

Same Sink, Different Plumbing: Reproducing and Refining the Mechanism of Attention Sinks

Abdullah Al Zubayer^{1*}, Syed Naveed Mahmood^{1*}, Md. Rezaur Rahman Bhuiyan¹,
Mir Md Mohiuddin Abrar¹, Jareen Tasneem Khondaker¹, Tasfia Zaman¹,
Farig Sadeque¹

¹Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
{abdullah.al.zubayer, syed.naveed.mahmood}@g.bracu.ac.bd
{rezaur.rahman.bhuiyan, mir.md.mohiuddinabrar}@g.bracu.ac.bd
{jareen.tasneem.khondaker, tasfia.zaman}@g.bracu.ac.bd
farig.sadeque@bracu.ac.bd

Abstract

Attention sinks assign large, largely content-independent attention to the first position. [Ran-Milo et al. \(2026\)](#) explain GPT-2’s sink with a parameter-level circuit: a learned query bias b_Q aligns with the key projection of the first position’s effective positional encoding, whose few massive coordinates amplify the score. They infer that other architectures build sinks differently but do not measure this claim. We reproduce the original intervention effects and stress-test their causal interpretation. Baseline plus nine interventions and four structural analyses reproduce under the original evaluation, with two scope qualifications; the massive coordinates are exactly $\{138, 378, 447\}$. Yet nullifying b_Q leaves 44.5% of the sink, which follows an activation-dependent query route into the same first-position key. This *anchor-route* distinction unifies the extensions: with GPT-2 scale, the bias route fades while the activation-dependent route concentrates into fewer carrier heads; across 16 checkpoints in four families, no qualitative profile over mutually available conditions spans all families; and across five pre-training runs, anchoring recurs despite unstable coordinate identities. In GPT-2-small at 40 tokens, Swap EPE incurs about one quarter of the cross-entropy cost of Remove First PE. The stable phenomenon is an anchoring computation, not a universal parameter-level circuit, so mitigation must be validated against the target model’s wiring.

1 Introduction

Autoregressive Transformers ([Vaswani et al., 2017](#)) often devote a large, largely content-independent share of attention to the first position: the *attention sink* ([Xiao et al., 2024](#); [Gu et al., 2025](#)). It shapes streaming and KV-cache policy ([Xiao et al., 2024](#)),

produces activation outliers that complicate quantization ([Sun et al., 2024](#); [Lin et al., 2024](#)), interacts with long-context behaviour and over-mixing ([Barbero et al., 2025](#); [de Llano et al., 2026](#)), and can dominate attention maps and mislead interpretation ([Guo et al., 2024](#); [Jain and Wallace, 2019](#); [Yu et al., 2024](#)).

[Ran-Milo et al. \(2026\)](#) give the most concrete mechanistic account to date. In GPT-2 ([Radford et al., 2019](#)), a learned query bias b_Q aligns with the key projection of the first position’s effective positional encoding; a few *massive* coordinates ([Sun et al., 2024](#)) amplify a source-agnostic score shift toward position 0. Their baseline and nine interventions show substantial causal contributions and editability, but do not establish a complete, strictly necessary mechanism.

The paper also infers *distinct circuits across architectures*: every identified component is absent from some architecture that still has a sink, but this consequential claim rests on citation rather than direct evidence. Mechanistic results based on one checkpoint, run, scale, metric, and short inputs carry known reproducibility risks ([Pineau et al., 2021](#); [Bouthillier et al., 2021](#); [D’Amour et al., 2022](#); [Tigges et al., 2024](#)). Through an end-to-end reproduction and new data, we address the original paper’s four acknowledged limitations: architecture and scale, learning dynamics, an unexplained residual sink, and mechanism versus function.

The original intervention effects reproduce, but their causal interpretation does not fully survive stress testing (§2). The residual reveals a *shared positional anchor* reached through distinct routes (§3); this anchor persists across GPT-2 scale, training runs, input lengths, and seven content regimes, while its supporting objects and routes vary across architectures (§3–§5). At 40 tokens, Swap EPE is

*These authors contributed equally.

substantially less costly than Remove First PE (§6). Two execution paths and parity gates validate the tested GPT-2 quantities within the coverage in Appendix H. We release code, configurations, seeds, and the complete result archive¹.

2 The Circuit, and Whether It Reproduces

Preliminaries and terminology. We index sequence positions from zero, so position 0 is the first content token; this corresponds to position 1 in Ran-Milo et al. (2026). For layer ℓ and head h , x_i denotes the residual-stream representation at position i , W_q and W_k the query and key projections, and b_Q the learned query bias. A *head-cell* is one (ℓ, h) pair. Our sink metric is mean attention from second-half source tokens to position 0, averaged over heads and a *mid-layer band*: layers 4–11 for 12-layer models and $[3, L - 1)$ otherwise.

GPT-2 adds a learned positional embedding (PE) p_j to the token embedding. Following Ran-Milo et al. (2026), the *effective positional encoding* (EPE) is

$$\text{EPE}_j = p_j + \text{MLP}^{(1)}(p_j),$$

a proxy for the positional signal written by the first-layer MLP. We call coordinates satisfying $|\text{EPE}_0[d]| > \mu(|\text{EPE}_0|) + 3\sigma(|\text{EPE}_0|)$ *massive coordinates*, re-identifying them per checkpoint. When evaluating EPE validity, the *MLP-only* comparison uses the net positional contribution of the MLP; the *full-layer* comparison instead uses that of the entire first block, including attention and normalization.

Two interventions recur throughout. REMOVE FIRST PE replaces p_0 by p_1 at position 0. SWAP EPE swaps the MLP-processed positional direction between positions 0 and 1 in the first-layer MLP output. A *manifest* is the fixed set of sampled and truncated evaluation sequences. *Per-example uncertainty* denotes $\pm 2\text{SE}$ over example-level measurements; cross-seed SD, reported separately, measures variation across resampled manifests.

The object under test. For a layer-head pair, let $a_{\ell,h}$ be the attention-score scale specified by the model configuration. Expanding the pre-softmax

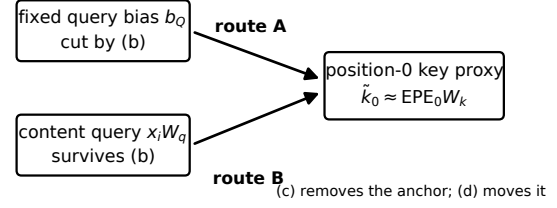


Figure 1: **Two routes into the shared first-position key.** Route A is bias-driven and route B is activation-dependent. Letters denote the interventions in Table 1.

score gives

$$s_{i \rightarrow j}^{\ell,h} = a_{\ell,h} \left[\underbrace{(x_i W_q)(x_j W_k)^\top}_{T1} + \underbrace{(x_i W_q)b_K^\top}_{T2} + \underbrace{b_Q(x_j W_k)^\top}_{T3=\Delta_j} + \underbrace{b_Q b_K^\top}_{T4} \right].$$

For a fixed source i , $T2$ and $T4$ are constant across targets and cancel under the softmax. Only $T1 + T3$ decides which target wins. The original paper isolates $T3$. The source-independent shift $\Delta_j = b_Q W_k^\top x_j^\top$ is unusually large at position 0 because $\text{EPE}_0 = p_0 + \text{MLP}^{(1)}(p_0)$ carries massive coordinates exactly where $b_Q W_k^\top$ is large. We call $T3$ **route A** and $T1$ **route B**, and write $k_0 \approx \text{EPE}_0 W_k$ for the shared first-position key component (Figure 1). Route B is the activation-dependent query route: $(x_i W_q)$ depends on the contextual residual stream rather than isolated token-semantic content.

Related work. Other accounts trace sinks to spectral structure (Cancedda, 2024), variance dynamics (Qiu et al., 2026; Hong and Lee, 2025; Li et al., 2026), geometry (Ruscio et al., 2025), or the anatomy of massive activations (Sun et al., 2026; Shi et al., 2026). Sinks are provably needed for some computations (Ran-Milo, 2026) and can be removed by re-parameterizing attention (Qiu et al., 2025). We follow standard practice for causal interventions (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024; Geiger et al., 2025) and multi-run reproduction (Sellam et al., 2022).

Setup and execution. The direct reproduction uses a fixed 300-sequence manifest, 100 examples each from SST-2 (Socher et al., 2013), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021), with per-example uncertainty. Unless stated otherwise, extensions use three resampled 300-sequence data seeds; maximum cross-seed SD

¹<https://github.com/AlZubayer/SameSinkDiffPlumbing>

Intervention	Paper ($\pm 2\text{SE}$)	Ours ($\pm 2\text{SE}$)
a Baseline	.563 $\pm$.004	.5630 $\pm$.0042
b Nullify b_Q	.251 $\pm$.003	.2514 $\pm$.0032
c Remove first PE	.017 $\pm$.001	.0168 $\pm$.0012
d Swap EPE	.035 $\pm$.002	.0352 $\pm$.0018
e Swap PE (ctrl)	.557 $\pm$.006	.5573 $\pm$.0056
f Zero token-0 emb. (ctrl)	.561 $\pm$.004	.5614 $\pm$.0042
g No MLP	.283 $\pm$.009	.2832 $\pm$.0094
h No PE	.099 $\pm$.015	.0987 $\pm$.0146
i Zero top-3 W_k	.367 $\pm$.004	.3672 $\pm$.0038
j Zero random W_k (ctrl)	.563 $\pm$.004	.5633 $\pm$.0042

Table 1: **Reproduction of the original Table 1** (GPT-2-small, $n=300$, layers 4–11; per-example $\pm 2\text{SE}$). The baseline agrees with the original result at the reported precision.

is 2.18 pp. We retain examples of at least 40 tokens and truncate to 40. Tokenization sets `add_special_tokens=False`; position 0 is the first content token. The metric averages attention from second-half sources to position 0 over heads and a mid-layer band. Table 1 defines baseline (a) and nine interventions (b)–(j); outside GPT-2, directly corresponding edits are used where possible and explicitly labeled route-level analogues elsewhere. Appendix A gives layer bands, coordinate selection, analogue definitions, HumanEval overlap, and seven-seed stability.

We implemented R1, a hand-written weight-level forward pass, and R2, Hugging Face eager attention under NNsight (Fiotto-Kaufman et al., 2025) with traced activation edits. All 1,157 tested GPT-2 quantities pass parity. R2 exposed an R1 attention-scaling error on CRFM checkpoints; after correction, 68 of 68 gate quantities pass and 765 paired measurements differ by at most 2.4×10^{-7} . Emergence was rerun on both paths. R2 revalidates residual analysis at 124M; 355M and 774M remain corrected R1 results. Large OPT and Qwen checkpoints use bf16. Appendix H gives complete coverage and executor-dtype confounds.

The intervention effects reproduce. The baseline and all nine intervention and control rows reproduce (Table 1). Our .5630 baseline falls within the reported $.563 \pm .004$ interval, and cross-entropy matches the Hugging Face reference within 5×10^{-8} nats.

All four structural analyses also reproduce, with two qualifications the original figures could not show. First, the pooled separation of Δ_0 reproduces, but a stricter per-layer test does not: requiring the position-0 tenth percentile to exceed every other position’s ninetieth percentile passes in only 3 of the 8 evaluated band layers (Table 4, strict fraction .375). The claim that Δ_0 is system-

	124M	355M	774M
baseline pos.-0	.564	.575	.501
after Nullify b_Q (%)	44.5	70.4	74.0
Δ -score share of route A	.561	.450	.247
Nullify b_Q + Remove First PE (%)	2.25	1.64	3.65
Nullify b_Q + Swap EPE (%)	4.23	3.47	2.17
Nullify b_Q + Zero Massive W_k (%)	24.8	19.1	31.2
Swap-EPE relocation ratio	.897	.922	.944
ΔCE of Nullify b_Q (nats)	+1.59	+0.36	+0.12

Table 2: **Residual-sink analysis across GPT-2 scale** (3 seeds; $\text{SD} \leq .23$ pp on intervention rows; % of own baseline).

atically larger *when pooled* is supported; complete layer-by-layer separation is not. Second, the EPE abstraction is excellent for the part it was defined on, with median cosine to the net MLP-only positional signal of at least .8218 at every position and .9962 at position 0. It also matches the original’s full-first-layer check at position 0 (.7569 against the reported .76). Away from position 0 the full-layer proxy is weak, with medians between .30 and .34. The original restricted that check to position 0, so this extends the claim rather than disagreeing with it.

The remaining two analyses reproduce quantitatively: massive-coordinate $|\gamma|$ is $7.79\times$ the value at other coordinates on average and $8.83\times$ in medians, and the massive coordinates are $\{138, 378, 447\}$, exactly as reported. **Verdict: reproduced.**

3 Following the Residual: One Anchor, Two Routes

The original paper is explicit that its strongest actuator does not finish the job. Nullifying b_Q leaves 44.7% of the sink standing, and the paper lists this residual as an open limitation. We start there because the residual reorganizes the picture.

Because only $T1$ and $T3$ survive the softmax, two query-side terms address the same key k_0 : route A is the fixed bias term $b_Q \cdot k_0$, the original circuit, and route B is the activation-dependent term $(x_i W_q) \cdot k_0$. We test whether the residual follows route B at three GPT-2 scales. R2 revalidates 124M; the validated R1 executor evaluates 355M and 774M.

The residual is a second route into the same key. Four observations point to one shared object (Table 2). Removing the first-position identity on its own leaves 1.7–3.9% at every scale. Removing only the bias leaves 44.5, 70.4, and 74.0%. Removing both together collapses the residual to 2.25, 1.64, and 3.65%. So whatever survives bias removal still needs the first-position object.

The transplant gives complementary evidence. Under Swap EPE the sink does not vanish, it *moves*: between .897 and .944 of the baseline mass reappears at the transplant target. After nullifying b_Q , between .260 and .360 absolute attention still follows the transplanted key. Route B tracks the key wherever the key goes.

These are softmax counterfactuals and do not add linearly. Combining all three components can leave slightly more residual than combining two, and a floor of roughly 2–6% survives everything we tried. Within this decomposition, the surviving attention is primarily explained by two query-side terms addressing the same key.

Graded causality. The original paper used on-off edits. We add dose-response tests. Scaling b_Q and scaling the massive W_k columns are perfectly monotone at all three scales (Spearman $\rho=1$, $p=4.0 \times 10^{-4}$ ($n=7$)). Interpolating the positional *identity* from p_0 toward p_1 behaves differently, akin to a threshold. At $\alpha \leq .25$ the sink is 2–4% of baseline, and at $\alpha \geq .5$ it is back at baseline.

Deleting p_0 is not the same as replacing it. At 124M, deletion *strengthens* the sink to 110.0%, and at 774M it collapses it to 36.6%. The sink is sensitive to positional identity, and deletion is not equivalent to replacement. This is consistent with sinks appearing in models trained without positional encodings (Gu et al., 2025), though positional presence clearly matters more at larger scale.

With scale, route A fades and route B specializes. The residual after nullifying b_Q grows from 44.5 to 74.0%. Independently, and without any softmax, the Δ -score share of route A falls from .561 to .247. Two different measurements agree that the bias route matters less at larger scale.

Route B does not simply take over. It concentrates. The mean alignment between the activation-dependent query and k_0 collapses from +.228 to +.023 to −.061, yet it stays enriched over matched controls at +.023, −.070, and −.108 (Cohen’s d of 1.79, .42, .30). The head-cell distribution explains this. Cells with alignment above .1 fall from 89 to 18%, while the strongest single cell rises from .465 to .719. A pooled mean understates a mechanism that lives in a few carrier heads, so the original alignment claim generalizes as *relative* enrichment rather than positive absolute cosine.

The functional result is also unexpected. Removing b_Q becomes almost free with scale, costing +1.59, +0.36, and +0.12 nats, even as it re-

	b	c	d	e	f	g	h	i	j	base
GPT-2 124M	44	3	6	99	100	50	18	65	100	0.56
GPT-2 355M	70	2	8	100	100	14	9	28	100	0.58
GPT-2 774M	24	4	3	100	100	21	3	41	100	0.50
Neo 125M	n/a	6	30	100	93	4	51	98	101	0.27
Neo 1.3B	n/a	2	80	100	40	7	26	88	100	0.33
Neo 2.7B	n/a	2	25	100	80	16	19	66	100	0.41
OPT 125M	83	2	61	100	18	4	71	100	0.68	0.68
OPT 1.3B	98	32	0	4	100	24	5	38	100	0.63
OPT 2.7B	98	36	13	34	100	18	8	39	100	0.62
OPT 6.7B	99	6	3	53	100	12	3	56	100	0.63
OPT 13B	99	7	94	98	99	50	59	112	98	0.34
Qwen 0.5B	5	100*	96*	96*	96	6	81	100*	100	0.49
Qwen 1.5B	7	95*	83*	83*	1	5	82	99*	100	0.50
Qwen 3B	12	100*	95*	95*	66	3	92	100*	100	0.48
Qwen 7B	6	99*	96*	96*	1	7	92	97*	100	0.53
Qwen 14B	42	100*	100*	100*	4	6	96	100*	100	0.64

Residual position-0 attention (% of each checkpoint baseline); lower is stronger removal

Figure 2: **Intervention fingerprints on 16 checkpoints** (R2; 3 seeds, $SD \leq 2.18$ pp; cells show one-decimal % of baseline surviving; darker means greater removal; right: absolute baseline). “n/a” marks GPT-Neo’s missing b_Q , ° Qwen analogues, and * bf16 checkpoints. OPT-13B (i) is the sole increase above 110%. No profile over mutually available conditions spans all four families.

moves less of the sink. The bias route becomes less causally dominant with scale, so bias-only mitigation becomes cheaper but leaves most of the sink intact.

4 Related Anchoring Functions Across Architectures

We next ask how other architectures support concentrated first-position attention. All tested models exhibit anchoring, but its supporting objects and routes differ. We evaluate baseline plus nine intervention labels across 16 checkpoints: GPT-2 (124M to 774M), GPT-Neo (125M to 2.7B; Black et al., 2021), OPT (125M to 13B; Zhang et al., 2022), and Qwen2.5 (0.5B to 14B; Qwen Team, 2024). Each applicable residual is classified as strong reduction ($< 50\%$), partial reduction (50% to $< 90\%$), negligible change (90% to 110%), or increase ($> 110\%$). The ordered categories define a profile over available conditions. Crucially, outside GPT-2—especially for RoPE-based Qwen2.5—shared intervention labels denote route-level architectural analogues, not identical tensor edits or identical causal objects; missing edits remain unavailable. No such profile spans all four families. Figure 2 and Appendix B give the results.

Every checkpoint has a sink, with baselines from .271 to .677. Beyond that the two pillars of the GPT-2 circuit come apart. GPT-2 uses both routes, with route A weakening as scale grows, and both Remove First PE and Swap EPE collapse its sink at every scale. Qwen2.5 is carried by its query bias: nullifying it leaves 5.3–12.2% at 0.5B to 7B and 41.8% at 14B, while the positional analogues are mostly inert. OPT is the mirror image through 6.7B, where nullifying the query bias leaves 83–99% but zeroing the massive key columns leaves 38–56%.

GPT-Neo has no query bias at all, so (b) can do nothing, yet its sink is still positionally keyed and Remove First PE leaves only 2.2–5.9%.

Controls are also model-dependent. Zeroing the token-0 embedding is nearly inert in GPT-2 and OPT, but leaves 40.1% in Neo-1.3B and 0.9–3.5% in Qwen 1.5B, 7B, and 14B, showing token-driven sinks within families. The raw-PE swap leaves 4.0–53.1% in mid-size OPT, and one fixed random- W_k draw raises the sink by 3.47% although the resampled control stays near 100%. Control validity must therefore be established per architecture and distribution. No-MLP reduces the sink to 3.4–50.2%, strongly in 14 of 16 checkpoints.

One checkpoint stands apart. In OPT-13B, zeroing the massive key columns *increases* the sink to 111.9%, and the effect is stable across seeds. A single-sentence probe suggests a two-position anchor that our position-0 metric under-measures. We treat this as an open observation and describe the checks it needs in Appendix D.

The original structural diagnostics do not reliably identify causal importance outside GPT-2. This raises a question. If the routes differ per family, do the original paper’s *structural* diagnostics still tell us where the sink is? We computed per-model analogues of all of them (Appendix C). The diagnostics produce strong signals in several models, but those signals dissociate from causal importance in both directions.

First, a signature can appear without causal dominance. Inside GPT-2, the Δ_0 diagnostics show different scale trends: relative median separation increases, whereas strict layer coverage rises from .375 to .950 and then recedes to .844; meanwhile, the causal share of b_Q roughly halves, from 55.5 to 26.0%. OPT 1.3B to 6.7B separates reliably and shows the strongest coordinate co-adaptation in the sweep at $10.1\times$, yet b_Q removes only 1–2% of its sink there. Its absolute separation is roughly ten times smaller than GPT-2’s, which is too small to matter once it passes through a softmax. Thus OPT can show strong route-A geometry while bias removal changes little of the sink. The five Mistral runs of §5 converge to a bias-key cosine of .795 while nullifying b_Q removes 7.7%.

Second, causation can appear without the expected signature. At OPT-6.7B the MLP-only EPE proxy collapses to a cosine of .05, and yet Swap EPE still leaves only 2.6% of the sink. The lever works even where the abstraction that motivated it

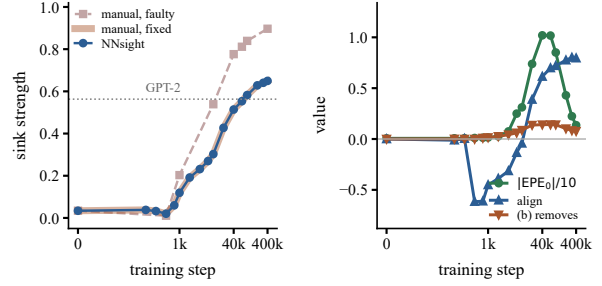


Figure 3: **Circuit emergence over five independent pre-training runs** (17 checkpoints; run means). The corrected R2 trajectory is shown with the OpenAI GPT-2-small reference. *Right*: $\max |EPE_0|/10$, bias-key alignment, and the sink share removed by b_Q . EPE peaks near 40k, while route A peaks near 70k and then nearly halves.

does not.

This is not a criticism of the original paper, which did intervene rather than rely on its figures (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019). It is a caution about a class of diagnostic. A high cosine does not mean load-bearing, and clean separation does not mean dominance. Structural signatures identify correlates of the sink computation, not the causal load carried by the corresponding components.

5 The Same Anchor Across Training Runs

Architecture is one way to vary the wiring. Training is another. We now ask whether the same architecture, trained again from a different initialization, rebuilds the same circuit. We use five independently pre-trained GPT-2-small runs from Stanford CRFM Mistral, 17 checkpoints each, and all 85 evaluations completed on the same fixed 300-example set.

Under R2 the converged sink is $.650 \pm .005$, moderately above OpenAI GPT-2-small at .563. Bias-key alignment reaches $.795 \pm .027$ and query alignment $.304 \pm .021$. The isolated activation-dependent route carries $.552 \pm .004$ attention while delta-only attention is only $.059 \pm .001$. All five runs recover the anchoring behaviour, and their converged sink is mainly carried by route B. Both execution paths agree on the converged values reported here to within 2.4×10^{-7} , but only after we corrected a scaling fault in the first path that had inflated the converged sink by 38% (Appendix F). Three findings follow (Figure 3):

Assembly is not monotone. At step 400 route A is strongly *anti*-aligned at $-.616 \pm .155$, and the sink falls from .034 to .020. After that it grows at

every stored checkpoint. The maximum $|EPE_0|$ peaks near 10.20 around 40k to 70k steps and then relaxes about $7.5\times$ to 1.36, while delta-only attention peaks near .079 and settles at .059. A mid-training checkpoint would give a different mechanistic reading than the final model.

The anchoring behaviour recurs across all five tested runs. The five final top-3 coordinate sets are pairwise disjoint, and none contains $\{138, 378, 447\}$. We do not read the disjointness as evidence: under a uniform null over the 768 dimensions, five random 3-sets are all pairwise disjoint with probability .889 (Appendix F), so zero Jaccard is the expected outcome, not a surprising one. The supported claim does not depend on that statistic. Across the five runs, exact coordinate identities were not stable: no run reproduces the published coordinates, no two runs agree, and three of the five do not lock their final labels until the last checkpoint at 400k. Methods that hard-code these coordinates should not be assumed to transfer across independently trained runs.

Route A’s signal rises while its causal share falls. Removing the first PE removes $98.5 \pm .4\%$ of the sink. Zeroing the massive W_k columns removes $42.0 \pm 11.7\%$. Nullifying b_Q removes only $7.7 \pm .5\%$, in runs whose bias-key alignment is .795. The trajectory is sharper than the endpoint. Between 70k and 400k steps the alignment climbs from .698 to .795, while the share of the sink that b_Q actually carries falls from 14.2% to 7.7%, a drop of nearly half (Figure 3). Signal and causation move in opposite directions in the same runs over the same checkpoints. This repeats the dissociation in §4, now inside one architecture and along training time rather than across model families. It is also a third axis of variation: route A weakens with GPT-2 scale (§3), varies substantially across architectures (§4), and loses causal share during training.

We pre-specified an ordering test asking whether alignment rises before the sink does. It is directionally consistent but not significant. Alignment precedes the sink in all 3 non-tied runs (2 ties dropped; Wilcoxon $W=0$, $p=.25$). We do not build on the descriptive timing.

6 What the Anchor Is Worth

We stress-test GPT-2-small across metrics, length, and content, then measure each intervention’s

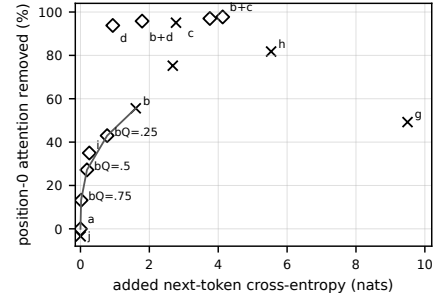


Figure 4: **Sink reduction versus LM cost** (GPT-2-small, 40 tokens).

language-model cost as added mean next-token cross-entropy.

Metrics mostly agree. Six sink metrics rank the ten interventions similarly (mean Kendall $\tau = .84$); all five pairs below .8 involve attention entropy, which measures concentration anywhere rather than specifically at position 0 (Hong and Lee, 2025). Swap EPE *relocates* the anchor, moving .552 of the mass to positions 1–4 at entropy .488; Remove First PE *diffuses* it, sending .930 to later positions at entropy .607. Position-0 attention cannot distinguish relocation from destruction.

Length. Using within-domain concatenations, we extend context from 40 to 1024 tokens. Despite a $25.6\times$ increase, the baseline sink falls only 23.7%, from .5625 to .4294, and 90.7% of head-cells still exceed .2. The key objects remain fixed to fp32 rounding (relative discrepancy $< 2 \times 10^{-6}$). Fair-share scaling predicts a logit slope of -1 , but the across-length slope is $-.166$ and within-length slopes flatten from $-.336$ to $-.107$; fixed logits therefore do not imply exchangeable competitors. Effects also vary with length: Remove First PE leaves 0.08% at 1024 tokens, Swap EPE is non-monotone, and (b) spans 30.5–57.9% by domain (Appendix G).

Content. Across seven regimes, the anchor’s *address* is stable: Remove First PE leaves 2.1–4.3% and Swap EPE 5.1–10.0%. Its *traffic* varies: after nullifying b_Q , residual attention rises from 43.1–45.5% in natural domains (44.4% pooled) to 55.1% on uniform and 57.5% on repeated tokens. We report effect sizes because paired tests ignore seed clustering and HumanEval overlap.

Relocation is cheaper than removal. Swap EPE removes 93.8% of position-0 mass for +.938 nats, while Remove First PE removes 97.0% for +3.757 nats, about four times the cost. Scaling b_Q to .75

and .5 removes 13.2 and 27.2% for +.019 and +.188 nats. No-MLP costs +9.497 nats to remove 49.2%.

Swap EPE, Zero Top- k W_k , $(b \wedge d)$, and b_Q scales .25 and .75 are Pareto-efficient in all nine seed-by-domain strata; No-MLP, No-PE, full bias removal, and the random control in none. Damage is also route-specific (Appendix G).

The ordering suggests that the model benefits from a concentrated anchor but not specifically at position 0, consistent with functional accounts of first-token attention (Barbero et al., 2025; Ran-Milo, 2026; de Llano et al., 2026). Mitigation should report where attention moves and what the model loses, not only position-0 reduction (Yu et al., 2024; Qiu et al., 2025).

7 Discussion

Generalization is hierarchical. The *phenomenon* is robust: baselines span .27–.68 across 16 checkpoints and remain .43 at $25.6\times$ context. The best abstraction is an *anchoring function*: concentrated first-position attention supported by architecture-specific objects and routes (Geiger et al., 2025). These routes recur, but their causal weights vary across architecture, scale, content, and run, and coordinates do not transfer. The computation is robust; its parameter-level implementation, structural signatures, and controls are not.

This refines Ran-Milo et al. (2026): the intervention effects reproduce, and 16 checkpoints in four families plus five GPT-2-small runs directly support the distinct-circuit conclusion. Their circuit is one instance of the broader anchoring computation.

Transfer follows the wiring: bias editing is strong in Qwen but generally weak in OPT, except at 125M; the massive-key edit works in GPT-2 and most OPT checkpoints, fails in Qwen analogue form, and reverses at OPT-13B. Shared labels do not license transfer. Pooled statistics can hide carrier heads; one draw is not a null; alignment need not imply causal load; controls require per-model validation.

Limitations

Sections 3, 4, and 6 vary data seeds rather than model-training seeds; only Section 5 varies training runs, and most experiments use three data seeds, so SD estimates remain noisy. Cross-model comparisons use route-level causal analogues rather than identical operations for Qwen’s (c), (d), (e), (h), (i) and OPT’s (e), so identical labels should not be interpreted as identical interventions. The 1024-token evaluation uses within-domain concatenations rather than natural documents, and scaled layer bands average more layers in deeper models. The Mistral checkpoints converge to a moderately stronger sink than OpenAI GPT-2-small, so Section 5 should be interpreted as a study of learning dynamics rather than a training replication; the pre-specified emergence ordering test is negative. R2 revalidates the residual-sink analysis only at 124M, leaving the larger GPT-2 scales on R1; for large OPT, executor and dtype changes are confounded, and Qwen lacks a family-specific parity gate, so observed differences cannot be causally attributed. The OPT-13B two-position result is a single-sentence diagnostic, intervention (i) uses three columns at 124M but six at larger GPT-2 scales, confounding its cross-scale trend, and we did not evaluate GPT-2-XL, multilingual models, or a denser interpolation grid between $\alpha = .25$ and $\alpha = .5$.

Ethical Statement

This work is a good-faith reproducibility study. We represented the original paper faithfully, followed its experimental setup wherever possible, clearly separated reproduction from extension, and reported confirmations, discrepancies, negative results, and implementation errors transparently. Our experiments use publicly available pretrained models and benchmark datasets and involve no new human participants or private data. While the studied interventions may support model analysis and efficiency, they may also be misapplied; our results therefore do not imply that attention-sink mitigation is universally safe or beneficial, and interventions should be revalidated for each target model and use case.

References

- Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. 2025. [Why do LLMs attend to the first token?](#) In *Second Conference on Language Modeling*.
- Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow](#).
- Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Asya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. 2021. [Accounting for variance in machine learning benchmarks](#). In *Proceedings of Machine Learning and Systems*, volume 3, pages 747–769.
- Nicola Cancedda. 2024. [Spectral filters, dark signals, and attention sinks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, et al. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- 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. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, et al. 2022. [Underspecification presents challenges for credibility in modern machine learning](#). *Journal of Machine Learning Research*, 23(226):1–61.
- Enrique Queipo de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael M. Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. 2026. [Attention sinks and compression valleys in LLMs are two sides of the same coin](#). In *The Fourteenth International Conference on Learning Representations*.
- Jaden Fiotto-Kaufman, Alexander Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. 2025. [Nnsight and ndif: Democratizing access to open-weight foundation model internals](#). In *International Conference on Learning Representations*, volume 2025, pages 92337–92370.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. [Causal abstraction: A theoretical foundation for mechanistic interpretability](#). *Journal of Machine Learning Research*, 26(83):1–64.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. [When attention sink emerges in language models: An empirical view](#). In *International Conference on Learning Representations*.
- Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael Jordan, and Song Mei. 2024. [Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs](#). In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.
- Stefan Heimersheim and Neel Nanda. 2024. [How to use and interpret activation patching](#). *arXiv preprint arXiv:2404.15255*.
- Jonghyun Hong and Sungyoon Lee. 2025. [Variance sensitivity induces attention entropy collapse and instability in transformers](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8360–8378. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556. Association for Computational Linguistics.
- Siquan Li, Kaiqi Jiang, Jiacheng Sun, and Tianyang Hu. 2026. [The structural origin of attention sink: Variance discrepancy, super neurons, and dimension disparity](#). In *Forty-third International Conference on Machine Learning*.
- Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. 2024. Duquant: distributing outliers via dual transformation makes stronger quantized llms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. 2021. [Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program](#). *Journal of Machine Learning Research*, 22(164):1–20.

- 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. 2026. [A unified view of attention and residual sinks: Outlier-driven rescaling is essential for large language model training](#). In *Third Conference on Language Modeling*.
- 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. 2025. [Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). Technical report, OpenAI.
- Yuval Ran-Milo. 2026. [Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 70–89, San Diego, California, United States. Association for Computational Linguistics.
- Yuval Ran-Milo, Hila Ofek, and Shahar Mendel. 2026. [A mechanistic account of attention sinks in GPT-2: One circuit, broader implications for mitigation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–98, San Diego, California, United States. Association for Computational Linguistics.
- Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. 2025. [What are you sinking? a geometric approach on attention sink](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D’Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. [The MultiBERTs: BERT reproductions for robustness analysis](#). In *International Conference on Learning Representations*.
- Zeru Shi, Zhenting Wang, Fan Yang, Qifan Wang, and Ruixiang Tang. 2026. [A single layer to explain them all: Understanding massive activations in large language models](#). *arXiv preprint arXiv:2605.08504*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.
- Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). In *Conference on Language Modeling*.
- Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. 2026. [The spike, the sparse and the sink: Anatomy of massive activations and attention sinks](#). *arXiv preprint arXiv:2603.05498*.
- Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. [Llm circuit analyses are consistent across training and scale](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 40699–40731. Curran Associates, Inc.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Sarah Wiegreffe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 11–20. Association for Computational Linguistics.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. [Efficient streaming language models with attention sinks](#). In *International Conference on Learning Representations*.
- Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. 2024. [Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *International Conference on Learning Representations*, volume 2024, pages 1651–1678.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [OPT: Open pre-trained transformer language models](#). *arXiv preprint arXiv:2205.01068*.

A Experimental Setup

Data. With `add_special_tokens=False`, position 0 is the first content token. Each resampling seed draws 300 sequences, 100 each from SST-2 (Socher et al., 2013), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021). We retain examples of at least 40 tokens and truncate to 40. Artifact manifests are byte-identical across R1 and R2. Because HumanEval’s finite pool creates about 60/100 overlap per seed, compared with 9–11 for SST-2 and 0–3 for GSM8K, the three seeds measure sampling stability rather than independent replication.

Metric and layer band. The sink metric averages attention from second-half source tokens to position 0 over heads and a mid-layer band. We use layers 4–11 for 12-layer models, matching the original paper, and $[3, L-1)$ otherwise. This depth-aware band averages more layers in deeper models and is not a fixed-layer comparison.

Interventions. Labels follow the original paper: (a) baseline, (b) nullify the query bias, (c) replace the first positional encoding with the second, (d) swap the MLP-processed positional direction between positions 0 and 1, (e) the same swap on the raw positional embedding, (f) zero the first token embedding before positional signals are added, (g) skip all MLP blocks, (h) remove positional information, (i) zero the key-projection columns at the massive EPE_0 coordinates, and (j) zero the same number of random columns.

Intervention (i) is re-identified per model rather than copied from GPT-2-small. We select the coordinates of EPE_0 whose absolute value exceeds the mean by at least three standard deviations, which yields 3 columns for GPT-2-small and between 3 and 12 for the other checkpoints. Each run records the selected coordinates in its configuration file. Control (j) always zeros the same number of columns, resampled per layer with seed 42, so it is size-matched at every scale.

Cross-family analogues. Outside GPT-2 some interventions cannot be the same tensor edit. GPT-Neo has no query bias, so (b) is a structural no-op and returns the baseline exactly. OPT applies its +2 positional offset. Qwen2.5 has no additive positional encoding, so (c), (d), (e), and (h) act on RoPE position ids, (d) and (e) become the same operation by construction, and (i) uses a per-sentence token-embedding stand-in rather than a fixed EPE object.

Equal labels therefore denote route-level analogues, not identical operations, and cross-family equality of a label is not equality of a causal object.

Sampling stability. A separate seven-seed study on GPT-2-small, using 300 fresh examples per seed, gives cross-seed SDs of .15, .09, .08, and .10 pp for the surgical edits (b), (c), (d), and (i), against .65 and 1.10 pp for the coarse edits (g) and (h). The coarse edits also carry the largest domain gaps, for example 41.6/52.4/57.2% for (g) on SST-2, GSM8K, and HumanEval. Surgical edits are therefore roughly an order of magnitude more sampling-stable than coarse ablations, which is one reason we treat coarse ablations as weak mechanistic evidence.

B Full Cross-Model Results

Table 3 gives the 16-model R2 sweep as percentages of each model’s baseline. Cross-seed SD is at most 2.18 pp, including $\leq .58$ for OPT-13B and $\leq .25$ for Qwen-14B; baseline SEs are .001–.006. Appendix A describes coordinate selection and the size-matched control. The $j-i$ specificity gap is +35/ + 72/ + 58 points for GPT-2, +60/ + 62/ + 61/ + 44/−13 for OPT, +3/ + 12/ + 34 for GPT-Neo, and about 0 for Qwen. Released domain splits show the largest sensitivity for coarse edits.

C Structural Signatures Across Models

From the raw per-model exports we computed analogues of the original paper’s structural diagnostics. Qwen is excluded where the object is undefined, since it has no additive EPE , and is recorded as not applicable in the artifact. Table 4 consolidates the results. Five readings follow. First, within GPT-2 the Δ_0 diagnostics show different scale trends: relative median separation increases with scale, whereas strict layer coverage rises from .375 to .950 and then recedes to .844, while the causal b_Q share falls. Second, OPT 1.3B to 6.7B separates reliably but with absolute gaps roughly ten times smaller than GPT-2’s, even though (b) changes little of the sink there. This shows that geometric co-adaptation need not imply substantial causal load. Third, OPT-13B has a median ratio of 1.01, and position 0 ranks first in 51.7% of head-cells. We treat this result as inconclusive because the corresponding chance level depends on the exact ranking statistic. Fourth, coordinate co-adaptation is present in every additive-PE model that has a b_Q ,

Model	base	b	c	d	e	f	g	h	i	j
GPT-2 124M	.5637	44.5	3.0	6.2	98.9	99.7	50.1	18.0	65.2	100.1
GPT-2 355M	.5753	70.4	1.7	8.3	100.1	100.1	13.8	9.3	28.2	100.2
GPT-2 774M	.5015	74.0	3.9	3.4	100.0	100.5	20.6	3.2	42.1	99.9
Neo 125M	.2714	100.0 [†]	5.9	30.3	99.5	92.7	3.8	50.6	97.7	100.7
Neo 1.3B	.3336	100.0 [†]	2.4	79.9	99.9	40.1	6.6	26.3	88.4	100.1
Neo 2.7B	.4121	100.0 [†]	2.2	25.2	100.0	79.6	16.1	19.4	66.3	100.0
OPT 125M	.6766	82.8	2.1	60.6	46.3 [‡]	99.8	17.8	3.5	39.7	99.8
OPT 1.3B	.6268	97.7	32.2	0.5	4.0 [‡]	99.6	23.9	4.6	37.7	100.0
OPT 2.7B*	.6222	98.2	36.5	13.3	33.5 [‡]	99.8	17.9	7.5	38.7	99.8
OPT 6.7B*	.6307	98.7	5.7	2.6	53.1 [‡]	99.7	11.7	2.9	55.6	99.9
OPT 13B*	.3430	98.8	6.8	93.5	98.5	99.2	50.2	58.6	111.9	98.3
Qwen 0.5B	.4926	5.3	99.6°	96.3°	96.3°	96.5	5.7	81.3	99.6°	99.5
Qwen 1.5B	.4999	6.9	95.1°	83.1°	83.1°	1.3	5.3	82.3	99.2°	100.0
Qwen 3B	.4817	12.2	99.7°	94.7°	94.7°	66.5	3.4	92.3	99.9°	99.9
Qwen 7B*	.5348	6.1	99.4°	96.5°	96.5°	0.9	6.7	91.9	97.2°	100.0
Qwen 14B*	.6366	41.8	100.0°	99.9°	99.9°	3.5	5.9	96.4	99.7°	99.5

Table 3: Complete cross-model sweep (% of own baseline; “base” is absolute position-0 attention over the scaled band $[3, L-1]$). [†]: structural no-op, since there is no b_Q . [‡]: the raw-PE swap is not a neutral control in mid-size OPT. °: RoPE position-id analogues, where (d) equals (e) by construction and (i) uses a per-sentence token-embedding stand-in. *: bf16; other rows are fp32. Relative percentages are more comparable across rows than absolute bf16 baselines.

including OPT-1.3B at $10.1\times$ where (b) is causally inert. Fifth, the EPE abstraction is architecture-dependent and scale-dependent. It is sound in GPT-2 at all scales, MLP-only-valid in GPT-Neo, whose full first block reverses the net positional direction at position 0, and it degrades with OPT scale to .05 at 6.7B. Lever potency does not track proxy fidelity, since Swap EPE still leaves only 2.6% at OPT-6.7B.

D The OPT-13B Anomaly

OPT-13B is the sweep’s qualitative outlier. Under R2 its baseline is .343; Swap EPE, No-PE, and Remove First PE leave 93.5%, 58.6%, and 6.8%. Zeroing massive W_k raises the sink to a seed-stable 111.9%. The single-sentence R1 probe in Table 5 is diagnostic only and suggests a two-position anchor.

The pair carries nearly equal mass, at .755 combined, so a position-0-only metric under-measures the anchor. Swap EPE only re-balances mass *within* the pair and leaves the total unchanged, which would mechanically explain why (d) leaves 93%. Remove First PE destroys both, taking the total to .038, so the pair appears jointly keyed to position-0 identity. Zeroing massive W_k raises total pair mass to .800, with a mid-band gain and a late-layer collapse from .42 to .10 in layers 31–39. We emphasise that this is one sentence and head-averaged, and the dataset-level metric remains authoritative. Three checks would settle it: a position-0-plus-1 metric over the existing

sweep, a fixed-band $[3, 11]$ re-run to remove the deep-band confound, and a threshold sweep over $n_\sigma \in \{2.5, 3, 3.5\}$.

E Residual-Sink Anatomy: Details

Decomposition values are separate softmax counterfactuals and do not add linearly. With only route B active, attention to position 0 is .210/.329/.323 at 124M/355M/774M. With only route A active it is .377/.216/.136. The Δ -score share of the position-0 advantage is .561/.450/.247. Query-to- k_0 alignment means are $+.228/+ .023/- .061$ against controls at $+.023/- .070/- .108$, with Cohen’s d of 1.79/.42/.30 and Mann–Whitney $p \leq 1.5 \times 10^{-22}$. Head-cells with alignment above .1 fall from 89 to 48 to 18% while the maximum rises from .465 to .521 to .719. This is a distributed-to-carrier-head transition, and it is correlational at the head level; ablating the carrier heads is the natural causal sequel.

Combined interventions, as % of baseline with cross-seed SD, are: $b \wedge c$ $2.25 \pm .06 / 1.64 \pm .20 / 3.65 \pm .14$; $b \wedge d$ $4.23 \pm .09 / 3.47 \pm .03 / 2.17 \pm .08$; $b \wedge i$ $24.8 \pm .10 / 19.1 \pm .07 / 31.2 \pm .23$; $b \wedge i \wedge d$ $4.97 \pm .05 / 5.58 \pm .08 / 2.47 \pm .06$. The last row exceeds $b \wedge d$ at 124M, 355M and 774M which is the non-additivity noted in §3. Relocation: attention to the swap target under (d) is .512/.537/.483 against baselines of .006/.007/.010, giving ratios $.8965 \pm .0040 / .9215 \pm .0006 / .9441 \pm .0007$; after (b), .262/.260/.360 absolute remains at the target.

Model	Δ_0 separation (original Fig. 1 analogue)			bias-key alignment (original Fig. 2 analogue)			EPE validity (original Fig. 3 analogue)			causal
	strict frac	p50 ratio	p50 gap	pos-0 mean	gap	frac max	MLP-only	full layer	b removes	
GPT-2 124M	.375	3.20	19.27	+.700	1.082	.990	.996	.757	55.5%	
GPT-2 355M	.950	3.58	12.96	+.710	.983	.994	.970	.932	29.6%	
GPT-2 774M	.844	3.83	9.54	+.051	.503	.753	.991	.859	26.0%	
Neo 125M	0	N/A	0	0 [†]	0	N/A	.997	-.757	0 [†]	
Neo 1.3B	0	N/A	0	0 [†]	0	N/A	.925	-.646	0 [†]	
Neo 2.7B	0	N/A	0	0 [†]	0	N/A	.989	-.749	0 [†]	
OPT 125M	1.000	6.09	12.70	+.864	1.498	1.000	1.000	1.000	17.2%	
OPT 1.3B	.750	3.88	1.57	-.285	.192	.766	.867	.851	2.3%	
OPT 2.7B	.750	3.68	1.32	-.314	.252	.875	.805	.875	1.8%	
OPT 6.7B	.679	3.59	1.14	-.291	.149	.779	.052	.570	1.3%	
OPT 13B	.000	1.01	0.01	-.547	.023	.517	.408	.404	1.2%	
Mistral ($\times 5$)	N/A	N/A	N/A	+.795	N/A	N/A	N/A	N/A	7.7%	

Table 4: Per-model structural signatures over each model’s mid band. “Strict frac” is the layer fraction whose position-0 tenth percentile exceeds every other position’s ninetieth percentile for min-recentred Δ ; “p50 ratio/gap” compares position 0 with the largest other-position median. “Gap” is the position-0 minus other-position mean cosine, and “frac max” the head-cell fraction ranking position 0 first. EPE validity is the position-0 median cosine to the net positional signal. “b removes” is the sink share removed by nullifying b_Q . [†]: zero because b_Q is absent. Massive-to-other $|b_Q W_k^\top|$ ratios are $7.79/7.62/3.84\times$ for GPT-2 and $7.21/10.09/7.94/6.96/3.79\times$ for OPT.

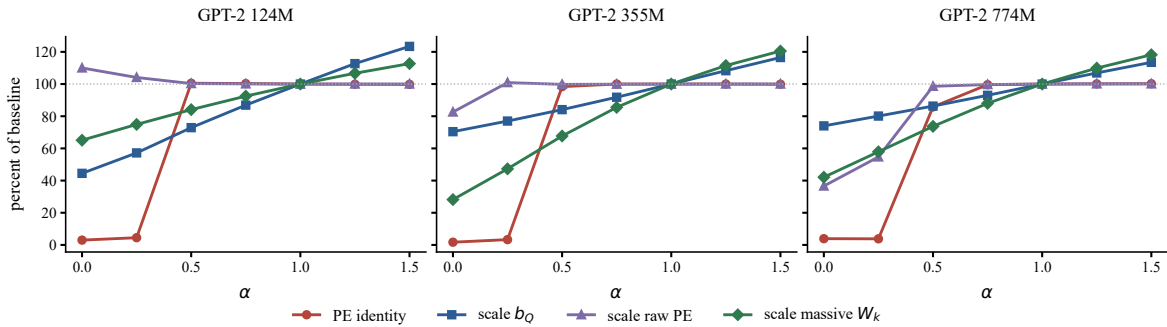


Figure 5: **Dose response across GPT-2 scale** (α grid $\{0, .25, .5, .75, 1, 1.25, 1.5\}$, % of each model’s baseline; cross-seed SDs are smaller than the markers). Scaling b_Q and scaling the massive W_k columns are perfectly monotone. PE-identity interpolation is threshold-like between $\alpha = .25$ and $\alpha = .5$. Raw-PE deletion flips sign with scale.

int.	pos. 0	pos. 1	pos. 0+1
a	.389	.366	.755
b	.385	.364	.749
c	.003	.035	.038
d	.355	.400	.755
h	.133	.217	.350
i	.423	.377	.800

Table 5: OPT-13B single-sentence probe (head-averaged attention mass).

Surgical against coarse edits: skipping only the first-layer MLP leaves 102.5/**6.9**/87.7%, while the all-layer No-MLP leaves 50.1/13.8/20.6%. The first-layer skip nonetheless costs +2.42/ + 4.54/ + 4.21 nats. This mechanism-versus-function dissociation is further evidence of per-checkpoint reorganization. Cross-entropy baselines are 4.009/3.727/3.539 nats, equal to the Hugging Face reference to under 5×10^{-8} . Δ CE for nullify- b_Q is +1.586/ + 0.355/ + 0.120, and for zero-massive- W_k it is +0.259/ + 1.058/ + 0.472. Note that the massive-coordinate count differs across scales, at $k=3$ against $k=6$, which confounds the cross-scale trend for (i) at 65.2/28.2/42.1%.

run	sink	align.	top-3 coords
alias	.6449	.812	683, 8, 541
battlestar	.6546	.824	618, 318, 341
caprica	.6515	.805	115, 229, 439
darkmatter	.6439	.762	610, 470, 192
expanse	.6530	.773	529, 233, 41
paper	.563	(pos.)	138, 378, 447

Table 6: Converged circuit per run at 400k steps, under R2. All ten pairwise top-3 Jaccard indices are 0, which matches the uniform null (see text). Mean query alignment is $.304 \pm .021$, delta-only attention is $.059 \pm .001$, and $\max|EPE_0|$ is $1.360 \pm .040$.

F Emergence: Details

Null test for coordinate disjointness. The five 3-sets have zero pairwise Jaccard and 15 unique coordinates. Under a uniform null over 768 dimensions, all five sets are disjoint with probability .8887; a 100,000-trial simulation gives .8895 and mean pairwise Jaccard .0023. Disjointness therefore has no evidential weight alone. Exact recovery of $\{138, 378, 447\}$ in one run has chance probability 1.3×10^{-8} , but no run shows it; conversely, no

overlap with that set across five runs has probability .943 and is also uninformative. The supported evidence for run-specific coordinates is no agreement between runs plus label mobility until the final checkpoint.

Coverage is 5 runs by 17 checkpoints, or 85 of 85 completed evaluations, on a fixed 300-example set with layers 4–11 in fp32. Massive coordinates are re-identified at every checkpoint. The step-0 controls pass: the sink is .0341, bias alignment is zero, and every intervention efficacy is zero. The sink falls to .0201 at step 400 and then grows at every stored checkpoint. Coordinate labels stay mobile, with caprica locking at 40k, battlestar at 100k, and the other three only at 400k. The median 50%-onset is 20k for the sink, EPE amplitude, and delta-only attention. Alignment onsets are unstable because those curves are non-monotone, which is why the pre-specified ordering test is negative (statistic 0, $p = .25$, 3 of 5 runs). Route A’s causal share peaks at 14.2% near 70k steps and then falls to 7.7% by 400k, while its structural alignment continues to rise.

A scaling fault found by the second execution path. This section is where the two paths first disagreed, and the disagreement turned out to be a fault in our own code rather than a property of the models. We record it because the way it was found and hidden is itself a reproducibility result.

The original R1 executor hard-coded the attention scale as $D_h^{-1/2}$ and never read the checkpoint configuration. The Stanford CRFM checkpoints set both `scale_attn_by_inverse_layer_idx` and `reorder_and_upcast_attn`, so Hugging Face additionally divides layer l scores by $(l+1)$ and upcasts the softmax. Omitting the divisor inflates logits, sharpens every softmax, and therefore inflates exactly the quantities that are read through a softmax. The converged sink came out at $.897 \pm .008$ instead of $.650 \pm .005$, an inflation of 38%, and delta-only attention at $.555 \pm .013$ instead of $.059 \pm .001$, a factor of 9.4. Weight-derived quantities were never affected, which is why EPE amplitude, bias-key alignment, and every coordinate set matched throughout and the fault was invisible in half the analysis.

The fault survived our parity gate because that gate ran only on `openai-community/gpt2`, where both flags are off and the two paths therefore agree to 10^{-7} by construction. A gate that

never varies the configuration cannot test configuration handling. After correcting the executor we added a CRFM-specific gate, which passes on 68 of 68 quantities with a worst relative deviation of 4.6×10^{-6} , and re-ran the whole study on both paths. Across 5 runs $\times$ 17 checkpoints $\times$ 9 forward-dependent quantities, 765 paired measurements, the largest absolute difference is 2.4×10^{-7} . Converged values agree to eight decimals, all five coordinate triples are identical, and the ordering test returns the same statistic and p -value from both paths. Every qualitative finding in §5 held under the faulty path as well, including the step-400 dip, the anti-aligned early phase, the EPE overshoot, and the negative ordering test, so no conclusion in this paper ever rested on the inflated numbers. All reported values are the corrected ones.

One quantity remains numerically unstable rather than wrong. The route-A share of the delta score advantage divides by a per-cell quantity that can pass through zero, and the two paths disagree in 19 of 85 checkpoints, occasionally by orders of magnitude. Its converged value is reliable, at .104299 from both paths, agreeing to 2.6×10^{-7} , and that is the only value we quote. Delta-only attention, the route-A share of the delta score advantage, and the fraction of the full-model sink removed by nullifying b_Q have different denominators, so they are not estimates of one common percentage. We do not plot the unstable trajectory. Intervention efficacies before step 4k divide by a near-zero sink and should likewise be read as undefined rather than small.

These checkpoints also differ from OpenAI GPT-2 in data and schedule, so this section studies learning dynamics rather than replicating a training run.

G Evaluation: Details

Additional records follow. **Concordance:** Position-0 attention gives $\tau = .956$ against sink-rate .5 and $\tau = .867$ against rates .2/.3 and rank; entropy gives $\tau = .644$ against position-0 attention and .600 against rate .5. All five of fifteen pairs below .8 involve entropy. **Intensive against extensive** at threshold .3: (c)/(d) deactivate 99.9/98.0% of active head-cells; (i) keeps 70.3% active but shrinking; (b) yields 57.3% deactivated and 40.9% shrinking. **Length invariance:** k_0 and Δ_0 agree across nested prefixes within 2×10^{-6} relative error. Extreme long-context slopes for (b)/(d) reflect nonlinear curves

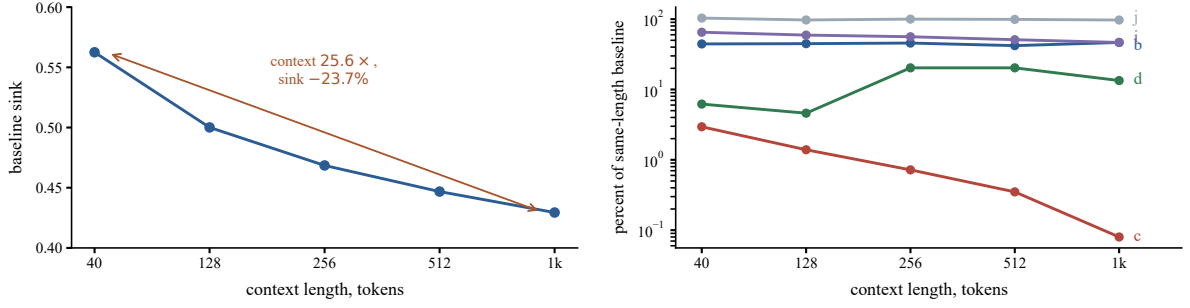


Figure 6: **Length generalization.** *Left:* the baseline sink falls only 23.7% while context grows 25.6 $\times$. *Right:* intervention residuals as a percentage of the same-length baseline, on a log scale, with each curve labelled at its right end. Remove First PE strengthens with length, Swap EPE is non-monotone, the key edit weakens steadily, and the random control stays near 100%.

	int.	pos. 0	rate >.3	rank	entr.	m_{1-4}	Gini
a	.5625	.886	1.36	.419	.021	.206	
b	.2498	.380	4.38	.563	.022	.421	
c	.0166	.001	19.8	.607	.053	.679	
d	.0350	.018	13.5	.488	.552	.690	
e	.5552	.887	1.36	.417	.033	.200	
f	.5606	.884	1.36	.421	.021	.207	
g	.2836	.375	6.22	.563	.051	.544	
h	.1024	.107	13.4	.611	.225	.624	
i	.3659	.630	1.77	.548	.034	.244	
j	.5820	.911	1.32	.395	.020	.188	

Table 7: Metric multiverse at 40 tokens (means over 3 seeds, SDs $\leq .008$). “ m_{1-4} ” is attention mass relocated to positions 1–4. Swap EPE relocates a concentrated anchor, while Remove First PE diffuses mass to positions 5 and beyond (.930). Fixed-draw (j); resampled control in Table 1. Band/seed set differs from Table 1, hence small baseline offsets.

regime	base	b	c	d	h	i
SST-2	.600	44.6	2.3	6.5	16.7	65.7
GSM8K	.548	43.1	2.4	7.1	23.8	66.0
HumanEval	.540	45.5	4.3	5.1	14.2	63.4
uniform-random	.602	55.1	2.1	5.7	20.6	68.0
Zipf/unigram	.523	45.7	2.6	8.6	26.8	63.3
shuffled natural	.508	46.0	2.9	8.3	28.4	63.7
repeated token	.533	57.5	4.3	10.0	6.4	72.8

Table 8: Content generality (% of each regime’s own baseline). The address of the anchor is stable across the seven tested regimes, since (c) leaves 2.1–4.3% everywhere. The route mixture is not, since (b) spans 43.1–57.5%.

and low-probability logits, not new scaling laws. **Frontier:** $b \wedge i$ removes 75.2% for +2.680 nats and $b \wedge i \wedge d$ 95.1% for +2.774; unstable reduction-per-nat values near zero cost are unused. **Position damage** in early/middle/late thirds: at 40 tokens, (b) 2.29/1.60/93, (c) 5.60/3.29/2.38, (i) .43/.20/.14; at 1024, (b) .55/.52/**1.09**, (c) 1.75/1.16/1.43, (i) .16/.14/.15. **Calibration and power:** bootstrap-SE/reseed-SD ratios are .94–3.63; paired MDEs

range from .0006 for (f) to .0222 for (h). The (e)-(f) gap .00536 is below its .00656 MDE and remains unresolved; fixed-draw (j), +.0195 versus MDE .00044, is a real draw-specific increase.

H Reference Execution and Parity

R2, our reference execution, uses NNsight 0.7.0 (Fiotto-Kaufman et al., 2025) over the native Hugging Face eager-attention path, with Transformers 5.3.0 and PyTorch 2.10.0+cu128. Eager attention is required because the experiments read realized attention probabilities, queries, keys, and logits. GPT-2, GPT-Neo, OPT-125M/1.3B, and Qwen2.5 up to 3B run in fp32. OPT-2.7B/6.7B/13B and Qwen2.5-7B/14B run in bf16. The emergence, residual-sink, and evaluation studies run in fp32.

Parity gates pass for every tested GPT-2 quantity: 10 of 10 intervention rows, 9 of 9 emergence quantities, 48 of 48 residual-sink quantities, and 1,090 of 1,090 evaluation rows. In the residual-sink gate, all primitive quantities passed the original tolerance, and one derived ratio needed a 5×10^{-4} relative-tolerance floor, which the report records. Sample manifests are identical between R1 and R2, so the comparison uses the same examples. All 16 model suites and all 85 emergence checkpoints completed, and the numeric audit found no NaN or infinite values.

The two paths agree to six displayed decimals for all GPT-2 scales, all GPT-Neo models, and OPT-125M/1.3B. They differ on OPT-2.7B/6.7B/13B, where the change of executor and the change of dtype are confounded, and on the five Qwen models, where the largest pooled residual change is 4.29 pp and the largest domain-level change is 5.34 pp. No family fingerprint reverses. Qwen

has no family-specific parity gate, so we report the observed values without assigning a unique cause. The residual-sink analysis was rerun under R2 only at 124M, so the 355M and 774M rows remain R1 results, and the OPT-13B single-sentence probe is R1 evidence. The emergence discrepancy is described in Appendix [F](#).

I Reproducibility Checklist

Every reported value traces to a CSV, JSON, or NPZ file in the released archive, which contains 2,069 files and 2,062,645,551 bytes. Each run records model geometry, dtype, layer band, seeds, and the exact coordinates ablated. All 16 model geometries were checked against the upstream configurations. Duplicate exports are documented, and a five-example timing tree is excluded from inference. The archive includes the sampled SST-2, GSM8K, and HumanEval manifests, pointers to the public Hugging Face checkpoints, and the five Stanford CRFM Mistral runs. It also includes per-model analogues of the original paper’s seven structural figures, and scripts that regenerate every table and figure in this paper from the archived results.