformly. We show that this prediction is wrong. Sink removal devastates morphosyntactic agreement while leaving other phenomena largely intact.

We block attention to sink positions in Pythia (1B–6.9B), Mistral-7B, and Qwen2.5-1.5B, evaluating all 67 BLiMP subtasks (Warstadt et al., 2020). The result is a three-way gradient organized by linguistic structure: determiner–noun agreement ($\Delta = -0.11$), non-agreement phenomena ($\Delta = -0.17$), subject–verb agreement ($\Delta = -0.44$). This gradient is stable across scales and architectures ($\mathrm{SE}_\Delta \in [0.011, 0.024]$).

Activation patching traces the pattern to a three-phase circuit. L0 attention writes subject number features to sinks. Intermediate MLPs propagate them. Cliff-layer heads read them back. The write phase replicates across four models and three architectures. The read phase is sparse at 1B (3 head indices) and distributed at 2.8B (32 heads), yet agreement-selective at both scales.

Our contributions:

1. **Agreement-selective dissociation.** A three-way behavioral gradient under sink removal, replicated across five models, three architectures, and three languages.
2. **Causal relay circuit.** A write–propagate–read pathway with mechanistic detail at 1B, representational alignment across scales (RSA $\rho = 0.87$), and a shared write phase with c-command binding—establishing positional relaying as a class of mechanisms.

2 Related Work

Three questions structure prior work on attention sinks: what they are, when they emerge, and whether they affect inference. Xiao et al. (2024) identified sinks and exploited them for streaming inference. Gu et al. (2025) traced their emergence to early training steps. Sun et al. (2024) linked them to outlier activations in hidden states. A fourth question—what linguistic function sinks serve—remains unanswered. The artifact account implicitly assumes the answer is "none."

A parallel literature probes syntactic knowledge in language models (Linzen et al., 2016; Warstadt et al., 2020; Hewitt and Manning, 2019), but localizes that knowledge to heads (Clark et al., 2019; Voita et al., 2019) or layers (Vig et al., 2020), not to positions. Causal methods—activation patching (Meng et al., 2022), path patching (Wang et al., 2023), circuit discovery (Conmy et al., 2023)—have characterized how factual recall reads directly from subject positions (Geva et al., 2023) and how indirect object identification flows from subject to output (Wang et al., 2023). Agreement computation, we show, takes a different route: through sink positions, via an indirect relay absent in these prior circuits.

These two lines of work have not been connected. We know sinks exist and when they form. We know models encode syntax. No prior work has asked whether sinks participate in syntactic computation.

3 Method

3.1 Intervention: 4D Attention Mask

The intervention must block information flow from sinks while preserving all lexical content and positional encodings. Truncating the first k tokens removes content that may carry the grammatical subject (*e.g.*, "*The keys*" in "*The keys to the cabinet are rusty*").

We construct a 4D attention mask $M \in \mathbb{R}^{1\times1\times T\times T}$:

$$M_{i,j} = \begin{cases} -\infty & \text{if } j < k \ \text{ (sink columns)} \\ -\infty & \text{if } j > i \ \text{ (causal constraint)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This mask is applied additively to pre-softmax attention logits at every layer and head. All tokens remain in the input: embeddings are computed, positional encodings are assigned, and residual connections are intact. The mask blocks only attention *to* sink positions, isolating relay-mediated information flow from content removal. We evaluate $k \in \{0, 1, 4\}$.

3.2 Evaluation: BLiMP

We use all 67 subtasks of BLiMP (Warstadt et al., 2020). Each item is a minimal pair; the model succeeds if it assigns higher log-probability to the grammatical sentence. We sample 200 pairs per subtask and organize them into seven categories: determiner–noun agreement (8 subtasks), subject–verb agreement (4), anaphor agreement (2), binding (7), NPI licensing (7), island constraints (8), and other (31).

3.3 Models

Pythia (Biderman et al., 2023): 1B, 2.8B, 6.9B. All share GPT-NeoX architecture (absolute positional embeddings, parallel attention/MLP) and training data (the Pile), differing only in scale. This isolates scale effects from data and architecture confounds.

Mistral-7B-v0.1 (Jiang et al., 2023): 7.3B parameters with Grouped-Query Attention, Sliding Window Attention, and RoPE, trained on a different corpus. This tests whether the relay generalizes beyond Pythia.

Pythia-1B/2.8B and Mistral use float16. Pythia-6.9B uses 8-bit quantization for behavioral experiments (§4) and float16 for patching (§5). Baseline performance is comparable across configurations.

3.4 Patching Procedure

The activation-patching experiments reported in §5 use a uniform protocol, which we describe here for clarity.

Item selection. From the two distractor-agreement BLIMP subtasks (the relational-noun and relative-clause variants),[1] we select pairs on which the un-patched model assigns higher log-probability to the grammatical sentence (i.e., the model is initially correct). We use $n = 100$ such pairs per architecture unless otherwise noted; the $n = 150$ figure in §4.1 refers to a combined pool used for the sink-corruption test.

Source-target pairing. Each target sentence is paired with three source sentences, one per condition. Let the target's grammatical subject have number $\nu \in \{\text{sg}, \text{pl}\}$.

- **Same-number** ($\nu_{\text{src}} = \nu$): a different pair from the same distractor-agreement pool whose grammatical sentence has the same subject number as the target.
- **Cross-number** ($\nu_{\text{src}} \neq \nu$): a different pair from the same pool whose grammatical sentence has the opposite subject number.
- **Random control**: a sentence drawn from an unrelated BLIMP subtask (a binding or

[1] Full identifiers in the BLIMP release: `distractor_agree-ment_relational_noun` and `distractor_agree-ment_relative_clause`.

anaphor pair), with no controlled number relation to the target.

Pairings are fixed by a deterministic offset over the item index (seed $= 42$), so that each target receives the same three sources across all layers and across runs.

Patching operation. For each (target, source) pair and each layer $\ell \in \{0, \ldots, L-1\}$ *independently*, we replace the hidden states at the sink positions $\{0, 1, 2, 3\}$ at the output of layer ℓ with the source's hidden states at the corresponding positions (layer ℓ output, same positions). All other hidden states are left untouched, the forward pass is run to completion, and we record the top-1 next-token prediction at the final position of the target. A *flip* is defined as a change in this top-1 prediction relative to the un-patched forward pass on the target.

The flip rate at layer ℓ in condition c is then the fraction of targets that flip, $f_c(\ell) \in [0, 1]$. Layers are patched one at a time; no cumulative or multi-layer patching is reported in §5.

Cliff definition. Reading the random-condition curve $f_{\text{rand}}(\ell)$ as the "upper envelope" of disruption, we define the *cliff layer* ℓ^* as the layer at which the disruption ceases to propagate to the prediction:

$$\ell^* = \arg\max_{\ell} \left[f_{\text{rand}}(\ell - 1) - f_{\text{rand}}(\ell + 1) \right],$$

i.e., the layer with the largest discrete second-derivative drop in flip rate.[2] For each model, the cliff defines the boundary between the *propagate* and *read* phases of the relay circuit; pre-cliff flips reflect disruption to features still in transit, while post-cliff flips require disrupting features after they have been consumed by the agreement-resolution heads. Cliff layers reported in §5 are L12–13 (Pythia-1B), L20–21 (Pythia-2.8B), L17–18 (Pythia-6.9B), and L22–23 (Mistral-7B).

Same vs. cross vs. random. The interpretive logic of the three conditions is as follows. A position that merely buffers an arbitrary signal would be equally disrupted by any non-identical source: same, cross, and random would yield equal flip rates. A position that carries a content-specific feature would be *less* disrupted by sources sharing that feature (same) than by sources lacking it (cross/random). The observed ordering "same $\ll$ cross/random" (§5.2) thus distinguishes a content-relay site from a generic positional artifact. The reversed cross/random ordering observed in Mistral-7B (cross slightly above random for early layers, then matched at the cliff) is consistent with the same logic provided one assumes that *incompatible* number features actively interfere, rather than merely being inert; we treat this as an architectural variant whose source is unknown.

Reproducibility. The same patching code, item selection, and random seed are used across all four architectures. Source code and item lists will be released upon publication.

4 Results

4.1 Agreement-Selective Collapse

Table 1 presents the central result. Sink removal ($k = 4$) produces a sharp dissociation: agreement degrades by $\Delta = -0.50$, non-agreement by $\Delta = -0.17$ (2.9×; Mann–Whitney $p < 10^{-6}$, Cohen's $d = -1.89$). The effect survives difficulty control: subtask-level regression yields $\beta_{\text{agree}} = -0.18$ ($p < 0.001$) after partialing out baseline accuracy.[3]

Within agreement, the dissociation sharpens into a three-way gradient: determiner–noun $\Delta = -0.11$, subject–verb $\Delta = -0.44$, anaphor $\Delta = -0.57$. The anaphor collapse is noteworthy: despite short linear distance, same-number patching flips only 6% of anaphor predictions vs. cross-number 60% ($n = 100$)—a steeper gradient than SV agreement. Number/gender asymmetry (-0.73 vs. -0.38) indicates feature-specific relay usage (Appendix D).

Sinks are the most critical single component. Corrupting only sinks collapses agreement to .220 ($\Delta = -0.71$, $n = 150$). This is worse than corrupting *all* positions ($\Delta = -0.41$): intact context without sink features actively misleads the agreement circuit. Restoring sinks alone within a fully corrupted context recovers only 10%. The relay is necessary but not sufficient—it operates within a globally coherent computation.

[2] Equivalently, the layer index that produces the sharpest negative bend in the curve; we use this rather than a fixed-threshold definition because the absolute flip-rate level varies with model scale, while the qualitative cliff is scale-invariant in the relative-depth sense (§5.2).

[3] Per-item GEE: agreement×masking $\beta = -3.12$, OR $= 0.044$, $p < 10^{-10}$.

Table 1: Category-level accuracy under sink removal ($k = 4$). Δ = change from baseline. Cross-model SE of Δ: 0.011–0.024. Per-subtask results in Appendix A.

	Pythia-1B			Pythia-2.8B			Pythia-6.9B			Mistral-7B		
Category	$k{=}0$	$k{=}4$	Δ	$k{=}0$	$k{=}4$	Δ	$k{=}0$	$k{=}4$	Δ	$k{=}0$	$k{=}4$	Δ
Agreement phenomena												
Det–noun	.971	.849	−.122	.979	.838	−.141	.979	.851	−.128	.949	.912	−.037
Anaphor	.998	.430	−.568	1.00	.435	−.565	.995	.452	−.543	.988	.390	−.598
Subj–verb	.915	.461	−.454	.909	.448	−.461	.915	.432	−.483	.854	.478	−.376
Non-agreement phenomena												
Binding	.804	.639	−.165	.799	.657	−.142	.815	.674	−.141	.838	.726	−.112
NPI licensing	.733	.579	−.154	.760	.534	−.226	.753	.561	−.192	.784	.536	−.248
Island	.649	.505	−.144	.682	.495	−.187	.732	.519	−.213	.710	.546	−.164
Other	.842	.692	−.150	.851	.678	−.173	.847	.663	−.184	.815	.705	−.110

Table 2: Masking intensity ($k = 1$ vs. $k = 4$), averaged across three Pythia scales. Mistral $k = 1$ results are consistent (supplementary material).

Category	$k{=}0$	$k{=}1$	$k{=}4$
Det–noun agree.	.976	.973	.846
Anaphor agree.	.998	.875	.439
Subj–verb agree.	.913	.713	.447
Binding	.806	.807	.656
NPI licensing[4]	.749	.519	.558
Island	.688	.556	.506
Other	.847	.785	.678

The dissociation is qualitative, not quantitative. Three-condition patching on binding yields reversed gradients (same .430, cross .550, random .740)—the opposite of agreement's same $\ll$ cross/random pattern. Non-agreement degradation does not correlate with sentence length ($\rho = -0.30$, $p = 0.20$, $n = 20$); agreement degradation does ($\rho = -0.70$, $p = 0.01$, $n = 12$).

Masking profiles. Agreement collapses monotonically as k increases (.945 $\rightarrow$ .350; $n = 200$). NPI shows a U-shaped profile (.580 $\rightarrow$ .485 $\rightarrow$ .600 $\rightarrow$ .490) with no content sensitivity. Sinks do not relay polarity features. Figure 1 visualizes the dissociation across scales.

4.2 Masking Intensity: $k = 1$ vs. $k = 4$

Masking only the first token ($k = 1$) already collapses agreement substantially while leaving non-agreement intact (Table 2). The first token is the primary relay.

[4]NPI shows a non-monotonic masking profile (Appendix B).

4.3 Scale and Architecture Invariance

Within Pythia, Δ values are stable across scales: subject–verb $\Delta \in [-0.45, -0.48]$, anaphor $\Delta \in [-0.54, -0.57]$, determiner–noun $\Delta \in [-0.12, -0.14]$.

Mistral-7B reproduces the pattern despite differing in attention mechanism (GQA vs. MHA), positional encoding (RoPE vs. absolute), and window structure (sliding vs. full causal): anaphor $\Delta = -0.60$, subject–verb $\Delta = -0.38$, determiner–noun $\Delta = -0.04$, non-agreement $\Delta = -0.16$. The replication extends to causal analysis: Mistral shows a patching cliff at L22–23 (§5.2).

Qwen2.5-1.5B (Team, 2024) provides a third architecture family. The three-way gradient replicates: long-range $\Delta = -0.53$, det–noun $\Delta = -0.22$, non-agreement $\Delta = -0.12$ (ratio: $4.4\times$; $n = 200$ per subtask).

5 Causal Evidence for the Relay

The behavioral dissociation establishes selectivity. We now identify the mechanism.

5.1 Number Probing

Linear probes (logistic regression, 10-fold CV) trained on 1,556 balanced sentences decode grammatical number with > 0.95 accuracy at sinks, subject positions, last tokens, and random content positions. Number features are encoded everywhere. The relay hypothesis therefore rests not on a *storage* privilege but on a *usage* privilege: the agreement circuit reads from sinks preferentially. The behavioral dissociation and causal patching below establish this usage privilege. Its co-emergence with sinks during training (§6.1) indicates the circuit co-opts sinks as a fixed-address interface.

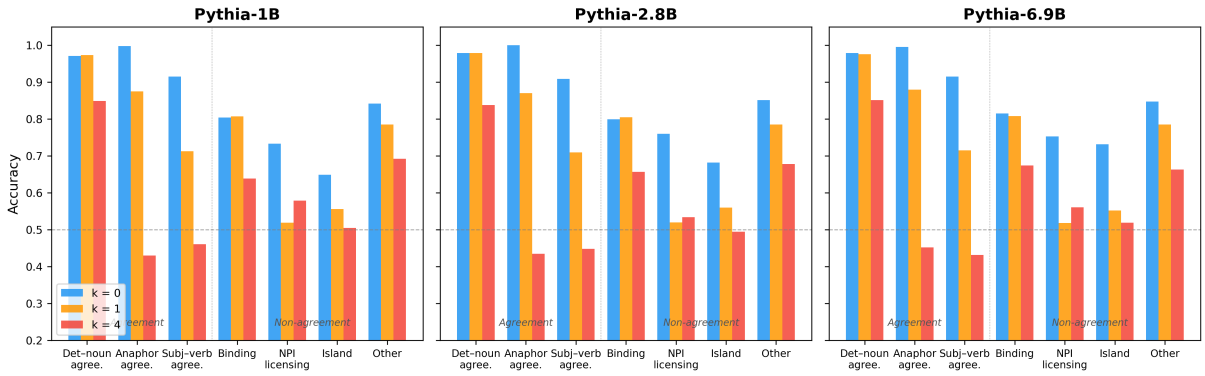

Figure 1: Category-level accuracy under three sink masking conditions ($k \in \{0, 1, 4\}$) across Pythia-1B, 2.8B, and 6.9B. Dashed line: chance (0.5). Agreement collapses to near-chance under $k = 4$; non-agreement remains above chance.

5.2 Activation Patching

Following the patching protocol of §3.4, we apply per-layer activation patching to BLIMP's distractor agreement subtasks: for each target item we replace sink hidden states (positions 0–3) at the output of one layer with those from a source sentence under three conditions (same-number, cross-number, random) and record whether the top-1 next-token prediction flips relative to the un-patched run.

Content sensitivity. All four models show a clear gradient (Table 3). For Pythia-1B: same 16%, cross 43%, random 58% (Cochran–Armitage $Z = 15.0$, $p < 10^{-10}$). Pythia-2.8B replicates ($Z = 20.3$, $p < 10^{-10}$). Mistral-7B shows same (14.8%) $<$ random (27.5%) $<$ cross (39.5%), all pairwise significant (Fisher's exact, $p < 0.02$). The reversed cross/random ordering in Mistral suggests incompatible number actively interferes; the architectural source is unknown. The key contrast—same $\ll$ cross/random—holds in all models, ruling out the alternative that sinks are merely fragile positions.

Cliff localization. All conditions share the same cliff—the layer at which the circuit consumes sink-stored features. The cliff is at L12–13 (1B), L20–21 (2.8B), L17–18 (6.9B), L22–23 (Mistral; Figure 2). Larger models complete agreement computation earlier in relative terms: 6.9B at 56% of depth vs. 2.8B at 63%.

Write phase: layer-selective masking. Masking sinks at early layers alone reproduces nearly the full collapse. Pythia-1B: L0–5 $\Delta = -0.56$ vs. full -0.55 ($n = 300$). Pythia-2.8B: L0–9 $\Delta = -0.42$ vs. full -0.46 ($n = 200$). Cliff-layer masking has negligible effect (1B L12–15: $\Delta = -0.01$; 2.8B L20–31: $\Delta = -0.04$). Qwen2.5-1.5B confirms: L0 masking alone collapses agreement from .870 to .210 ($\Delta = -0.66$, exceeding full-mask $\Delta = -0.63$); binding is unaffected ($\Delta = +0.03$). The write phase replicates across four models and three architectures.[5]

Table 3: Activation patching flip rates across three conditions and four models ($n = 100$; Mistral: 84 valid). Layer bins defined by cliff location (§3.4; see Figure 2).

Condition	**Pythia-1B** (16 layers)			
	0–5	6–10	11–12	13–15
Same-number	.157	.078	.040	.030
Cross-number	.433	.418	.400	.087
Random control	.580	.430	.395	.080
	Pythia-2.8B (32 layers)			
	0–7	8–20	21–28	29–31
Same-number	.111	.077	.056	.040
Cross-number	.401	.409	.081	.053
Random control	.601	.499	.135	.083
	Pythia-6.9B (32 layers)			
	0–7	8–17	18–28	29–31
Same-number	.081	.048	.028	.020
Cross-number	.482	.483	.193	.080
Random control	.696	.569	.260	.130
	Mistral-7B (32 layers)			
	0–7	8–22	23–28	29–31
Same-number	.265	.220	.037	.000
Cross-number	.573	.560	.100	.000
Random control	.532	.433	.110	.013

5.3 The Write–Propagate–Read Circuit

Write. Zeroing L0 attention output at sink positions ($n = 150$) collapses agreement selectively: 1B $\Delta_{\text{agree}} = -0.67$, $\Delta_{\text{bind}} = -0.07$; 2.8B $\Delta_{\text{agree}} = -0.64$, $\Delta_{\text{bind}} < 0.11$. L0–L2 MLPs con-

[5] Layer-selective masking was infeasible for 6.9B/Mistral due to GPU memory. Mistral's cliff at L22–23 is consistent but unconfirmed by direct masking.

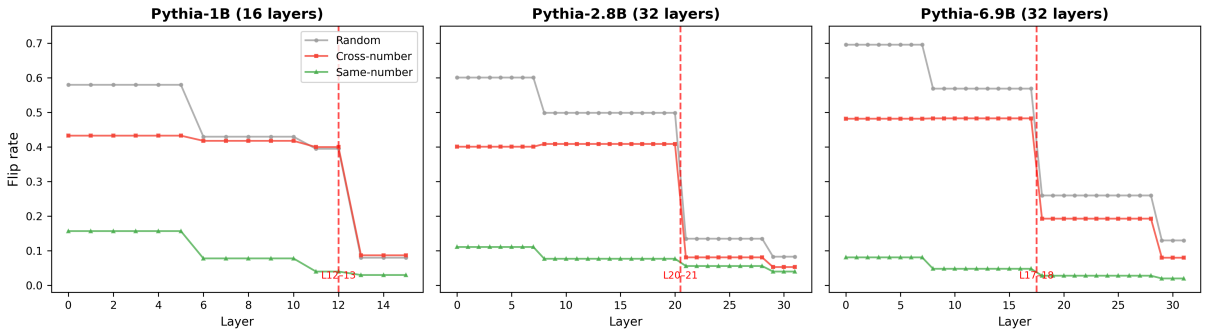

Figure 2: Per-layer activation patching across three conditions and three model scales. The ordering same-number $<$ cross-number $<$ random is consistent. Red dashed lines mark the cliff: L12–13 (1B), L20–21 (2.8B), L17–18 (6.9B).

solidate the written features; interventions at L3+ are inert. Directed measurement confirms the flow direction: corrupting the subject at L0 shifts sink activations at L1 ($\cos = 0.87$ with clean); corrupting a random position does not ($\cos = 1.00$; $p < 10^{-6}$, $n = 100$).

Propagate. Path patching in Pythia-1B identifies the transmission medium. Restoring MLP output at L12 recovers 57% of the sink-removal effect; restoring attention recovers nothing. Combined restoration of L0–12 recovers 77%; the remaining 23% falls to post-cliff decoding layers. Cosine similarity between clean and masked sink representations drops sharply in L0–L3 (L1: 0.62 $\rightarrow$ L3: 0.07), then plateaus at $\sim$0.48 through L14. Linear probing for number at sinks shows stable accuracy ($\sim$.76) across L0–L15, confirming number features persist without further transformation once consolidated.

Read at 1B: sparse. Ablating each of 128 heads (16×8) reveals 16 that selectively damage agreement ($\Delta_{\text{agree}} < -0.03$, $\Delta_{\text{bind}} > -0.01$). These map to three head indices—H0, H3, H6—repeated across layers. L12H0 is maximally disruptive ($\Delta_{\text{agree}} = -0.450$, $\Delta_{\text{bind}} = +0.005$), converging with the patching cliff. The sparsity is striking: 3 of 8 head indices carry the entire relay function.

Read at 2.8B: distributed. Exhaustive ablation at L20 yields $\Delta_{\text{agree}} = -0.007$ ($n = 150$, $>$80% power for $\Delta > 0.10$). The relay is present—a linear probe at L20 sinks decodes number with .95 accuracy—but the read is distributed. Flipping the number-direction component degrades agreement by $\Delta = -0.13$; zeroing yields only -0.02. Cumulative ablation quantifies the distribution: per-head contribution is $\sim$0.013, accumulating linearly to $\Delta = -0.40$ at full ablation ($n = 100$). Binding remains intact ($\Delta = +0.01$). RSA between 1B L12 and 2.8B L20 sink activations: $\rho = 0.87$ ($p < 10^{-25}$; $n = 13$, 78 pairs). Flipping top-3 PCA directions at L20: $\Delta = -0.35$ (78% of full-mask vs. 13% for single direction).

Sinks vs. subject: ruling out the masking confound. A natural alternative interpretation of the masking results is that positions 0–3 typically contain the determiner and head noun of the subject, so the agreement collapse reflects the loss of *subject* information rather than a sink-specific effect. Three observations already presented above rule this out. First, sink removal collapses agreement even in 3-token SVO sentences ($\Delta = -0.47, n = 250$), where the subject occupies position 0 and is by construction included in any sink mask: if subject inaccessibility were the operative cause, the binding control would collapse equally, which it does not. Second, masking four tokens at five different positions ($n = 100$) shows that only the sink positions and the subject positions produce agreement-selective collapse ($\Delta_{\text{agree}} = -0.39$ and -0.74; $\Delta_{\text{bind}} = -0.06$ and 0.00), while mid-sentence (4–7), last-four, and random four-position masking all degrade agreement and binding to a comparable extent ($\Delta \approx -0.20$–0.33 in each case, no selectivity). Sink and subject behave as components of a single agreement circuit, not as interchangeable sources of the same information; other positions, including the four *immediately following* the sink region, do not. Third, what occupies the sink matters: prepending "Meanwhile," increases the masking effect (-0.63 vs. -0.39), while content-rich "The story is that" reduces it (-0.34; $n = 100$). A generic positional-fragility account predicts no such content dependence. To-

gether with the directed-flow result (subject corruption shifts sink representations, cos $= 0.87$ with clean; random corruption does not, cos $= 1.00$; $p < 10^{-6}$, $n = 100$; above) these observations establish that sinks are a content relay rather than a positional artifact, and that the sink-masking effect is not reducible to subject loss.

6 Generalization

6.1 Training Dynamics

The relay hypothesis predicts no agreement effect before sinks emerge and abrupt onset at formation. Both predictions are confirmed. Pythia-1B ($n = 100$): step 512 $\Delta_{\text{agree}} = +0.06$; step 2000 $\Delta = -0.56$ (selectivity $3.1\times$). Pythia-2.8B ($n = 50$): $+0.06 \rightarrow -0.68$; binding stable ($\Delta \approx -0.06$). The coupling with sink formation (Gu et al., 2025) is not predicted by the artifact account.

6.2 Ecological Validity

BLiMP uses controlled minimal pairs. We test whether the effect holds in naturalistic settings. On 1,000 Linzen-style sentences (Linzen et al., 2016) with 0–3 attractors ($n = 250$ each), sink removal degrades agreement from .752 to .569 ($\Delta = -0.18$, 95% CI $[-0.22, -0.14]$, $p < 10^{-6}$). The effect persists without attractors ($\Delta = -0.46$). WikiText-103 (Merity et al., 2016): $\Delta = -0.29$ for $\leq$10-token prefixes, -0.04 for 36+, confirming dual-pathway attenuation.

6.3 Cross-Linguistic Evidence

German SVO. LeoLM-7B (Plüster, 2023), $n = 200$: $\Delta = -0.52$ (95% CI $[-0.59, -0.45]$; McNemar $p < 10^{-8}$). Random-source patching at L0 flips 51% of predictions ($n = 39$, CI $[35\%, 67\%]$), dropping to 8% at L8. German replicates both behavioral and causal signatures.

Korean SOV. Polyglot-ko-1.3B (Ko et al., 2023), $n = 150$: case-marking collapses from 1.00 to .40 ($\Delta = -0.60$, CI $[-0.68, -0.52]$; $p < 10^{-10}$). Non-agreement less affected (ratio $1.9\times$). Lower selectivity likely reflects coarser controls and ceiling compression; subject–sink overlap in SOV may inflate the effect. Korean replicates behavior; causal confirmation awaits.

6.4 Locality of the Relay

The behavioral data presented so far is unified by a single structural property: the relay is recruited as a function of how local the dependency is. Two converging observations establish this. The first concerns dependencies whose distance can be varied externally (context length); the second concerns dependencies whose internal structural complexity differs while their phenomenon type is held constant (c-command vs. local binding). Together they explain the apparent sparing of determiner–noun agreement reported in §4.1.

Context length attenuates the relay. The relay's importance diminishes with context (n = 80 per bin): $\Delta = -0.31$ at 6–8 tokens, -0.13 at 12–17, -0.05 at 25–34, near zero beyond 46. Verb attention to sinks drops from .82 (6 tokens) to .45 (66 tokens). This implies a dual-pathway architecture: the relay dominates short contexts; direct content-to-content pathways emerge with length. BLIMP items average ~ 12 tokens, placing the benchmark squarely in the relay-dominant regime. WikiText-103 confirms the pattern in naturalistic text: sink removal increases NLL by $+3.65$ nats for 11–20-token prefixes vs. $+2.41$ for ≥ 36 tokens ($n = 300$). The dual-pathway account connects to training: the relay emerges at steps 512–2000 (§6.1), suggesting it is the first agreement strategy acquired, with direct pathways consolidating later.

Within-phenomenon distance: c-command binding. If sinks relay structural features generally, the write phase should extend beyond agreement to other long-distance dependencies. We test this on binding, where structural distance varies systematically within a single phenomenon. L0-only masking ($n = 100$) produces a $4\times$ differential: c-command binding $\Delta = -0.24$ vs. local binding $\Delta = -0.06$. Under full sink masking, c-command collapses to $\Delta = -0.54$ while local binding remains essentially spared. By contrast, subject-verb agreement shows $\Delta_{L0} = -0.69 > \Delta_{\text{full}} = -0.39$: a shared write phase, but different read depths. Positional relaying is thus a class of mechanisms rather than a single circuit, but the common thread is locality: in both phenomena, the relay carries the load only when the dependency cannot be resolved by direct attention across the intervening material.

Det-noun agreement and the locality account. This locality structure explains the determiner–noun result reported in §4.1, where masking the sinks produces $\Delta = -0.11$ for det-noun vs. $\Delta = -0.44$ for subject–verb. In BLIMP, det-noun dependencies are predominantly adjacent (the determiner and its noun typically occupy consecutive

positions), placing them in the regime where direct attention across a one-token span is unobstructed and the relay is not recruited. Under the locality account, det-noun is not an exception to the relay claim but a prediction of it; the prediction that stretched det-noun dependencies should reintroduce the masking effect remains to be tested directly (see Limitations).

7 Discussion

Redistribution does not explain the data. Same-number patching disrupts at 5–15%; random at 28–58%. Entropy changes uniformly ($\Delta H \approx -0.48$ nats). If redistribution were the mechanism, replacement content should not matter.

Necessary bottleneck. Restoring sinks within fully corrupted context recovers only 10%. The relay carries critical features, but the receiving circuit requires intact context. Sinks are necessary but not standalone.

Dual function of sinks. The artifact account predicts uniform degradation under removal. We observe selective degradation. Sinks stabilize softmax (Xiao et al., 2024) *and* relay morphosyntactic features.

KV cache compression. Eviction collapses agreement: Pythia-1B .951 → .378; Mistral .890 → .505. Recovery is 3× agreement vs. non-agreement (Appendix B). StreamingLLM (Xiao et al., 2024) retains sinks—our results explain why. H2O (Zhang et al., 2023) may deprioritize sinks in long contexts; our results predict selective agreement degradation.

8 Limitations

(i) *Sparse-to-distributed read transition.* The transition coincides with a 4× head-count increase (8 at 1B, 32 at 2.8B). Pythia confounds head count with model size; varying heads at fixed parameters would isolate the cause. Full DAS (Geiger et al., 2023) could recover the complete distributed circuit.

(ii) *Write-phase confirmation in Mistral.* Layer-selective masking requires simultaneous clean/masked activations at all layers, exceeding our GPU budget for 6.9B and Mistral. The write phase is confirmed in 1B, 2.8B, and Qwen; Mistral's patching cliff is consistent but unconfirmed by direct masking.

(iii) *Cross-linguistic depth.* German replicates behavioral and causal signatures. Korean replicates behavior with a single model and coarser controls. Circuit characterization in morphologically rich languages—including SOV languages, agglutinative languages, and languages with case stacking—remains an open direction. A wider Korean test bed beyond case-marking and a causal-patching analysis adapted to SOV morphology would test the generality of the relay claim.

(iv) *Propagation dynamics.* Cosine analysis (§5.3) suggests active transformation in L0–L3 and passive maintenance thereafter. Subspace tracking (Geiger et al., 2023) would determine whether intermediate layers contribute further transformations or merely propagate consolidated features.

(v) *Additional controls on the subject-masking confound.* The position-set comparison (§5.3), the sink-only sufficiency test (§4.1), and the directed-flow analysis jointly establish that the agreement collapse cannot be reduced to subject masking: only sinks and subject produce agreement-selective collapse among five tested position sets; restoring sinks alone in a fully corrupted context recovers only 10%; and information flow is directional (subject → sink, not the reverse). Two further controls would sharpen this conclusion: a two-site joint sufficiency test (does jointly restoring sink and subject yield additive recovery beyond restoring either alone, or only the same recovery, as our circuit account predicts?) and a write-head/read-head dissociation test (are the heads that write subject features into sinks disjoint from the heads that read those features at the cliff layer?). Preliminary work on both points is consistent with the relay model but a systematic treatment is deferred to future work.

(vi) *The determiner–noun asymmetry.* Determiner–noun agreement shows a smaller masking effect ($\Delta = -0.11$) than subject–verb agreement ($\Delta = -0.44$) or anaphor agreement ($\Delta = -0.57$). The locality analysis of §6.4 suggests this is not a counterexample to the relay account but

a prediction of it: det-noun dependencies in BLIMP are typically adjacent (linear distance ≤ 1) and fall in the regime where direct content-to-content attention is sufficient, so the relay is not recruited. This prediction is testable: stretching det-noun dependencies via intervening material—for example, German articles separated from their nouns by attributive phrases, or English *the very tall old* NOUN configurations—should reintroduce the masking effect. We did not attempt such a stretched test in this work.

(vii) *NPI licensing mechanism.* NPI subtasks show a non-monotonic, U-shaped masking profile (§4.1) and no content sensitivity. We interpret this as evidence that NPI licensing does *not* use the sink relay; what licensing mechanism it does use, and whether NPI scope is computed via a distinct positional relay or a fundamentally different routing strategy, are left open.

(viii) *Fine-grained island variability.* The non-agreement category aggregates seven binding subtasks, seven NPI subtasks, and eight island subtasks. Within islands, individual subtasks show substantial variability (Appendix C): CSC object extraction collapses sharply ($\Delta = -0.70$), while complex-NP island is essentially unaffected ($\Delta = -0.04$). The claim that "non-agreement remains above chance" (Figure 1, caption) is correct in the categorical aggregate but smooths over this subtask-level variation; readers concerned with individual island phenomena are referred to Appendix C.

(ix) *Item-length stratification.* Our category-level averages do not stratify by item token length, which our own locality analysis (§6.4) shows is a relevant modulating variable. A length-stratified per-subtask table would expose whether the agreement-selective collapse holds uniformly across the length distribution or whether short and long items contribute differently to the average. We provide the unstratified table here and leave length stratification to follow-up work.

(x) *Architectural source of Mistral's cross/random reversal.* Mistral-7B's activation-patching curves (Table 3) show the cross-number condition exceeding the random condition in the early layers, reversing the ordering observed in the Pythia and Qwen families. The same-number condition remains the lowest, so the key content-sensitivity contrast holds. The source of the reversal is unknown; it may relate to Mistral's Grouped-Query Attention or Sliding Window Attention, but we do not isolate the cause.

9 Conclusion

Attention sinks are positional relays. They carry grammatical features through a write–propagate–read circuit. The write phase is shared across agreement and c-command binding. The read phase transitions from sparse to distributed while preserving representational geometry (RSA $\rho = 0.87$). The relay emerges with sinks during training and precedes its causal role.

The relay is a default short-context mechanism. Direct pathways supplement it as context grows. This trajectory—positional shortcut first, distributed computation later—may apply beyond agreement. C-command binding shares the write phase but reads at different depths, suggesting a class of relay mechanisms whose full inventory remains to be characterized. For practitioners: KV cache compression should evaluate agreement in short-context regimes.

Reproducibility. Code and evaluation scripts will be released upon publication. All models and benchmarks are publicly available.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2023. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the 2023 Conference on Causal Learning and Reasoning*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and 1 others. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Björn Plüster. 2023. Leolm: Linguistically enhanced open language model for german. *Hessian.AI*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Qwen Team. 2024. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, and 1 others. 2023. H2O: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*.

A Subtask-Level Results

Table 4 reports per-subtask $\Delta(k = 4)$ averaged across Pythia scales. No agreement subtask exceeds $\Delta = -0.30$; no non-agreement subtask falls below $\Delta = -0.35$ in the multi-model average.

Cliff validation. The L12–13 cliff is not an averaging artifact: 86/100 targets show decreased logit shift from L12 to L13 (agreement); 84/100 (random). Mean: $+2.86 \rightarrow +1.06$ (63% reduction).

Table 4: Representative per-subtask $\Delta(k = 4)$, Pythia average.

Category	Subtask	Δ
LR agree.	distractor_agree_rel_noun	-0.42
	distractor_agree_rel_clause	-0.48
	irreg_plural_subj_verb_1	-0.50
Anaphor	anaphor_number_agreement	-0.73
	anaphor_gender_agreement	-0.38
Det–noun	determiner_noun_agree_1	-0.09
	det_noun_agree_irreg_1	-0.17
Binding	principle_A_c_command	-0.15
	principle_A_domain_2	-0.25

B Additional Controls

Window size. Agreement recovery: $+0.68$ ($W = 4$) to $+0.74$ ($W = 8$–32). Non-agreement $\leq +0.02$.

Quantization. Cliff at L20→21 under 8-bit ($n = 50$).

NPI. Monotonic agreement collapse; U-shaped NPI (scope conflict).

KV cache. Simulated ($W = 8$): agreement $+.32$, non-agreement $+.12$ ($3\times$). Actual: 1B $.951 \rightarrow .378$; Mistral $.890 \rightarrow .505$.

C Full Subtask Results

Table 5: all 32 paradigms, Pythia-1B ($n = 50$). Outliers: princ. A c-command (-0.54), princ. A case 2 (-0.48), CSC object extr. (-0.70), wh-subj gap (-0.48). These involve structural computation overlapping with the relay (§6). Category means: agreement -0.43, non-agreement -0.21.

Table 5: Full per-subtask results, Pythia-1B ($n = 50$).

Subtask	Cat.	Base	Mask	Δ
dist. agree. rel. noun	SV	.94	.58	−.36
dist. agree. rel. clause	SV	.88	.30	−.58
irreg. pl. SV agree. 1	SV	1.0	.58	−.42
irreg. pl. SV agree. 2	SV	.90	.46	−.44
reg. pl. SV agree. 1	SV	1.0	.50	−.50
reg. pl. SV agree. 2	SV	.94	.44	−.50
anaphor gender	Ana	1.0	.22	−.78
anaphor number	Ana	1.0	.52	−.48
det–noun 1	DN	1.0	.68	−.32
det–noun 2	DN	1.0	.68	−.32
det–noun irreg. 1	DN	.94	.80	−.14
det–noun irreg. 2	DN	1.0	.64	−.36
princ. A c-command	Bind	.84	.30	−.54
princ. A case 1	Bind	1.0	.96	−.04
princ. A case 2	Bind	.94	.46	−.48
princ. A domain 1	Bind	1.0	.94	−.06
princ. A domain 2	Bind	.84	.84	.00
princ. A domain 3	Bind	.66	.50	−.16
NPI present 1	NPI	.62	.60	−.02
NPI present 2	NPI	.62	.48	−.14
only NPI licensor	NPI	.90	.72	−.18
only NPI scope	NPI	.76	.50	−.26
sent. neg. NPI lic.	NPI	1.0	.90	−.10
sent. neg. NPI scope	NPI	.82	.54	−.28
adjunct island	Isl	.86	.76	−.10
complex NP island	Isl	.56	.52	−.04
CSC complex left	Isl	.72	.30	−.42
CSC object extr.	Isl	.86	.16	−.70
left branch echo	Isl	.46	.62	+.16
left branch simple	Isl	.62	.44	−.18
wh-island	Isl	.78	.58	−.20
wh-subj gap	Isl	.96	.48	−.48

D Anaphor Agreement

Anaphor collapses $\Delta = -0.56$ despite $\sim$8-token distance. Matched pairs ($n = 100$): PP -0.49 (baseline .97), RC -0.27 (.76), cross-clause -0.04 (.56). Binding at $\sim$.80 baseline: only -0.15. Task type, not difficulty, determines vulnerability. Number/gender asymmetry (-0.73 / -0.38) and patching (same 6% vs. cross 60%) confirm feature-specific relay usage.