FEW NEURONS, MANY ROLES: SPARSE MLP NEURONS THAT CONTROL ATTENTION SINKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) send a disproportionately large share of attention to the first token, a pattern known as the attention sink. Although this pattern is widespread and functionally important, it is debated which upstream units build it and whether those units serve only the sink. We identify sparse populations of MLP neurons that causally control the sink across two model families. Suppressing these neurons weakens the sink by 12 to 51%, while amplifying their natural activity strengthens it by 1 to 6%. Exact softmax decomposition and live restoration show that these neurons act through later queries and keys, with a pathway mix that differs across models. Evaluating the models with the neurons suppressed reveals that these neurons significantly change downstream behavior differently across model families and benchmarks. Restoring the exact sink while the neurons remain suppressed recovers little or none of the behavioral effect, indicating that the sink explains little of this effect We further find that preserving their coordinates in high precision while quantizing the model to W8A8 recovers 65% of the loss, despite these neurons having little overlap with magnitude-selected neurons. A few neurons help build the sink, but their role reaches well beyond it.

1 INTRODUCTION

Language models often direct a disproportionate share of attention toward initial tokens even when those tokens carry little to no semantic information (Xiao et al., 2024; Gu et al., 2025). Attention sinks can play important roles: preserving the key–value states of sink tokens is critical for stable long-context streaming (Xiao et al., 2024), sinks provide a "no-op" place to send attention when no other token is useful (Bondarenko et al., 2023; Barbero et al., 2025; Ran-Milo, 2026), attention sinks can act as a stable positional anchor inside the model (Ruscio et al., 2025; de Llano et al., 2026). Additionally, modifying sink attention can affect downstream accuracy, activation outliers, and model compression behavior (Bondarenko et al., 2023; Yu et al., 2024; Zhang et al., 2025).

Existing accounts connect attention sinks to surplus attention under softmax normalization, avoidance of excessive token mixing, and the tagging of representations for later retrieval (Gu et al., 2025; Barbero et al., 2025; Zhang et al., 2025). Mechanistic studies the formation of sinks through residual activations and attention–MLP interactions (Sun et al., 2024; Ran-Milo et al., 2026). Neuron-level accounts identify attention-deactivation neurons, first-token amplifiers, and super-neuron structure associated with sink formation (Gurnee et al., 2024; Yona et al., 2025; Li et al., 2026). This coincides with prior work which show that small MLP populations can support particular behaviors like knowledge, skill, and language-specific abilities (Dai et al., 2022; Wang et al., 2022; Tang et al., 2024). Together, these findings motivate asking which neurons causally sustain the sink and how that role relates to their broader contribution to model behavior.

In this work, we identify sparse MLP populations that causally control attention sinks. We localize them independently in instruction-tuned Qwen2.5-1.5B, Qwen2.5-3B, Llama-3.2-3B, and Llama-3.1-8B by shortlisting neurons able to affect the strongest sink layers and directly measuring each neuron's causal sink effect. Our findings:

- **Sparse bidirectional control.** These *sink-control neurons* (SCNs), at most 0.05% of candidate MLP neurons, reduce sink strength by 12–51% when suppressed and increase it by 1–6% when amplified.

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

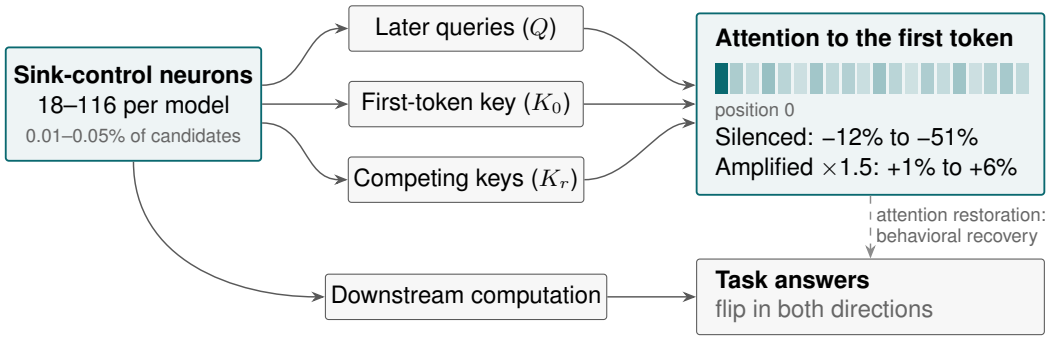

Figure 1: Overview. Sink-control neurons influence first-token attention through later queries and keys (§3 and 7) and shape task-dependent behavior (§4–5). Attention restoration tests whether recovering the sink also recovers behavior while the neurons remain suppressed (§6).

- **Query–key pathways.** SCNs act through later queries and keys, with a pathway mix that differs across checkpoints.
- **Behavior beyond the sink.** Suppression changes predictions on 12 benchmarks in both directions, persists at matched sink reduction, and involves neurons largely distinct from task-supporting neurons.
- **Sink restoration.** Restoring the original sink attention while SCNs remain suppressed recovers little or none of the behavioral effect.
- **Precision.** Keeping SCNs in high precision recovers 65% of the W8A8 quantization loss.

Thus, a tiny MLP population helps construct the sink while also supporting computation beyond it.

2 Preliminaries and Related Work

Attention sinks. Xiao et al. (2024) showed that autoregressive models concentrate attention on initial tokens and that preserving those key–value states stabilizes streaming. The phenotype recurs across families, scales, and stages (Gu et al., 2025; Zhang et al., 2025; Su et al., 2026); proposed roles include a softmax no-op that limits over-mixing (Barbero et al., 2025; Bondarenko et al., 2023), an emergent bias (Gu et al., 2025; Ran-Milo, 2026), and a positional or tagging anchor (Ruscio et al., 2025; de Llano et al., 2026; Zhang et al., 2025); formation studies implicate massive activations and attention–MLP interactions (Sun et al., 2024; Guo et al., 2024; Ran-Milo et al., 2026); and the sink shapes KV-cache policy (Su & Yuan, 2025; Son et al., 2024) and motivates sink-free designs (Zuhri et al., 2026; Qiu et al., 2025; Fu et al., 2026). We add the missing unit-level causal account: which neurons control the sink, through which representations, and at what cost.

An MLP coordinate is a residual write. Decoder block j of a pre-norm transformer applies causal self-attention and then a gated MLP to the residual stream x. Attention head h computes

$$A^{(j,h)} = \mathrm{softmax}\Big(Q^{(j,h)}K^{(j,h)\top}/\sqrt{d_k} + M\Big), \qquad d_k = 128, \tag{1}$$

with causal mask M, under grouped-query attention (12/2, 16/2, 24/8, 32/8 query/key–value heads in our four checkpoints); $A^{(j,h)}_{t,0}$ is the probability that query position t assigns to the first key. Writing r^ℓ_t for the post-attention residual and $u^\ell_t = \mathrm{RMSNorm}(r^\ell_t)$, the gated MLP computes

$$a^\ell_t = \mathrm{SiLU}(W^\ell_g u^\ell_t) \odot (W^\ell_u u^\ell_t), \qquad x^{\ell+1}_t = r^\ell_t + \sum_i a^\ell_{t,i} w^\ell_{d,i}. \tag{2}$$

A neuron is coordinate (ℓ, i) immediately before `down_proj`. Its activation weights a residual direction (Geva et al., 2021; 2022). Since attention precedes the MLP, its write can affect attention only at $j > \ell$. We change natural activations to $\alpha a^\ell_{t,i}$ at every token position, without editing weights. This identifies a population's intervention effect, not a feature basis or an exclusive semantic function (Elhage et al., 2022; Marks et al., 2025).

Table 1: Bidirectional causal control by independently frozen SCN populations: suppression on the locked test split, amplification on a fresh disjoint corpus with the same frozen neurons and controls. Δ_{spec} is the paired target-minus-random effect (pp); P95 the 95th percentile of 20 layer/count-matched random sets; ρ the dose-response Spearman. Baseline sinks 0.618, 0.672, 0.841, 0.843.

		Suppression, $\alpha = 0$				Amplification, $\alpha = 1.5$			
Model	k (frac.)	RSR	P95	Δ_{spec} [95% CI]	ρ	RSG	P95	Δ_{spec} [95% CI]	ρ
Qwen2.5-1.5B	116 (0.05%)	**11.747%**	0.089%	11.7 [11.200, 12.244]	1.0	**+3.110%**	0.017%	+3.113 [2.990, 3.242]	1.0
Qwen2.5-3B	28 (0.01%)	**28.022%**	0.025%	28.2 [27.149, 29.198]	1.0	**+1.003%**	0.005%	+1.006 [0.957, 1.058]	1.0
Llama-3.2-3B	18 (0.01%)	**51.052%**	0.00317%	51.055 [49.879, 52.183]	1.0	**+6.399%**	0.00409%	+6.399 [5.946, 7.211]	1.0
Llama-3.1-8B	42 (0.01%)	**30.278%**	0.008%	30.3 [29.534, 31.127]	1.0	**+5.136%**	0.007%	+5.135 [5.036, 5.237]	1.0

Related neuron populations and feature-level analyses. The closest accounts identify attention-deactivation neurons (Gurnee et al., 2024), first-token amplifiers (Yona et al., 2025), massive residual activations (Sun et al., 2024), and super-neuron structure (Li et al., 2026). Recent work also separates massive activations from sink function (Sun et al., 2026). Other accounts show that sparse MLP units are tied to knowledge, skills, languages, positions, repetition, and tasks (Dai et al., 2022; Wang et al., 2022; Tang et al., 2024; Voita et al., 2024; Hiraoka & Inui, 2025; Song et al., 2024). These proposals rest on different models and observational or single-direction evidence. Our localization follows the attribute-then-verify logic of attribution patching and circuit analysis (Syed et al., 2024; Kramár et al., 2024; Conmy et al., 2023; Wang et al., 2023) at the neuron level (Bricken et al., 2023; Templeton et al., 2024).

3 Sparse Populations that Control Sinks

Checkpoints and data. We study instruction-tuned Qwen2.5-1.5B, Qwen2.5-3B, Llama-3.2-3B, and Llama-3.1-8B in bfloat16 at batch size 1. Localization uses a neutral OpenWebText corpus of 40-token sequences in disjoint discovery, validation, and test splits of 100 examples each; the test split stays locked until every later choice is frozen, and no benchmark example, label, or score influences localization.

Sink layers and metric. For attention layer l we register

$$S_l = \mathbb{E}_{h,\, t\in\mathcal{T}_2} A_{t,0}^{(l,h)}, \qquad \mathcal{T}_2 = \{20, \ldots, 39\}, \tag{3}$$

the position-0 attention averaged over heads and second-half queries. A layer is registered as a sink layer if its discovery-split mean $\bar{S}_l$ lies in the top quartile across layers and $\bar{S}_l \geq 0.15$, yielding 7, 9, 7, and 8 sink layers in the four models. For an activation multiplier α applied to the selected neurons, we report relative sink reduction $\mathrm{RSR}(\alpha) = (S^0 - S_\alpha)/S^0$ and relative sink gain $\mathrm{RSG}(\alpha) = (S_\alpha - S^0)/S^0$.

Candidate neurons and causal screening. For each causally eligible neuron n in layer l and sink layer $j > l$, we compute the signed activation–gradient score

$$I_{j,l,n} = \mathbb{E}_{x,t}\left[a_{l,t,n}(x)\frac{\partial S_j(x)}{\partial a_{l,t,n}(x)}\right]. \tag{4}$$

The score only shortlists candidates ($I_{j,l,n} > 0$); localization rests on measured effects. Each shortlisted neuron is silenced individually on 20 discovery passages, and only neurons with positive measured sink reduction $D_{j,\ell,i} = \mathbb{E}_x[S_j^0(x) - S_j^{(-\ell i)}(x)]$ are retained and ranked.

Sparse MLP neuron population construction. For each registered fraction $f \in \{0.01, 0.05, 0.10, 0.25, 0.50, 1.00\}\%$ of the causally eligible neuron pool of size P, we define

$$k = \max\left(1, \mathrm{ROUND}\left(P\frac{f}{100}\right)\right). \tag{5}$$

The k neurons are distributed across frozen sink layers in proportion to their baseline sink strength,

$$w_j = \frac{\bar{S}_j}{\sum_{u\in\mathcal{L}}\bar{S}_u}, \qquad q_j \approx k w_j, \tag{6}$$

Table 2: Downstream specificity under full and sink-matched SCN suppression. Entries are target-minus-layer-matched-random accuracy changes (pp). Full suppression uses $\alpha = 0$; matched suppression uses α^* calibrated on independent neutral text to ≈10% RSR. **Bold** matched effects survive Holm correction (12 tests per model); † marks nominal matched effects that do not survive the family-wise threshold. n/a marks model–task pairs below the registered baseline-viability floor (Random Chance + 5%; App. H); their results are not reported or interpreted. Holm families retain all 12 registered tests, which is conservative.

Model	Intervention	MMLU	ARC	CB-E	CB-H	GSM8K	BLEnD	HSwag	Wino	PIQA	CSQA	GMLU	NormAd
Qwen2.5-1.5B	Full	−0.4	−0.4	−1.7	−2.4	n/a	+3.3	+0.2	+0.5	+0.1	+0.1	−1.4	0.0
	Matched	−0.300	−0.085	−0.815	−2.608†	n/a	**+3.562**	−0.067	+0.947	+0.109	0.000	**−1.250**	−0.459
Qwen2.5-3B	Full	−3.3	−2.6	−5.7	−8.3	−1.5	+1.6	−3.7	−1.7	−2.3	−3.9	−5.4	n/a
	Matched	−1.333†	−0.512	−1.711†	**−4.482**	−0.303	**+2.312**	**−1.033**	−0.631	−0.816	−0.819	**−1.696**	n/a
Llama-3.2-3B	Full	−25.3	−17.3	−15.6	−19.2	−64.2	−32.3	−15.6	−10.4	−11.6	−25.1	−20.7	−26.5
	Matched	−0.567	−1.109	−0.081	−0.489	**−4.701**	−0.031	+0.267	−2.131†	−0.381	−0.246	−0.893	−0.880†
Llama-3.1-8B	Full	−23.8	−42.6	−20.6	−20.1	−72.0	−23.3	−14.6	−12.5	−8.1	−18.9	−17.7	−22.3
	Matched	**−6.567**	**−13.908**	**−5.379**	**−7.009**	**−29.492**	**−14.688**	**−2.067**	−1.579	**−1.795**	**−6.061**	**−4.600**	**−2.141**

where integer quotas q_j are obtained using the largest-remainder method such that $\sum_{j\in\mathcal{L}} q_j = k$. Each quota is filled with the highest-ranked measured candidates satisfying the causal constraint $\ell < j$.

Loss and gain of function. The smallest tested MLP coordinates passing the causal gate contain 116, 28, 18, 42 neurons which are 0.05%, 0.01%, 0.01%, 0.01% of 232,960, 275,200, 180,224, 415,744 eligible coordinates (Table 1). Their suppression reduces the sink by at least 10%, while suppressing a randomly selected layer matched number of neurons show negligible change in sink strength (0.09% change at most). Furthermore, both directions are perfectly dose-monotonic (= 1.0). Amplification growing through the grid in every model (Llama-3.1-8B: +1.216, +2.855, +5.136% at 1.10, 1.25, 1.50; App. E), and every registered sink layer (7/7, 9/9, 7/7, 8/8) moves in the predicted direction. This strongly indicates that these neurons have a causal ability to modulate sink strength across different model families.

However, the sink control is not free. At $\alpha = 1.5$ the intervention also perturbs the output distribution (CE +0.201, +0.066, +0.07185, +0.116; forward KL 0.278, 0.105, 0.09636, 0.179; top-1 flips 7.56, 6.36, 17.33, 14.10%). This indicates that SCNs are not sink-exclusive or behaviorally inert.

4 Downstream Behavior of SCN Suppression

We evaluate the four models with frozen SCNs on 12 benchmarks covering knowledge, reasoning, commonsense, culture, multilingual questions, and normative judgments. We test baseline, full and half target suppression, and dose-matched random controls. The primary behavioral contrast is target minus layer/count-matched random.

Full suppression. We set all SCN activations to zero ($\alpha = 0$). The sink phenotype transfers to task prompts: task-prompt RSR ranges from 6.5–13.2% in Qwen2.5-1.5B, 14.1–30.7% in Qwen2.5-3B, 27.8–53.2% in Llama-3.2-3B, and 21.7–42.1% in Llama-3.1-8B. Behavior differs sharply across models. Qwen2.5-1.5B changes by only -2.36 to $+3.00$ pp, with BLEnD improving by $+3.00$ pp (specificity $+3.28^{\dagger}$). Qwen2.5-3B ranges from -9.13 to $+2.31$ pp and also improves on BLEnD. Both Llama models degrade on all 12 tasks, including -64.29 and -71.80 pp on GSM8K for Llama-3.2-3B and Llama-3.1-8B, respectively. The same benchmark can therefore move in opposite directions: BLEnD changes by $+3.00$, $+2.31$, -32.16, and -23.16 pp across the four checkpoints.

However, full suppression does not provide a fair cross-model comparison because each model produces different RSR: 11.7%, 28.0%, 51.1%, and 30.3%.

Matched suppression. We therefore calibrate one activity multiplier α^* per model on a fresh 300-example neutral corpus. All other experimental choices remain unchanged. The target is $|\mathrm{RSR}(\alpha^*) - 0.10| \leq 0.005$.

We use a deterministic bracket-and-midpoint search with at most 17 evaluations and a monotonicity tolerance of 0.002 (App. I). The resulting multipliers are $\alpha^* = 0.0625, 0.078125, 0.671875, 0.40625$ (neutral RSR 10.200, 9.593, 10.033, 9.983%) for Qwen2.5-1.5B, Qwen2.5-3B, Llama-3.2-3B, and Llama-3.1-8B, respectively. Thus, similar sink reduction requires substantially different amounts of SCN suppression.

The behavioral differences remain after matching sink reduction (Table 2). Qwen2.5-1.5B has two Holm-significant effects: BLEnD improves by $+3.562$ pp ($p = 0.0012$), while Global-MMLU-Lite declines by -1.250 pp. Qwen2.5-3B has four significant effects: CulturalBench Hard (-4.482 pp), BLEnD ($+2.312$ pp), HellaSwag (-1.033 pp), and Global-MMLU-Lite (-1.696 pp). Llama-3.2-3B retains one significant effect, a -4.701 pp loss on GSM8K, whereas Llama-3.1-8B retains significant losses on 11 of 12 tasks, including -29.492 pp on GSM8K and -14.688 pp on BLEnD. WinoGrande is the only task whose confidence interval includes zero (-1.579 [−3.594, 0.399]).

Matching sink reduction therefore does not equalize behavior. SCN effects remain model- and task-dependent, and their direction can reverse across checkpoints.

Net accuracy hides substantial prediction churn. Aggregate accuracy understates how strongly SCN suppression changes individual answers. In the held-out transition analysis (§5, Table 4), the sink-population arm changes 6.0–37.6% of predictions across eight task–model pairs, in both directions: on Qwen2.5-3B BLEnD, 9.10% of answers turn wrong while 11.02% turn correct, and on Qwen2.5-3B WinoGrande 6.00% change although accuracy equals the matched control.

5 Dissociation From Task Neurons and Answer Flipping

The downstream effects in §4 raise a natural question: **do SCNs affect downstream behavior because of overlap with task supporting neurons?**

Task-neuron localization. We test this by localizing the task supporting neurons with the same attribute-then-measure procedure used in §3. For each task we optimize the correct-answer margin,

$$M = \log P(y^*) - \log \sum_{y \neq y^*} e^{\log P(y)}, \qquad I_{l,i} = a_{l,i}\, \partial M / \partial a_{l,i}, \tag{7}$$

then individually silence the shortlisted neurons and rerank them by measured margin loss.

Membership overlap is minimal. At matched primary set sizes, SCN and task-supporting populations share only 0–1 of 116 neurons in Qwen2.5-1.5B and 0–2 of 28 neurons in Qwen2.5-3B across the four tasks. Overlap increases for larger prespecified sets, indicating some broader alignment, but the full attribution rankings remain essentially uncorrelated ($|\rho| \leq 0.0072$). Thus, the strongest sink-supporting and task-supporting neurons are largely distinct, yet suppression of SCNs significantly affect the selected tasks. (full overlap statistics, layer-matched nulls, and graded-k results in Appendix J).

Crossed suppression dissociates the functions. Each subset, task-only $B = T \setminus C$ and SCN-only $C' = C \setminus T$, is suppressed at $\alpha = 0$ against its own size- and layer-matched control on held-out clusters (Table 3). The pattern is consistent: task-only suppression produces large task effects with almost little sink movement (single-model ARC -34.97 pp at RSR $\leq 4.86\%$; BLEnD -31.23 pp at RSR -1.55%), while sink-only suppression moves the sink by up to 25.84% with small, sign-varying task effects (-5.07 to $+2.71$ pp). This indicates that large changes in task behavior can occur with little sink change, and large sink changes can occur with only small or inconsistent behavioral effects.

Answer flipping. Net accuracy understates what these interventions do. On the held-out splits we partition every example by baseline and post-intervention prediction into

$$\{\mathrm{CC},\ \mathrm{CW},\ \mathrm{WC},\ \mathrm{WW}^{=},\ \mathrm{WW}^{\neq}\}, \qquad \mathrm{Flip} = \mathrm{CW} + \mathrm{WC} + \mathrm{WW}^{\neq}, \tag{8}$$

where C/W mark correct/wrong before→after, the five shares sum to 100%, and $\mathrm{WW}^{\neq}$ counts answers that stay wrong but *change*. Table 4 gives this partition for the sink-population arm (the frozen SCN set minus its ≤ 2 task-shared neurons) against its matched-random control. Three regularities emerge. First, suppression rewrites predictions at 1.5 to 12.9 times the matched-random rate, reaching 37.6% of all held-out examples. Second, most of the churn is invisible to accuracy. On Qwen2.5-1.5B CulturalBench Hard the accuracy gap versus random is only -2.42 pp, yet

Table 3: Crossed subset suppression on held-out clusters (cross-model extension): accuracy change versus each arm's own layer/count-matched control (pp) and the arm's task-prompt sink RSR. Benchmark selection criterion, shared arms are underpowered by rule ($|C \cap T| < 10$); single-model results and CIs in App. J.

		Task-only $T \setminus C$		Sink-only $C \setminus T$	
Model	Task	Δacc (pp)	sink RSR	Δacc (pp)	sink RSR
Qwen2.5-1.5B	CulturalBench Hard	−22.612	+0.649%	−2.418	8.844%
	Global-MMLU-Lite	−4.976	+0.655%	−1.262	8.113%
	BLEnD	−31.232	−1.551%	+2.714	7.527%
	WinoGrande	0.000	+0.256%	+1.153	13.086%
Qwen2.5-3B	CulturalBench Hard	−28.295	+0.192%	−0.242	1.937%
	Global-MMLU-Lite	−0.071	−0.040%	−5.071	25.840%
	BLEnD	−16.117	−0.276%	+1.294	24.470%
	WinoGrande	0.000	+0.006%	0.000	1.646%

Table 4: Prediction-transition structure under sink-population suppression on held-out splits. Each cell is target / matched-random, in % of all N examples; Flip = CW + WC + $\mathrm{WW}^{\neq}$ (Eq. 8). WinoGrande's $\mathrm{WW}^{\neq}$ is structurally zero (binary choice).

Model	Task	N	CW	WC	CC	$\mathrm{WW}^{=}$	$\mathrm{WW}^{\neq}$	Flip	Accuracy
Qwen2.5-1.5B	CulturalBench Hard	827	6.29 / 1.09	4.23 / 1.45	17.53 / 22.73	44.86 / 66.38	**27.09 / 8.34**	**37.61 / 10.88**	21.77 / 24.18
	Global-MMLU-Lite	4,200	4.60 / 0.76	3.64 / 1.07	44.55 / 48.38	41.98 / 48.52	5.24 / 1.26	**13.48 / 3.10**	48.19 / 49.45
	BLEnD	2,395	5.59 / 0.75	9.23 / 1.67	50.31 / 55.16	26.35 / 41.21	8.52 / 1.21	**23.34 / 3.63**	59.54 / 56.83
	WinoGrande	867	3.69 / 1.73	5.54 / 2.42	59.05 / 61.01	31.72 / 34.83	0.00 / 0.00	**9.23 / 4.15**	64.59 / 63.44
Qwen2.5-3B	CulturalBench Hard	827	3.75 / 3.02	2.78 / 2.30	25.27 / 26.00	54.17 / 59.85	14.03 / 8.83	**20.56 / 14.15**	28.05 / 28.30
	Global-MMLU-Lite	4,200	12.86 / 0.79	7.79 / 0.79	44.81 / 56.88	25.31 / 40.81	9.24 / 0.74	**29.88 / 2.31**	52.60 / 57.67
	BLEnD	2,395	9.10 / 1.21	11.02 / 1.84	55.24 / 63.13	19.08 / 32.69	5.55 / 1.13	**25.68 / 4.18**	66.26 / 64.97
	WinoGrande	867	2.54 / 0.81	3.46 / 1.73	65.86 / 67.59	28.14 / 29.87	0.00 / 0.00	**6.00 / 2.54**	69.32 / 69.32

27.09% of examples switch to a *different wrong* answer versus 8.34%; on Qwen2.5-3B BLEnD, target accuracy is *higher* than random (66.26 vs 64.97%) while 25.68% of predictions change versus 4.18%. Third, flip composition is structured by the task: Qwen2.5-3B Global-MMLU-Lite churn is dominated by newly wrong answers (CW 12.86 vs 0.79%), whereas WinoGrande has $\mathrm{WW}^{\neq} = 0$, a changed wrong answer being necessarily correct with two options. These results suggest that SCN suppression perturbs internal computations involved in answer selection, rather than simply removing task capability.

6 Restoring Sink Does Not Restore Behavior

Because the sink can shape behavior as an anchor, we test whether the behavioral effects run through it by restoring sink attention while keeping every SCN suppressed.

Rescue intervention. We replace the scalar S_l with a full vector of position-0 probabilities $p_{l,h,q,0}$ for each sink layer, head, and second-half prompt queries $q \in [\lfloor P/2 \rfloor, P)$. For each exact input we cache the baseline value $b = p^0(0)$ per row, keep the neurons fully suppressed, and overwrite each row after the softmax and before the value contraction:

$$p^{\mathrm{res}}(0) = b, \qquad p^{\mathrm{res}}(k) = (1-b)\,\frac{e^{z_k}}{\sum_{j\neq 0} e^{z_j}} \;\; (k \neq 0), \qquad O = p^{\mathrm{res}}\,V, \tag{9}$$

where z are the current logits; sink layers are patched sequentially, so later layers respond naturally to earlier rescues. This restores the *full registered position-zero vector*, not just its average, while preserving causal support and nonzero-key conditional odds. It is the unique such normalized row and, for an interior recipient distribution, its minimum-KL projection under $p_0 = b$ (proof in Appendix C). Values and all unpatched computations remain those of the evolving suppressed run. We define the behavioral loss from SCN suppression as:

$$L = A_{\mathrm{base}} - A_{\mathrm{sup}}, \qquad G = A_{\mathrm{res}} - A_{\mathrm{sup}}, \qquad J = A_{\mathrm{res}} - \frac{1}{5}\sum_{s=0}^{4} A_{\mathrm{TS}_s}. \tag{10}$$

where A_{TS_s} is the accuracy under suppressed sham condition s.

Table 5: Attention rescue with neurons suppressed; sink restoration is exact (100%) on every task. L, G, J follow Eq. 10 (pp, [95% CI]); **bold** survives Holm per model. The sham-adequacy gate fails on all seven tasks, so no position-0-specific mediation is claimed.

Model	Task	A_{base}	A_{sup}	A_{res}	L [95% CI]	G [95% CI]	J [95% CI]
Qwen2.5-1.5B	CulturalBench Easy	61.451	59.250	59.006	+2.200 [0.733, 3.667]	−0.244 [−1.141, 0.652]	−0.782 [−1.516, −0.081]
	CulturalBench Hard	23.390	21.027	21.190	+2.363 [0.570, 4.238]	+0.163 [−1.304, 1.630]	+1.043 [−0.179, 2.266]
	MMLU	58.795	58.275	58.332	+0.520 [0.128, 0.919]	+0.057 [−0.278, 0.392]	+0.060 [−0.221, 0.343]
	ARC-Challenge	74.317	74.147	74.147	+0.171 [−1.024, 1.365]	0.000 [−0.853, 0.853]	−0.085 [−0.751, 0.597]
Qwen2.5-3B	CulturalBench Easy	67.24	61.21	63.00	**+6.03** [4.16, 7.91]	**+1.79** [0.49, 3.10]	+1.24 [−0.02, 2.49]
	CulturalBench Hard	31.13	23.80	25.75	**+7.33** [4.89, 9.86]	+1.96 [−0.57, 4.48]	**+4.63** [2.93, 6.36]
	MMLU	66.900	63.400	63.567	**+3.500** [2.200, 4.767]	+0.167 [−0.800, 1.167]	**+3.920** [2.840, 5.007]
	ARC-Challenge	83.362	80.631	80.631	**+2.730** [1.024, 4.437]	0.000 [−1.195, 1.195]	**+1.758** [0.631, 2.901]

In Qwen2.5-1.5B, exact sink restoration produces no resolved behavioral recovery. In Qwen2.5-3B, it recovers only 17–30% of the accuracy lost under SCN suppression, leaving most of the effect unchanged. Thus, restoring the sink is not sufficient to restore the behavior produced by the intact SCNs.

This result does not show that attention is irrelevant to behavior. The registered sham controls fail their adequacy criterion, so we cannot attribute the partial recovery specifically to position-0 attention. These results indicate that: *the registered attention sink does not explain most of the behavioral effect of SCN suppression*.

7 Query–Key Pathways Carry Sink Control

Through which representations do SCNs act? For every frozen sink layer we capture the native post-RoPE tensors entering Eq. 1 on a fresh 200-example neutral corpus and write the position-0 probability as $F(Q, K_0, K_r)$, where Q collects second-half queries, K_0 the anchor key, and K_r all competing keys. Under full SCN suppression, each factor takes its baseline or perturbed value, giving 2^3 factorial cells evaluated exactly from cached tensors. We convert the eight cells into exact three-player Shapley values and report their signed contributions to sink loss,

$$\phi_X = \sum_{S \subseteq N \setminus \{X\}} \frac{|S|!\,(2-|S|)!}{3!} \big[F(S \cup \{X\}) - F(S)\big], \qquad C_X = -\phi_X \tag{11}$$

where $N = \{Q, K_0, K_r\}$. The decomposition closes exactly, $C_Q + C_{K_0} + C_{K_r} = S_0 - S_{\alpha=0}$, with audited error $\leq 10^{-12}$, thus interactions introduced by the softmax are allocated without leaving an unexplained residual. Negative C_X is meaningful: it indicates that suppression changes that pathway in a direction that *opposes* the net sink loss. We validate each decomposition with live interventions while keeping the SCNs suppressed. At one sink layer at a time, we restore the baseline Q, K_0, or K_r before native RoPE and measure

$$G_X = S_{\alpha=0,\text{restore } X} - S_{\alpha=0}, \qquad R_X = \frac{\text{mean}(G_X)}{\text{mean}(S_0 - S_{\alpha=0})}.$$

Full Q/K restoration reproduces the intact same-layer attention matrix exactly, whereas restoring the position-0 value vector leaves the attention weights unchanged. A pathway is supported only when both its Shapley contribution and live-restoration gain are positive with 95% confidence intervals above zero, their conjunction survives Holm correction over the three pathways, the contribution exceeds the corresponding matched-random P95, and the total sink loss is resolved. Inference uses 10,000 paired whole-example bootstrap resamples; heads, queries, and layers are not treated as independent observations.

The three analyzed checkpoints reach the same sink phenotype through sharply different pathway mixtures (Table 6). **Qwen2.5-1.5B** loses 0.0759 sink attention ($0.637 \rightarrow 0.561$); Q, K_0, and K_r contribute 33.9%, 34.3%, and 31.8% of this loss, respectively, and all three are supported. Their individual live restorations recover 38.7%, 34.9%, and 36.6%; the sum exceeds 100% because each restoration is a separate nonlinear softmax counterfactual. **Qwen2.5-3B** instead shows a large positive K_0 contribution (135.8%) and a smaller K_r contribution (4.1%), opposed by Q (-39.9%).

Table 6: Exact pathway decomposition of the suppression-induced sink loss and its live single-pathway restoration. Share is C_X relative to the net loss; ×P95 is the multiple of the matched-random 95th percentile; negative entries oppose. Net losses: 0.0759 [0.0738, 0.0779], 0.1894 [0.1849, 0.1936], 0.2639 [0.2589, 0.2695].

Model	Path	C_X [95% CI]	Share	G_X [95% CI]	R_X	×P95	Supported
Qwen2.5-1.5B	Q	**0.0258** [0.0246, 0.0269]	33.9%	**0.0293** [0.0281, 0.0306]	38.7%	34	✓
	K_0	**0.0260** [0.0253, 0.0267]	34.3%	**0.0265** [0.0258, 0.0272]	34.9%	133	✓
	K_r	**0.0242** [0.0231, 0.0252]	31.8%	**0.0278** [0.0267, 0.0289]	36.6%	202	✓
Qwen2.5-3B	Q	−0.0756 [−0.0773, −0.0737]	−39.9%	−0.0529 [−0.0547, −0.0511]	−27.9%	fails	
	K_0	**0.2572** [0.2519, 0.2620]	135.8%	**0.2660** [0.2603, 0.2712]	140.4%	8334	✓
	K_r	**0.0078** [0.0050, 0.0099]	4.1%	**0.0111** [0.0086, 0.0132]	5.9%	75	✓
Llama-3.1-8B	Q	**0.2703** [0.2666, 0.2739]	102.4%	**0.2806** [0.2771, 0.2840]	106.3%	3251	✓
	K_0	**0.0027** [0.0014, 0.0040]	1.0%	**0.0039** [0.0023, 0.0055]	1.5%	194	✓
	K_r	−0.0090 [−0.0136, −0.0036]	−3.4%	0.0005 [−0.0042, 0.0059]	0.2%	fails	

Accordingly, restoring intact queries lowers the sink further ($R_Q = -27.9\%$), while restoring K_0 overshoots baseline recovery (140.4%). **Llama-3.1-8B** shows the complementary pattern: Q accounts for 102.4% of the net loss with 106.3% live recovery, K_0 contributes a small but supported 1.0%, and K_r opposes the loss and is unsupported. Supported contributions exceed matched-random P95 by $34\times$ to $8{,}334\times$.

These aggregate mixtures also vary with depth: competing keys contribute most strongly at Qwen2.5-1.5B layer 6 while K_0 grows toward layer 26; Qwen2.5-3B develops a strong compensating query effect around layers 6–9; and Llama's layer-25 sink loss (0.5132) is almost entirely query-borne (0.5109). As a secondary direction check, amplification at $\alpha = 1.5$ reverses all three supported Qwen2.5-1.5B pathways and both supported Qwen2.5-3B pathways; in Llama, Q reverses but the small supported K_0 term does not. We did not register pairwise pathway-dominance tests, so magnitude differences are descriptive. The supported conclusion is therefore a common causal chain,

$$\boxed{\text{sparse MLP neurons} \rightarrow \text{later query/key representations} \rightarrow \text{position-0 attention}}$$

implemented through a checkpoint-specific mixture of query, anchor-key, and competing-key effects.

8 SCNs ALSO MATTER FOR PRECISION

Sinks are frequently linked to activation outliers and quantization difficulty (Bondarenko et al., 2023; Dettmers et al., 2022; Sun et al., 2024). We test the connection causally at the Qwen2.5-1.5B SCN coordinates on two fresh zero-overlap corpora (400×128, 384×512 tokens), while recording the inputs to all 196 decoder linear operators and 28 post-block residual streams (App. L).

Propagated outlier bridge. Suppressing an SCN trivially changes the tensor that contains it, so we exclude effects that can arise directly at the intervention site. We measure only operators that are causally downstream of the SCN population ($l < m$ for linear inputs and $l \leq m$ for post-block residuals), and we mask selected same-layer coordinates identically in the baseline and intervention runs. For the masked input Z_o of eligible operator o, we define

$$B_{\text{prop}}(x) = \frac{1}{|O_{\text{prop}}|} \sum_{o \in O_{\text{prop}}} \log \max(\|Z_o(x)\|_\infty, \varepsilon)\,, \qquad E = \overline{B^0_{\text{prop}} - B^{\alpha=0}_{\text{prop}}}. \tag{12}$$

On the test split, suppressing the SCNs produces a small but specific reduction in downstream activation extremes ($E = 0.001103$), well above matched-random controls (P95 $= 0.000300$, median $= 2.5 \times 10^{-5}$; Holm $p = 0.0004$). The effect is much stronger at downstream W_d inputs (0.0478, vs. random P95 $= 0.00106$), suggesting that the propagated scale change is concentrated in later MLP inputs. By contrast, amplifying the SCNs ($\alpha = 1.5$) does not produce a clear increase in the downstream activation-extreme score defined in Eq. 12. We also test whether SCNs coincide with recurrent position-0 massive activations. Following the massive activation criterion of Sun et al. (2024), we require $|r| > 100$ and $|r|/\operatorname{median}(|r|) \geq 10^3$ on at least 50% of passages. No coordinate satisfies these criteria. Thus, SCNs causally affect downstream activation scale, particularly at later

Table 7: W8A8 fake quantization on held-out text (Qwen2.5-1.5B); each protected set holds 116 coordinates; FP32 reference CE 2.8393 (PPL 17.10). Gain is CE recovered versus naive W8A8 [95% CI]; **bold** survives Holm; global magnitude is descriptive (it abandons the SCN layer histogram).

Condition	CE	PPL	Gain over naive W8A8 [95% CI]
Naive W8A8	5.0722	159.53	gap = 2.2330 [2.1965, 2.2706]
+ protect SCN	3.6151	37.15	**+1.4572** [1.4257, 1.4883] (65.3%)
+ protect layer-matched magnitude	3.3343	28.06	+1.7380 [1.7079, 1.7693]
+ protect global magnitude	3.0575	21.27	+2.0147 [1.9809, 2.0492]
+ protect super-neuron proxy	4.8976	133.97	**+0.1746** [0.1511, 0.1979]
+ protect random (median of 20)	5.0920	162.72	P95 = −0.0105

MLP inputs, but they are distinct from the recurrent massive-activation phenomenon studied by Sun et al. (2024; 2026).

Sparse quantization protection. We fake-quantize decoder linear weights and activations with symmetric max-absolute per-tensor scaling, $Q_b(X) = \mathrm{clamp}(\mathrm{round}(X/s), -q_{\max}, q_{\max})\, s$, where $s = \max|X|/q_{\max}$ and $q_{\max} = 2^{b-1} - 1$. The FP32 reference has no resolved drift (+0.000299 CE). For protected coordinates P, we compute $y = Q(W_U)Q(x_U) + W_P x_P + b_{\mathrm{lin}}$ with separate protected/unprotected scales.

Naive W8A8 increases CE by 2.2330. Keeping only the 116 SCNs in high precision recovers 1.4572 CE (65.26%) while protecting just 0.0228% of activation coordinates and 0.0136% of weight elements (Table 7). Matched-random sets are null or slightly harmful, while a structural super-neuron proxy (Li et al., 2026) (App. F) recovers little (SCN advantage +1.2825).

Magnitude does not explain the effect: SCNs overlap only $9/116$ coordinates with the layer-matched magnitude set, which nevertheless recovers more loss (difference −0.2808; SCN efficiency 83.8%), and global-magnitude allocation performs better still. Thus, SCNs are strongly quantization-relevant but not more optimal than outlier and activation signals. The effect is also bounded: W8A4 benefits from neither SCN nor magnitude protection, whereas W8A8 restricted to W_d modules in SCN-bearing layers reproduces the SCN gain (+1.3494 vs. random P95 = 0.0141).

9 Limitations and Future Work

We study four instruction-tuned checkpoints (≤8B) from two families. Owing to compute limits, Llama-3.2-3B was evaluated on localization, bidirectional control, and full and sink-matched suppression only, and the costlier analyses (pathways, dissociation, rescue, quantization) cover subsets of the other models. Because the pipeline is model-agnostic and uses no benchmark data, extending it to more models, base checkpoints, and long contexts (we localize on 40-token passages) requires only compute. Our populations are sufficient for bidirectional control but not claimed to be minimal or unique, and zero-ablation may be off-distribution; graded doses, amplification, and matched random controls mitigate but do not eliminate this. Because the sham gate fails, rescue supports only the narrower claim that the sink does not explain most of the behavioral effect; tracing the non-sink route from SCNs to answers is the main open question. Finally, the quantization analysis (§8) is a single-model, per-tensor probe rather than a proposed method.

AI use statement

Generative AI tools were used to assist with drafting, formatting tables, fixing grammar and spelling, checking LaTeX formatting and internal cross-references. AI-assisted suggestions were treated as editorial or analytical input rather than as independent evidence. All reported experimental results, numerical values, statistical conclusions, and scientific claims were checked against the underlying experimental outputs and source materials by the authors. The authors reviewed and edited all AI-assisted text and take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts.

Reproducibility statement

All data splits, random seeds, neuron populations, matched random sets, intervention doses and decision rules were frozen before evaluation, and every test split was opened once (Section 3). Statistical procedures, including paired bootstrap resampling and Holm correction, are described with each experiment. An anonymized snapshot of the complete code repository, including experiment scripts, configuration files, frozen data manifests, and run receipts, is provided as a zipped supplementary attachment to the OpenReview submission.

References

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=tu4dFUsW5z.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3282. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/edbcb7583fd8921dad78adecfe06a99b-Abstract-Conference.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Enrique Queipo de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1734b19d9afe7d2c7f1154954eaf0d5a-Abstract-Conference.html.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c3ba4962c05c49636d4c6206a97e9c8a-Abstract-Conference.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Zichuan Fu, Wentao Song, Guojing Li, Yejing Wang, Xian Wu, Yimin Deng, Hanyu Yan, Yefeng Zheng, and Xiangyu Zhao. Attention needs to focus: A unified perspective on attention allocation, 2026. URL https://arxiv.org/abs/2601.00919.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f1b04face60081b689ba740d39ea8f37-Abstract-Conference.html.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs, 2024. URL https://arxiv.org/abs/2410.13835.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ZeI104QZ8I.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.41. URL https://aclanthology.org/2025.naacl-short.41/.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Siquan Li, Kaiqi Jiang, Jiacheng Sun, and Tianyang Hu. The structural origin of attention sink: Variance discrepancy, super neurons, and dimension disparity. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2605.06611.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3345. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/904e89bb4e632e75fb47f093b620b257-Abstract-Conference.html.

Yuval Ran-Milo. 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)*, pp. 70–89. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-short.8. URL https://aclanthology.org/2026.acl-short.8/.

Yuval Ran-Milo, Hila Ofek, and Shahar Mendel. 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), pp. 90–98. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-short.9. URL https://aclanthology.org/2026.acl-short.9/.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b2c4b7d34b3d96b9dc12f7bce424b7ae-Paper-Conference.pdf.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2242–2252. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.134. URL https://aclanthology.org/2024.emnlp-main.134/.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7101–7113. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.403. URL https://aclanthology.org/2024.emnlp-main.403/.

Zunhai Su and Kehong Yuan. Kvsink: Understanding and enhancing the preservation of attention sinks in KV cache quantization for LLMs. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=gIqb6zWZoO.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, Keyu Fan, Weihao Ye, Jing Xiong, Hui Shen, Chaofan Tao, Taiqiang Wu, Zhongwei Wan, Yulei Qian, Yuchen Xie, and Ngai Wong. Attention sink in transformers: A survey on utilization, interpretation, and mitigation, 2026. URL https://arxiv.org/abs/2604.10098.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026. URL https://arxiv.org/abs/2603.05498.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum Stuart McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp.

1288–1301. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.75. URL https://aclanthology.org/2024.findings-acl.75/.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11132–11152. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.765. URL https://aclanthology.org/2022.emnlp-main.765/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Itay Yona, Ilia Shumailov, Jamie Hayes, Federico Barbero, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. URL https://proceedings.mlr.press/v267/yona25a.html.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57659–57677. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24l.html.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A 'catch, tag, release' mechanism for embeddings. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-2786. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/77cd5a17edf97293eab706caf9bfee39-Abstract-Conference.html.

Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18179–18213. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.905. URL https://aclanthology.org/2026.findings-acl.905/.

A Protocol, Checkpoints, and Execution Environment

Checkpoints and protocol. All four models are instruction-tuned chat checkpoints loaded at pinned revisions in bfloat16 with eager attention, batch size 1, and deterministic serial execution. Model geometry, eligible coordinate universes, and the intervention coordinate system are reported in Appendix B. Sink layers are selected on the discovery split alone by the top-quartile rule with the absolute floor $S_l \geq 0.15$; the rule was registered together with a sink-negative exit that would have stopped the study for any model without a qualifying layer.

Stage map. The experiments are organized into nine frozen stages, each consuming the artifacts of earlier stages and writing append-only outputs: (i) sink-layer registration and neutral-corpus construction; (ii) attribution, measured reranking, and validation-set selection of the SCN populations; (iii) held-out loss-of-function confirmation (§3); (iv) gain-of-function confirmation on a fresh corpus (§3); (v) re-implementation and comparison with prior neuron definitions (Appendix F); (vi) query–key decomposition and live pathway restoration (§7); (vii) downstream evaluation under full and sink-matched SCN suppression (§4); (viii) task-neuron localization, crossed suppression, and answer-transition analysis (§5); and (ix) attention-sink rescue (§6) together with the downstream-outlier and quantization experiments (§8).

Corpora. Neutral text is drawn from OpenWebText at a pinned revision. Localization uses 40-token sequences in three disjoint splits of 100 examples (discovery, validation, test) drawn with seed 0. Every later stage draws from a window of the same corpus that is disjoint from all earlier windows, and each fresh corpus passes an overlap filter against all previously used text: exact normalized-string equality, exact token-sequence equality, and 13-gram containment are all rejected. Benchmark corpora are downloaded from their canonical sources at pinned revisions and never enter localization.

Execution. All runs execute serially on two NVIDIA RTX 4080 SUPER cards and one RTX 3090 (24 GB), with intervention hooks installed and removed per condition and an identity audit ($\alpha = 1$ reproduces baseline logits exactly) before each interpretable condition. Every stage writes an append-only audit record containing input digests, seeds, condition identifiers, and result hashes; replays must reproduce recorded results bit-for-bit for scalar statistics and within 10^{-12} for aggregated floating-point summaries.

B Model Geometry and Coordinate System

Table 8 gives the checkpoint geometry used throughout the interventions. Indices are zero-based. The eligible pool contains MLP intermediate coordinates in layers strictly upstream of at least one registered sink layer, rather than every neuron in the model. Its size is therefore the number of eligible MLP layers multiplied by the MLP intermediate width. A selected coordinate is multiplied at every sequence position immediately before the down projection, preserving its sign, input dependence, and natural relative magnitude when $\alpha > 0$.

Property	Qwen2.5-1.5B	Qwen2.5-3B	Llama-3.2-3B	Llama-3.1-8B
Decoder layers	28	36	28	32
MLP intermediate width	8,960	11,008	8,192	14,336
Query / KV heads	12 / 2	16 / 2	24 / 8	32 / 8
Head dimension	128	128	128	128
Eligible MLP layers	0–25	0–24	0–21	0–28
Eligible coordinates	232,960	275,200	180,224	415,744
Frozen SCNs	116	28	18	42
Registered sink layers	7	9	7	8

Table 8: **Checkpoint geometry.** Grouped-query attention is preserved in capture and restoration. Eligible layers depend on the deepest registered sink layer, not solely on model depth.

The frozen sink-layer sets are

$$\mathcal{L}_{\text{Q-1.5B}} = \{6, 14, 19, 23, 24, 25, 26\}, \tag{13}$$
$$\mathcal{L}_{\text{Q-3B}} = \{3, 4, 5, 6, 7, 8, 9, 20, 25\}, \tag{14}$$
$$\mathcal{L}_{\text{L-3B}} = \{2, 3, 16, 18, 20, 21, 22\}, \tag{15}$$
$$\mathcal{L}_{\text{L-8B}} = \{2, 19, 20, 23, 24, 25, 27, 29\}. \tag{16}$$

The head dimension is $d_k = 128$ in all four checkpoints. Under grouped-query attention, key–value groups are expanded to the full query-head count before computing the sink statistic of Eq. 3, so every query head contributes exactly once.

C Mathematical Properties of the Intervention Operators

This appendix collects identities that the experiments rely on. Each follows from the model and intervention definitions of Sections 2–8; none is an additional empirical claim. Where an identity can be checked against a reported table, we give the check. Throughout, n denotes the number of token positions of an input.

Causal reachability. The eligibility rule of Section 3 rests on the following ordering fact. [Within-pass causal ordering] In a decoder whose blocks compute attention before the MLP, rescaling coordinates of $a^{(l)}$ during a forward pass leaves every attention probability $A_{t,s}^{(j,h)}$ with $j \le l$ unchanged. The attention of block j reads only the residual stream entering block j. For $j \le l$ that stream is produced by blocks $0, \dots, j-1$, none of which contains the MLP of block l. The rescaled coordinates enter the stream only through $W_d^{(l)} a^{(l)}$, which is added after the attention of block l has been computed. Blocks $j > l$ read the updated stream and may change. The statement concerns one teacher-forced pass without a key–value cache. Hence a coordinate of $a^{(\ell)}$ can affect S_j only if $j > \ell$. The eligible universe therefore consists of the MLP intermediate coordinates in layers strictly upstream of the deepest registered sink layer; the resulting eligible layer ranges and coordinate counts are reported in Table 8. The same ordering is why each sink layer's share of the selection budget is filled only from MLP layers $\ell < j$.

The signed attribution score. Let a gate g multiply coordinate (l, i) at every position, $\tilde{a}_{t,i}^{(l)} = g\, a_{t,i}^{(l)}$, with the rest of the model fixed, and write $S_j(x; g)$ for the resulting sink of layer j on input x. The activation $a_{t,i}^{(l)}(x)$ is computed before the gate acts and does not depend on g, so the chain rule gives

$$\left.\frac{\partial S_j(x; g)}{\partial g}\right|_{g=1} = \sum_{t=0}^{n-1} a_{t,i}^{(l)}(x)\, \frac{\partial S_j(x)}{\partial a_{t,i}^{(l)}}. \tag{17}$$

The registered per-target form of Eq. 4 averages the summand over examples and positions, $I_{l,i\to j} = \mathbb{E}_x\big[\frac{1}{n}\sum_t a_{t,i}^{(l)}\, \partial S_j / \partial a_{t,i}^{(l)}\big]$. Every localization sequence has exactly $n = 40$ tokens, so $I_{l,i\to j} = \frac{1}{n}\,\mathbb{E}_x\big[\partial_g S_j|_{g=1}\big]$, and the positive factor $1/n$ changes neither signs nor rankings. If $|\partial_g^2 S_j(x; g)| \le M$ for every x and every g between 1 and α, Taylor's theorem with Lagrange remainder gives

$$\Big|\, \mathbb{E}_x\big[S_j(x; \alpha) - S_j(x; 1)\big] - (\alpha - 1)\, n\, I_{l,i\to j} \Big| \le \tfrac{1}{2} M (\alpha - 1)^2. \tag{18}$$

A positive score therefore predicts a sink decrease under a small reduction of α. It does not fix the effect at $\alpha = 0$, where the remainder need not be small, which is why the gradient shortlist is reranked by the measured effect of $a_i' = 0$, in line with the role of approximate attribution in causal localization (Syed et al., 2024; Kramár et al., 2024). For a set of coordinates scaled by one shared coefficient, the derivative at $g = 1$ is the sum of the individual gate derivatives, so first-order predictions add. Full-ablation effects need not add, and every population effect in the paper is measured directly.

Position-0 attention as anchor versus competition. [Anchor–competition form] Consider one attention row with finite logits z_s on the causally allowed keys $s \in \mathcal{K}$, where $\mathcal{K}$ contains key 0 and at least one key $r \ne 0$, and let $\tilde{p}_r = e^{z_r} / \sum_{s\in\mathcal{K}\setminus\{0\}} e^{z_s}$ be the conditional distribution over nonzero keys. Then

$$p_0 = \sigma(d), \qquad d = z_0 - \log \sum_{r\in\mathcal{K}\setminus\{0\}} e^{z_r}, \qquad \frac{\partial p_0}{\partial z_0} = p_0(1-p_0), \qquad \frac{\partial p_0}{\partial z_r} = -p_0(1-p_0)\,\tilde{p}_r. \tag{19}$$

Dividing numerator and denominator of $p_0 = e^{z_0}/\big(e^{z_0} + \sum_{r\neq 0} e^{z_r}\big)$ by e^{z_0} gives $p_0 = 1/(1 + e^{-d}) = \sigma(d)$. The derivatives follow from $\sigma' = \sigma(1-\sigma)$, $\partial d/\partial z_0 = 1$, and $\partial d/\partial z_r = -\tilde{p}_r$. With the post-RoPE logits of Section 7, $z_s = q^\top k_s/\sqrt{d_k}$, the contrast d has one gradient per factor of the decomposition:

$$\nabla_q d = \frac{1}{\sqrt{d_k}}\Big(k_0 - \sum_{r\neq 0} \tilde{p}_r\, k_r\Big), \qquad \nabla_{k_0} d = \frac{q}{\sqrt{d_k}}, \qquad \nabla_{k_r} d = -\frac{\tilde{p}_r\, q}{\sqrt{d_k}} \quad (r \neq 0). \tag{20}$$

A query change thus acts through the gap between the anchor key and the attention-weighted mean competing key, so $q^\top k_0$ alone cannot sign the change in p_0. For finite changes the logit change is bilinear, $\Delta z_s = \big((\Delta q)^\top k_s + q^\top \Delta k_s + (\Delta q)^\top \Delta k_s\big)/\sqrt{d_k}$, and softmax normalization couples all keys; both are reasons to decompose with exact factorial cells rather than first-order logit terms. Since S_j averages p_0 over heads and second-half queries, these statements hold row by row and average linearly.

Exact three-factor accounting. For one row, let $F(U)$, $U \subseteq N = \{Q, K_0, K_r\}$, be the position-0 probability when the factors in U take their suppressed-run values and the others their baseline values, all as post-RoPE tensors of the same input. We index cells by $\kappa(U) = \mathbf{1}[Q \in U] + 2\cdot\mathbf{1}[K_0 \in U] + 4\cdot\mathbf{1}[K_r \in U]$ and write F_κ, so F_0 is the baseline and F_7 the fully suppressed probability. Define the interaction dividends $D(\varnothing) = F(\varnothing)$ and, for $U \neq \varnothing$,

$$D(U) = \sum_{V\subseteq U} (-1)^{|U|-|V|} F(V), \qquad \text{so that} \qquad F(V) = \sum_{U\subseteq V} D(U) \tag{21}$$

by Möbius inversion. Let $\Delta = F(\varnothing) - F(N)$ be the net sink loss and $G_X = F(N\setminus\{X\}) - F(N)$ the gain from restoring factor X alone. [Dividend form of shares and restoration gains] For arbitrary finite cell values and $C_X = -\phi_X$ as in Eq. 11,

$$C_X = -\sum_{U\ni X} \frac{D(U)}{|U|}, \qquad \sum_{X\in N} C_X = \Delta, \tag{22}$$

$$G_X = -\sum_{U\ni X} D(U), \qquad G_X - C_X = -\sum_{U\ni X,\, |U|\geq 2} \frac{|U|-1}{|U|}\, D(U), \tag{23}$$

$$\sum_{X\in N} G_X = \Delta - \sum_{|U|\geq 2} (|U|-1)\, D(U). \tag{24}$$

In an ordering of N where the set P precedes X, the marginal contribution of X is $F(P\cup\{X\}) - F(P) = \sum_{U\ni X,\, U\setminus\{X\}\subseteq P} D(U)$ by Eq. 21. Thus $D(U)$ is credited to X exactly when X is the last member of U in the ordering, which happens with probability $1/|U|$ under a uniformly random ordering. Averaging gives $\phi_X = \sum_{U\ni X} D(U)/|U|$. Summing over X credits each nonempty U exactly once, so $\sum_X \phi_X = F(N) - F(\varnothing)$; negation gives Eq. 22. Next, $F(N\setminus\{X\}) - F(N) = \sum_{U\subseteq N\setminus\{X\}} D(U) - \sum_{U\subseteq N} D(U) = -\sum_{U\ni X} D(U)$. Subtracting Eq. 22 gives the second identity in Eq. 23, and summing G_X over X counts each U with weight $|U|$, which gives Eq. 24 because $\Delta = -\sum_{U\neq\varnothing} D(U)$. Written out with the subset weights $\frac{1}{3}, \frac{1}{6}, \frac{1}{3}$ of Eq. 11,

$$\phi_Q = \tfrac{1}{3}(F_1 - F_0) + \tfrac{1}{6}\big[(F_3 - F_2) + (F_5 - F_4)\big] + \tfrac{1}{3}(F_7 - F_6), \tag{25}$$

and analogously for K_0 (bit 2) and K_r (bit 4). The live restorations of Section 7 evaluate single cells: restoring queries, the anchor key, or the competing keys gives cells 6, 5, and 3, so $G_Q = F_6 - F_7$, $G_{K_0} = F_5 - F_7$, and $G_{K_r} = F_3 - F_7$; restoring all queries and keys gives cell 0, and restoring position-0 values leaves the cell at 7. Each G_X is therefore the one marginal term of ϕ_X in which X is suppressed last. It equals C_X when the dividends involving X vanish, in particular for additive F. Otherwise single restorations need not sum to Δ, and when one factor opposes the loss, restoring another can overshoot the baseline ($F_5 > F_0$ is possible although both are probabilities); a recovery above 100% is a ratio of gains, not a probability outside $[0, 1]$.

The live patches act on one sink layer at a time, are read at the patched layer only, and are required by the implementation to reproduce the cached cell exactly. The identities are linear in F, so they hold for the row averages behind Table 6, and the table satisfies them to rounding. The shares sum to 100.0% in every model ($33.9 + 34.3 + 31.8$, $-39.9 + 135.8 + 4.1$, $102.4 + 1.0 - 3.4$). The recoveries

sum to 110.2%, 118.4%, and 108.0%, so by Eq. 24 the weighted interaction $\sum_{|U|\geq 2}(|U|-1)\bar{D}(U)$ equals -10.2%, -18.4%, and -8.0% of the net loss: on balance, jointly suppressed factors lower the sink more than their separate effects add up to. In addition, G_X exceeds C_X in all nine rows of the table.

The rescue operator. Fix one edited row with finite current logits z on the allowed keys $\mathcal{K}$ (at least one nonzero), so that the current probabilities satisfy $p_0 \in (0,1)$, and let b be the baseline donor value. [Uniqueness and minimal divergence] For $b \in [0,1)$, the distribution p^{res} of Eq. 9 is the unique distribution q on $\mathcal{K}$ with $q_0 = b$ and $q_r/q_{r'} = e^{z_r - z_{r'}}$ for all nonzero $r, r' \in \mathcal{K}$. Among distributions on $\mathcal{K}$ with $q_0 = b$, it uniquely minimizes $D_{\mathrm{KL}}(q \,\|\, p)$, and for $b \in (0,1)$ it also uniquely minimizes $D_{\mathrm{KL}}(p \,\|\, q)$. For $b = 1$ the only admissible distribution is the point mass on key 0, which Eq. 9 also returns. The odds constraints force $q_r = c\, e^{z_r}$ on nonzero keys, and $\sum_{r\neq 0} q_r = 1 - b$ fixes c, giving $q_r = (1-b)\tilde{p}_r$. For the divergences write $q_r = (1-b)u_r$ with u a distribution on the nonzero keys, and $p_r = (1-p_0)\tilde{p}_r$. Grouping the sums gives

$$D_{\mathrm{KL}}(q \,\|\, p) = b \log \tfrac{b}{p_0} + (1-b) \log \tfrac{1-b}{1-p_0} + (1-b)\, D_{\mathrm{KL}}(u \,\|\, \tilde{p}), \tag{26}$$

$$D_{\mathrm{KL}}(p \,\|\, q) = p_0 \log \tfrac{p_0}{b} + (1-p_0) \log \tfrac{1-p_0}{1-b} + (1-p_0)\, D_{\mathrm{KL}}(\tilde{p} \,\|\, u), \tag{27}$$

with $0 \log 0 = 0$. The first two terms are fixed by b, and the last is nonnegative with equality only at $u = \tilde{p}$, with positive prefactors $1 - b$ and $1 - p_0$. Eq. 9 forms $\tilde{p}$ from logits rather than as $p_r/(1-p_0)$. The two agree algebraically, but the former stays well defined when $1 - p_0$ underflows. Rows whose requested change $b - p_0$ is exactly zero keep their original values, so the self-patch control is an exact identity by construction. [Matched allocation does not fix the head output] Let two runs $c \in \{1,2\}$ of the same head and query have the same position-0 probability b, values v_s^c, and nonzero-key conditionals $\tilde{p}^{\,c}$. Then

$$O^1 - O^2 = b\left(v_0^1 - v_0^2\right) + (1-b)\left(\mu^1 - \mu^2\right), \qquad \mu^c = \sum_{r\neq 0} \tilde{p}_r^{\,c}\, v_r^c. \tag{28}$$

Expand $O^c = \sum_s p_s^c v_s^c$ with $p_0^c = b$ and $p_r^c = (1-b)\tilde{p}_r^{\,c}$ for $r \neq 0$. Matching b constrains neither term. After rescue the recipient keeps its own suppressed values and its own conditional $\tilde{p}$, so exact restoration of the registered sink vector leaves both terms free, and first-half queries, continuation positions, and non-sink layers are not edited at all. This is the formal reason exact sink restoration and behavioral restoration can come apart (Section 6).

Sham capacity. A sham row moves mass to or from one hash-selected nonzero key r through $u = \mathrm{clip}(p_r + d, 0, 1 - p_0)$. The realized displacement $u - p_r$ equals the requested d exactly when $-p_r \leq d \leq 1 - p_0 - p_r$; otherwise it is truncated to $-p_r$ or to $1 - p_0 - p_r$. The capacity depends on the recipient row, not on the donor, so the same requested schedule can be realized well in one state and poorly in another. This capacity clipping is what fails the 90/90 adequacy gate (Table 17).

Layer-matched overlap null. The null of Section 5 holds the SCN set C fixed and redraws a task-sized set layer by layer: in layer l it draws t_l coordinates (the task population's own count) without replacement from all d_l coordinates of that layer, independently across layers. With $c_l = |C_l|$ and $X = \sum_l X_l$, the per-layer intersections are hypergeometric,

$$X_l \sim \mathrm{Hypergeometric}(d_l, c_l, t_l), \qquad \mathbb{E}[X] = \sum_l \frac{c_l t_l}{d_l}, \qquad \mathrm{Var}[X] = \sum_l t_l\, \frac{c_l}{d_l}\left(1 - \frac{c_l}{d_l}\right)\frac{d_l - t_l}{d_l - 1}. \tag{29}$$

The moments follow from writing X_l as a sum of c_l indicators with common mean t_l/d_l and pairwise covariances fixed by sampling without replacement; independence across layers makes the variances add. For reference, in Qwen2.5-1.5B ($d_l = 8{,}960$) a uniform draw over the whole universe shares $116^2/232{,}960 \approx 0.058$ coordinates with the SCN set in expectation, whereas two sets with the SCN layer histogram share $\sum_l c_l^2/8{,}960 = 1{,}430/8{,}960 \approx 0.160$. Shared depth preference alone nearly triples chance overlap, which is why the null matches layers. An intersection of one or two coordinates can exceed such a small expectation while remaining far from identity under the registered population-identity criteria.

Resolution of resampling p-values. Every add-one empirical p-value computed from K resamples or controls satisfies $p \geq 1/(K+1)$, and after Holm correction over m tests the smallest attainable value is $m/(K+1)$. With $K = 10{,}000$ this is 0.0004, 0.0009, and 0.0011 for families of $m = 4$, 9,

and 11 tests: the values reported for the overlap nulls and the outlier bridge, the Qwen2.5-3B rescue family, and the matched-suppression BLEnD contrast. Such values mean that no resample reached the observed statistic; they are floors, not estimates of a smaller p. With 20 matched-random controls the floor is $1/21 \approx 0.0476$ (Appendix F).

Accuracy change and answer transitions. For conditions u and v with correctness indicators $c_u(x) = \mathbf{1}[\hat{y}_u(x) = y^*(x)]$, let $\mathrm{CW} = P(c_u = 1, c_v = 0)$ and $\mathrm{WC} = P(c_u = 0, c_v = 1)$. Since $A_v = \mathrm{CC} + \mathrm{WC}$ and $A_u = \mathrm{CC} + \mathrm{CW}$,

$$A_v - A_u = \mathrm{WC} - \mathrm{CW}, \qquad \mathrm{Flip} = \mathrm{CW} + \mathrm{WC} + \mathrm{WW}^{\neq} \;\geq\; \mathrm{CW} + \mathrm{WC} \;\geq\; |A_v - A_u|. \tag{30}$$

With one correct label and two options, any changed prediction changes correctness, so $\mathrm{WW}^{\neq} = 0$ and $\mathrm{Flip} = \mathrm{CW} + \mathrm{WC}$, as for WinoGrande. With more options, or with grouped signatures such as the four decisions of CulturalBench Hard, $\mathrm{WW}^{\neq}$ can be positive, so equal accuracy does not imply unchanged predictions. The identity also fixes how the two transition tables relate. Table 4 takes the intact baseline as u for both the target arm T and its matched control R, so

$$A_T - A_R = \big(\mathrm{WC}_T - \mathrm{CW}_T\big) - \big(\mathrm{WC}_R - \mathrm{CW}_R\big), \tag{31}$$

and all eight rows reproduce the sink-only column of Table 3 to rounding. For example, $(4.23 - 6.29) - (1.45 - 1.09) = -2.42$ against -2.418 (Qwen2.5-1.5B, CulturalBench Hard), and $(11.02 - 9.10) - (1.84 - 1.21) = +1.29$ against $+1.294$ (Qwen2.5-3B, BLEnD). The registered arm-versus-control transitions of Appendix J instead take $u = R$ and $v = T$ directly. Both frames satisfy Eq. 30, but their individual cells need not coincide.

Error of the protected linear operator. In this paragraph b denotes the bit width. For the symmetric max-absolute quantizer of Section 8, $|X/s| \leq q_{\max}$ elementwise and $q_{\max}$ is an integer, so rounding never leaves $[-q_{\max}, q_{\max}]$ and the clamp is inactive. Nearest rounding then gives $|Q_b(X) - X| \leq s/2$ elementwise, ignoring floating-point roundoff. Activation scales are computed per passage over all token positions, and weight scales per matrix. For a down-projection with protected columns P and unprotected columns U, the protected operator and the FP32 operator on the same input are

$$\hat{y} = Q_b(W_U)\, Q_b(x_U) + W_P x_P + b_{\mathrm{lin}}, \qquad y = W_U x_U + W_P x_P + b_{\mathrm{lin}}, \tag{32}$$

with scales s_W and s_x computed on U only. With $E_W = Q_b(W_U) - W_U$ and $E_x = Q_b(x_U) - x_U$, the protected term and the bias cancel, and for every output row r and token

$$\hat{y} - y = E_W x_U + W_U E_x + E_W E_x, \qquad |\hat{y}_r - y_r| \leq \frac{s_W}{2}\,\|x_U\|_1 + \frac{s_x}{2}\,\|W_{r,U}\|_1 + \frac{|U|\, s_W s_x}{4}. \tag{33}$$

Protection acts only at down-projection inputs; every other operator is quantized identically in the naive and protected conditions. Removing coordinates never raises a scale, since a maximum over U cannot exceed the maximum over all coordinates, and the inequality is strict exactly when every coordinate attaining the maximum lies in P. This is how protection can shrink every term of the bound. For activations, the setter rate measures how often it happens: an SCN coordinate is the all-token argmax of the down-projection activation in 22.2222% of SCN-bearing layer/example pairs, against 49.0278% for the layer-matched magnitude set, 83.4500% for the global-magnitude set, 5.7500% for the super-neuron proxy, and a random median of 0.0278%. Eq. 33 is an upper bound, not a predictor of realized cross-entropy, and it does not imply that SCNs minimize error; the magnitude sets indeed recover more (Table 7). Finally, activation-only protection adds $\big(Q_b(W_P) - W_P\big)x_P$ with W_P quantized under its own scale, and weight-only protection adds $W_P\big(Q_b(x_P) - x_P\big)$ with x_P quantized under its own scale, while both keep the U scales computed without P. Neither is a pure single-operand test, so their cross-entropies (3.615326 and 3.630764, against 3.615055 for joint protection) are descriptive.

D Localization and Loss of Function

Per-layer membership. The Qwen2.5-1.5B population of 116 neurons distributes across eligible layers 0–25 as

$$[3, 4, 2, 1, 2, 17, 0, 0, 0, 0, 0, 0, 0, 16, 1, 1, 1, 3, 15, 6, 0, 4, 20, 4, 11, 5],$$

that is, concentrated immediately upstream of the registered sink layers, with layer 22 contributing the largest single block (20 neurons) and layers 6, 13, and 18 contributing 17, 16, and 15. The same

largest-remainder allocation produces the 28-, 18-, and 42-neuron populations of the other three models; all three are correspondingly sparser per layer, and all four histograms are reused verbatim for every matched random control in the paper.

Screening and selection. Attribution scores of Eq. 4 are computed per target sink layer on the discovery split and averaged over examples and positions, after which the top candidates are reranked by their measured single-neuron sink effect under $a'_i = 0$ on 20 discovery examples. Selection at each registered fraction then proceeds by largest-remainder allocation across sink layers in proportion to baseline S_l. Validation freezes the sink layers, the membership at the chosen k, the 20 random controls, and the dose grid before the test split is unlocked.

Dose monotonicity and the low-distortion search. The loss-of-function criterion requires Spearman $\rho \geq 0.8$ over $\alpha \in \{1, 0.75, 0.5, 0.25, 0\}$; all four models achieve $\rho = 1.0$, i.e., perfectly ordered sink reduction across the dose grid. A registered search for a low-distortion sub-population (a smaller k that retains the sink effect at reduced output distortion) qualified no candidate for Qwen2.5-3B, Llama-3.2-3B, or Llama-3.1-8B on the validation split, so the registered k was carried forward unchanged in every model and no claim of behavioral inertness is made anywhere in the paper.

E Gain of Function and Health Metrics

The gain-of-function stage reuses the frozen populations, frozen controls, and frozen layer histograms on a fresh 100-example neutral corpus that passes the three overlap filters of App. A against the localization corpus. The decision dose is $\alpha = 1.5$; the registered grid is $\{1, 1.10, 1.25, 1.50\}$.

Table 9: Amplification dose response (relative sink gain over baseline, %) and output-distortion health metrics at the decision dose $\alpha = 1.5$ on the fresh corpus.

	RSG at α				Health at $\alpha = 1.5$		
Model	1.10	1.25	1.50	Baseline sink	ΔCE	forward KL	top-1 flips
Qwen2.5-1.5B	+0.753	+1.664	+3.110	0.618	+0.201	0.278	7.56%
Qwen2.5-3B	+0.173	+0.485	+1.003	0.672	+0.066	0.105	6.36%
Llama-3.2-3B	+1.601	+3.783	+6.399	0.841	+0.072	0.096	17.33%
Llama-3.1-8B	+1.216	+2.855	+5.136	0.843	+0.116	0.179	14.10%

Per-layer positivity holds everywhere: all 7, 9, 7, and 8 registered sink layers move upward at $\alpha = 1.5$ and downward at $\alpha = 0$, so neither aggregate is produced by a single anomalous layer. Figure 2 shows the full bidirectional dose response for three models with the matched-random bands.

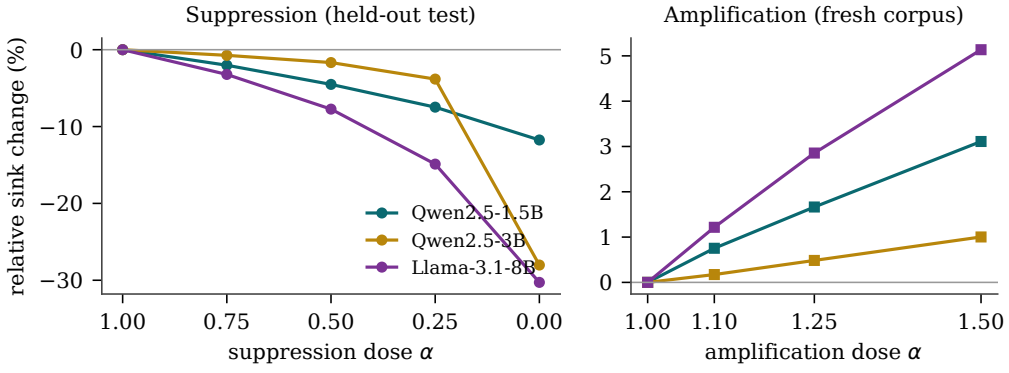

Figure 2: Bidirectional dose response of the frozen populations. Sink attention decreases monotonically as the scaling coefficient falls from 1 to 0 and increases monotonically as it rises to 1.5, in each plotted model (Llama-3.2-3B values in Tables 1 and 9); matched-random controls with identical size and per-layer counts stay flat.

F Differentiation from Prior Neuron Populations

Comparator implementations. Every comparator is re-implemented on the identical Qwen2.5-1.5B checkpoint, in the identical coordinate system (MLP intermediate coordinates entering W_d), and frozen before any common outcome is inspected.

- *Attention-deactivation neurons* (Gurnee et al., 2024): scored by the effect of a neuron's activation on downstream position-0 attention mass, adapted to grouped-query attention by expanding key–value groups to full head count before scoring and by applying rotary position embeddings natively, so the comparator sees exactly the tensors the model uses.
- *Repeated-token sink neurons* (Yona et al., 2025): scored by the product of activation magnitude and the norm of the corresponding down-projection column, on a repetition-structured corpus.
- *Structural super-neuron proxy* (Li et al., 2026): a proxy combining activation magnitude, down-projection column norm, and residual-stream write alignment, since the original definition is architecture-specific.
- *Positional neurons* (Voita et al., 2024): selected by variance of activation explained by token position on the neutral corpus.
- *Repetition neurons* (Hiraoka & Inui, 2025): selected by differential activation on a repetition dataset against matched non-repetitive text.
- *Task-specific neurons* (Song et al., 2024): gradient-selected on a sentiment classification task at a pinned dataset revision.

Acquisition hierarchy and auxiliary corpora. Auxiliary datasets are acquired by a registered hierarchy that prefers the canonical source at a pinned revision, falls back to a mirrored revision, and, failing both, to a deterministic constructed corpus; every fallback is recorded in the audit log. The repetition dataset is constructed from a reservoir of neutral passages in window $[3600, 7000)$ with seeds in $[43010, 143010)$, splits of 128 training and 64 evaluation passages, a repeated 10-gram appearing at least 3 times, and a ± 30-token context window. The sentiment corpus is SST-2 at revision `8d51e7e4887a4caaa95b3fbebbf53c0490b58bbb` with 128 and 64 examples and a 5% native selection rate. The neutral reservoir supplies 256, 64, 64, and 200 passages to the positional, repetition, proxy, and panel stages respectively.

Native versus matched populations. Each comparator is evaluated twice. The *native* variant uses that method's own threshold and yields its own population size, and is used only to verify that our re-implementation reproduces the method's own phenotype (positional neurons are positionally selective, repetition neurons fire on repeats, and so on). The matched variant is truncated or extended to exactly 116 neurons with the SCN per-layer histogram. Its interpretation is summarized in Table 10.

Panel and inference. The causal panel evaluates suppression ($\alpha = 0$) and amplification ($\alpha = 1.5$) for every population on a fresh 200-example neutral test set, each against 20 layer- and count-matched random controls drawn with seeds 43000–43019; permutation nulls use seed 43021 and bootstraps seed 43022. Specificity p-values are one-sided with the empirical floor $p \geq 1/21 = 0.0476$, which is why bidirectional P95 exceedance can hold without family-wise significance across the 22 registered specificity tests.

Table 10: Identity requires all four registered population-identity criteria; functional convergence requires a resolved common effect in the same direction.

Population	Identity	Same dir.	Resolved	Registered interpretation
Sink neurons		✓	✓	convergent function, disjoint set
Attention deactivation		✓	✓	weak convergence, distinct population
Super-neuron proxy		partial	partial	suppression-only convergence
Positional			✓	opposite-direction population
Repetition		n/a		no resolved common effect
Task-specific (gradient)			✓	opposite-direction population
Task populations (4 benchmarks)		n/a	n/a	near-disjoint membership

Task-population overlap. Task-neuron populations are also near-disjoint from SCNs; full membership, graded-overlap, and crossed-suppression analyses are reported in Appendix J.

G Query and Key Pathways

Corpus and conditions. The decomposition uses a fresh neutral corpus from window $[2400, 3600)$ with 200 examples of 40 tokens. The Qwen2.5-3B run reuses the same 200 items token-identically; the Llama-3.1-8B run draws an independent sample because its tokenizer differs. Conditions are the eight factorial cells of the three-factor design, evaluated exactly from cached post-RoPE tensors, plus four live restoration conditions (query only, anchor key only, competing keys only, all query and key), an amplification endpoint at $\alpha = 1.5$, a value-restoration control, and a full query–key restoration audit. The factorial cells are indexed by a three-bit code in which the first bit marks queries, the second the anchor key, and the fourth competing keys, with the all-baseline and all-suppressed cells as endpoints; both endpoints are verified against directly recomputed attention.

Audits. Full query and key restoration reproduces the baseline attention matrix to within floating-point tolerance on every example, head, and query; restoring position-0 *values* leaves attention exactly unchanged, confirming that the entire effect is carried by attention probabilities. The exact-closure identity in Eq. 22 is verified numerically per row with a maximum absolute residual of 10^{-12}. Forward-pass counts are 12,000 for each Qwen model and 13,000 for Llama-3.1-8B (65 passes per example).

Amplification reversal. At $\alpha = 1.5$ every supported suppression pathway reverses sign, as required by the registered secondary check: Qwen2.5-1.5B gives -0.0098, -0.0019, and -0.0082 for queries, anchor key, and competing keys. The one exception is Llama-3.1-8B's anchor-key term, which remains positive at $+0.004694$ [0.004085, 0.005318] on the Shapley endpoint and $+0.0016$ [0.0016, 0.0017] live; the registered rule treats this as a failed secondary check that does not revoke the primary support for that pathway.

Depth structure. In Qwen2.5-1.5B the competing-key pathway dominates the earliest sink layer (layer 6, contribution 0.0557) while the anchor-key pathway grows with depth and dominates layer 26 (0.0541); layers 14 and 19 are query-leaning. In Qwen2.5-3B, layers 3 and 5 show mildly positive query contributions, layers 6–9 carry the opposing query effect (layer 9: -0.1588), and layers 20 and 25 carry most of the competing-key term. In Llama-3.1-8B the query pathway dominates layers 19, 24, and 25, with layer 25 alone carrying a net loss of 0.5132 of which 0.5109 is query-borne; layer 2 shows the opposite pattern, the anchor-key sign varies by layer, and the competing-key sign flips at layers 20, 23, and 29.

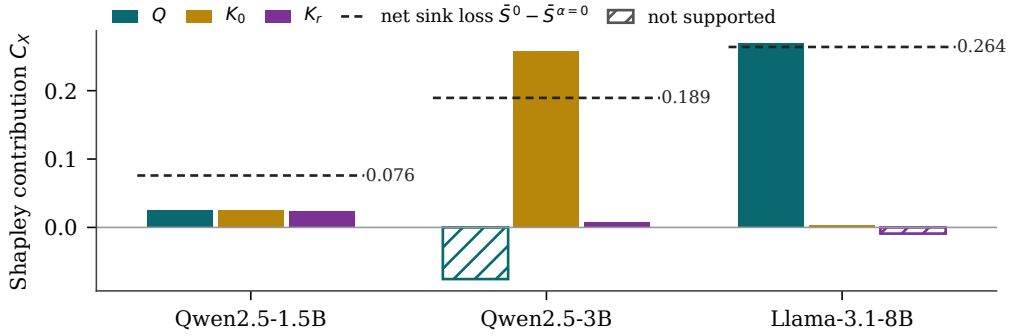

Figure 3: Exact Shapley decomposition of the suppression-induced sink loss into query, anchor-key, and competing-key contributions, per model. Contributions sum exactly to the net loss (Eq. 22); negative bars are opposing pathways.

H Full-Suppression Behavioral Suite

Suite and scoring. Twelve benchmarks with 27,086 baseline examples in total are evaluated under four conditions per model: baseline, target suppression at $\alpha = 0$, target at $\alpha = 0.5$, and the layer- and count-matched random control at $\alpha = 0$.

Table 11: Benchmark suite, example counts, and scoring protocols.

Benchmark	N	Shots	Scoring
MMLU (subsample)	3,000	5	letter-option likelihood
ARC-Challenge	1,172	0	option likelihood
CulturalBench Easy	1,227	0	option likelihood
CulturalBench Hard	1,227	0	all four binary items correct
GSM8K	1,319	5	exact match on final answer
BLEnD	3,200	0	short-answer match
HellaSwag	3,000	0	character-length-normalized likelihood
WinoGrande	1,267	0	suffix likelihood after the blank
PIQA	1,838	0	character-length-normalized likelihood
CommonsenseQA	1,221	0	option likelihood
Global-MMLU-Lite	6,000	0	15 languages, 200 + 200 items each
NormAd	2,615	0	country-only condition

The MMLU subsample of 3,000 items is drawn from the 14,042-item parent set by largest-remainder allocation across subjects with SHA-256 ordering at seed 0, so the subject histogram is preserved and the draw is reproducible. Viability floors exclude two model–task pairs from interpretation under both full and matched suppression: Qwen2.5-1.5B on GSM8K reaches only 2.578% exact match at baseline, below the registered 5% floor, and Qwen2.5-3B falls below the floor on NormAd. A pre-standardization caveat applies to two rows: the MMLU results were produced on the 14,042-item parent set and GSM8K was revised mid-study; both are reported with their own baselines and are never pooled with the standardized runs.

Table 12: Full-suppression results. Each cell is accuracy drift versus baseline (pp) / task-prompt RSR (%) / target-minus-random specificity (pp). † marks nominal CIs excluding zero. All twelve Llama-3.1-8B and Llama-3.2-3B specificity contrasts survive Holm within their respective model families.

Benchmark	Qwen2.5-1.5B	Qwen2.5-3B	Llama-3.2-3B	Llama-3.1-8B
MMLU	–0.520 / 7.450 / –0.356	–2.293 / 18.402 / –2.121	–25.367 / 53.183 / –25.333†	–19.180 / 26.727 / –19.163†
ARC-Challenge	–0.171 / 6.507 / +0.043	–1.451 / 15.021 / –1.280	–17.406 / 37.482 / –17.321†	–41.894 / 24.061 / –41.809†
CulturalBench Easy	–1.386 / 8.192 / –1.223	–6.031 / 19.117 / –5.786	–15.892 / 37.227 / –15.648†	–26.895 / 28.117 / –26.813†
CulturalBench Hard	–2.363 / 8.192 / –2.037	–7.335 / 19.117 / –7.172	–18.745 / 36.342 / –19.234†	–17.685 / 28.117 / –17.604†
GSM8K	n/a (floor)	–9.128 / 14.086 / –9.098	–64.291 / 51.610 / –64.215†	–71.797 / 21.744 / –71.949†
BLEnD	+3.000 / 13.153 / +3.281†	+2.312 / 30.652 / +2.312†	–32.156 / 40.642 / –32.313†	–23.156 / 42.077 / –23.031†
HellaSwag	–0.633 / 9.104 / –0.500	–2.100 / 21.243 / –1.933	–15.667 / 34.799 / –15.600†	–12.267 / 30.049 / –12.200†
WinoGrande	+0.947 / 9.544 / +1.026	–0.631 / 22.204 / –0.552	–10.576 / 31.986 / –10.418†	–9.077 / 31.128 / –8.998†
PIQA	–0.435 / 8.860 / –0.272	–1.088 / 20.597 / –0.925	–11.643 / 27.811 / –11.643†	–8.052 / 29.115 / –7.943†
CommonsenseQA	–2.359 / 8.596 / –2.211	–3.030 / 19.991 / –2.866	–25.307 / 34.352 / –25.061†	–27.273 / 28.283 / –27.191†
Global-MMLU-Lite	–1.617 / 9.917 / –1.450	–3.883 / 23.062 / –3.717	–20.571 / 41.938 / –20.696†	–22.117 / 32.339 / –22.033†
NormAd	–0.459 / 8.738 / –0.306	n/a (floor)	–27.036 / 36.037 / –26.539†	–22.600 / 28.590 / –22.330†

I SINK-MATCHED SUPPRESSION

Calibration. The matching coefficient α^* is calibrated on a fresh 300-example neutral corpus disjoint from all prior windows under the three overlap rules of App. A. The search anchors at the registered dose grid, brackets the target neutral RSR of 10%, and bisects, accepting the first coefficient satisfying $|\mathrm{RSR}(\alpha^*) - 0.10| \leq 0.005$ with at most 17 forward evaluations and a monotonicity tolerance of 0.002. What is matched is exactly one quantity: neutral-corpus relative sink reduction. Population size, per-layer counts, the coefficient itself, output KL, top-1 flip rate, and task-prompt RSR are *not* matched, and differences in those quantities are therefore not evidence about the matched comparison.

Task-prompt RSR at the matched coefficient ranges over 5.93–11.81%, 6.35–12.94%, 4.19–8.31%, and 7.54–13.99% for the four models, and accuracy drift over -2.53 to $+3.06$, -4.65 to $+2.75$, -4.85 to $+0.20$, and -29.34 to -1.74 pp. Registered intervals for the two resolved Qwen2.5-1.5B contrasts are BLEnD $+3.562$ [2.234, 4.937] at $p = 0.001200$ and Global-MMLU-Lite -1.250 [–2.071, –0.429] at $p = 0.035196$; CulturalBench Hard reaches $p_{\mathrm{Holm}} = 0.080992$ and is therefore reported with a dagger rather than in bold. The single resolved Llama-3.2-3B contrast is GSM8K -4.701

Table 13: Calibration outcomes on the fresh neutral corpus. Error is achieved minus target neutral RSR (pp).

Model	α^*	achieved neutral RSR	error
Qwen2.5-1.5B	0.0625	10.200%	+0.200
Qwen2.5-3B	0.078125	9.593%	−0.407
Llama-3.2-3B	0.671875	10.033%	+0.033
Llama-3.1-8B	0.40625	9.983%	−0.017

Table 14: Matched-suppression results. Each cell is accuracy drift versus baseline / target-minus-random specificity (pp) at α^*.

Benchmark	Qwen2.5-1.5B	Qwen2.5-3B	Llama-3.2-3B	Llama-3.1-8B
MMLU	−0.367 / −0.300	−1.367 / −1.333	−0.567 / −0.567	−6.600 / −6.567
ARC-Challenge	−0.085 / −0.085	−0.597 / −0.512	−1.280 / −1.109	−13.994 / −13.908
CulturalBench Easy	−0.815 / −0.815	−1.793 / −1.711	+0.081 / −0.081	−5.461 / −5.379
CulturalBench Hard	−2.526 / −2.608	−4.564 / −4.482	−0.570 / −0.489	−7.009 / −7.009
GSM8K	n/a (floor)	−0.303 / −0.303	−4.852 / −4.701	−29.492 / −29.492
BLEnD	+3.562 / +3.562	+2.312 / +2.312	−0.094 / −0.031	−14.750 / −14.688
HellaSwag	−0.100 / −0.067	−1.067 / −1.033	+0.200 / +0.267	−2.100 / −2.067
WinoGrande	+0.947 / +0.947	−0.631 / −0.631	−2.131 / −2.131	−1.658 / −1.579
PIQA	+0.109 / +0.109	−0.816 / −0.816	−0.326 / −0.381	−1.850 / −1.795
CommonsenseQA	0.000 / 0.000	−0.819 / −0.819	−0.491 / −0.246	−6.143 / −6.061
Global-MMLU-Lite	−1.250 / −1.250	−1.696 / −1.696	−0.911 / −0.893	−4.633 / −4.600
NormAd	−0.459 / −0.459	n/a (floor)	−0.841 / −0.880	−1.874 / −2.141

[−6.899, −2.502] at $p_{\text{Holm}} = 0.0024$; WinoGrande and NormAd exclude zero nominally but not after Holm correction. Figure 4 plots the full specificity profile.

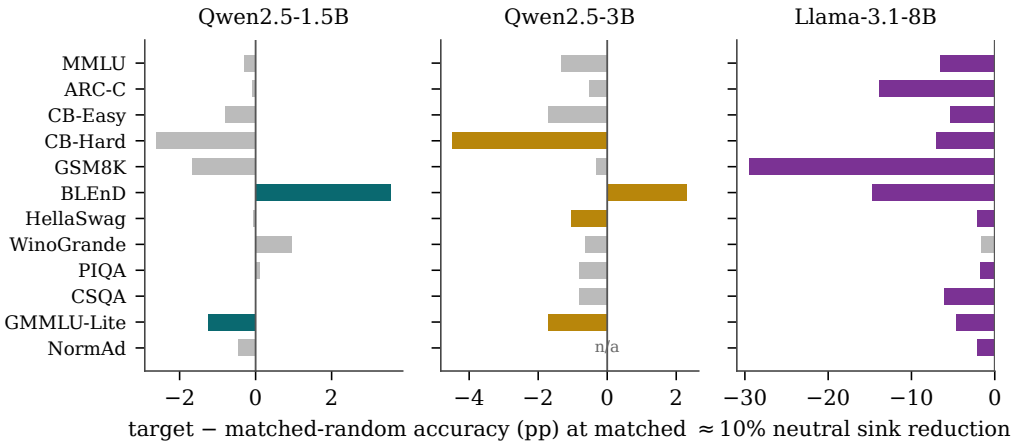

Figure 4: Target-minus-random accuracy at matched neutral sink reduction, by benchmark and model (Llama-3.2-3B values in Table 14). Positive values indicate the target population is less damaging than its dose-matched random control.

J Task-Neuron Overlap, Crossed Suppression, and Transitions

Single-model design. Task populations for MMLU, ARC-Challenge, CulturalBench Easy, and CulturalBench Hard are localized on Qwen2.5-1.5B with cluster-safe splits: MMLU 2,000 discovery items against 12,042 held out, ARC 400 against 772, and CulturalBench Easy and Hard 400 against 827 with shared question identifiers so that no held-out item shares a cluster with a discovery item. The measured reranking pool contains 2,330 shortlisted coordinates screened on 20 examples each.

Graded overlap at shortlist depths 500, 1,000, and 2,330 gives 3/23/136 (MMLU), 15/35/128 (ARC), 9/25/124 (CulturalBench Easy), and 10/29/120 (CulturalBench Hard), all at $p_{\text{Holm}} = 0.0004$. Rank

Table 15: Primary membership overlap between the 116-neuron SCN population and each 116-neuron task population, with layer-matched permutation p-values (10,000 draws, Holm-corrected within the family).

Task population	Overlap	p_{Holm}	Registered status
MMLU	1/116	0.0958	underpowered (< 10)
ARC-Challenge	2/116	0.0033	underpowered (< 10)
CulturalBench Easy	3/116	0.0004	underpowered (< 10)
CulturalBench Hard	1/116	0.0958	underpowered (< 10)

correlations over the whole universe are 0.0030, 0.0070, 0.0072, and -0.0008; restricted to the measured shared pool they are -0.0070, $+0.1013$, -0.0553, and $+0.0002$.

Table 16: Single-model crossed suppression on held-out clusters (Qwen2.5-1.5B). Accuracies are baseline $\rightarrow$ intervened; Δ is versus that arm's own layer- and count-matched control.

	Task-only $T \setminus C$			Sink-only $C \setminus T$	
Task	accuracy	Δ (pp)	sink RSR	Δ (pp)	sink RSR
MMLU	58.994 $\rightarrow$ 47.409	–11.585	+0.985%	–0.407	7.52%
ARC-Challenge	74.223 $\rightarrow$ 39.249	–34.974	+4.858%	+0.518	6.93%
CulturalBench Easy	61.427 $\rightarrow$ 46.675	–14.752	+2.31%	–1.330	7.63%
CulturalBench Hard	$\rightarrow$ 0.000	–23.337	+0.21%	–1.088	8.42%

Cross-model extension. The extension shortlists approximately 1% of each universe (2,330, 2,752, and 4,157 coordinates) and uses registered cluster units: 400 question identifiers for CulturalBench Hard, 100 source questions for Global-MMLU-Lite, 80 templates for BLEnD, and 400 rows for WinoGrande. Yoruba is excluded from Global-MMLU-Lite because its baseline accuracy of 29.5% falls below the registered 30% floor, leaving 14 languages and a held-out set of 300 source questions expanding to 4,200 scored rows. Graded overlaps at depths 582/1165/2330 (Qwen2.5-1.5B) are 10/37/119, 11/30/118, 12/29/95, and 9/35/119; at 688/1376/2752 (Qwen2.5-3B) they are 9/24/83, 3/13/48, 8/24/74, and 10/22/77. All contrasts reach $p_{\mathrm{Holm}} = 0.0004$ except Global-MMLU-Lite at depth 688 (0.086991). Whole-universe rank correlations lie in $[-0.00094, 0.00202]$ for Qwen2.5-1.5B and $[-0.0058, 0.00064]$ for Qwen2.5-3B; on the measured shared pool they are -0.1703, -0.0148, -0.0580, -0.0936 and $+0.0481$, -0.1102, -0.0622, $+0.2444$.

Prediction transitions. Table 4 in the main text reports the five-way partition for the sink-population arm. The corresponding two-column summary for the task-only arm, reported as changed-prediction rate for target and matched-random control, is: Qwen2.5-1.5B $22.612/0.000$, $13.857/8.881$, $33.236/2.004$, and $4.614/4.614$ for CulturalBench Hard, Global-MMLU-Lite, BLEnD, and WinoGrande; Qwen2.5-3B $28.416/0.121$, $1.738/1.667$, $18.580/2.463$, and $2.537/2.537$. The sink-only arm gives $6.167/3.748$, $4.810/3.548$, $5.846/8.559$, $4.729/5.882$ and $3.507/3.265$, $12.714/7.643$, $9.353/10.647$, $2.999/2.999$ under the same two-column summary, which aggregates differently from the five-way partition of Eq. 8 and should not be compared cell-by-cell with it.

Output-channel caveat. CulturalBench Hard scores an item as correct only when all four binary sub-items are correct. Under task-only suppression Qwen2.5-1.5B produces strictly 0% accuracy, and in Qwen2.5-3B 775 of 827 items collapse to the all-false response pattern, so accuracy on this benchmark partly measures an output-channel collapse rather than graded knowledge loss. Registered arms with fewer than 10 shared neurons are declared underpowered and are not interpreted causally; the Llama-3.1-8B extension arm was registered but not completed.

K ATTENTION RESCUE

Conditions. Fifteen conditions are evaluated per input: baseline; full suppression; rescue; the self-patch identity control (the model's own donor, which must reproduce the baseline exactly); five sham rescues with seeds 0–4 applied to the suppressed model; the matched sham family applied to the healthy model; the reverse imposition of the suppressed allocation on the intact model; and the

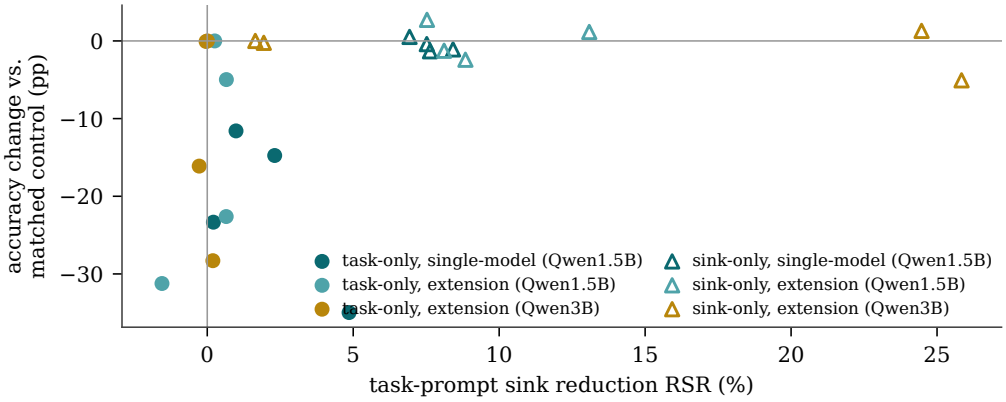

Figure 5: Dissociation between sink control and task performance. Task-only suppression moves accuracy with little sink movement; sink-only suppression moves the sink with small, sign-varying accuracy effects.

registered donor-binding controls, which verify that each rescued row receives the baseline value computed from the identical input, layer, head, and query.

Estimands and families. L, G, and J follow Eq. 10. Holm families contain 12 tests for Qwen2.5-1.5B (four tasks $\times$ three estimands) and 9 for Qwen2.5-3B (three tasks $\times$ three estimands). Loss-fraction ratios G/L are reported only when the family is stable, defined as a positive lower bound on L and at least 97.5% positive bootstrap denominators; unstable ratios are withheld.

Task protocols. MMLU uses the 14,042-item parent set at five shots; ARC-Challenge and both CulturalBench splits are zero-shot; GSM8K is excluded by the viability floor. The Qwen2.5-3B arm covers CulturalBench Easy, CulturalBench Hard, and a 5,600-row Global-MMLU-Lite subset.

Sham adequacy. The registered gate requires a sham family to realize at least 90% of the requested displacement on at least 90% of active rows. In Qwen2.5-1.5B the realized fractions are 0.884–0.930 (CulturalBench Easy), 0.891–0.953 (CulturalBench Hard), 0.828–0.879 (MMLU), and 0.899–0.945 (ARC), with row coverage 0.753–0.887. In Qwen2.5-3B the gate fails on every task in both the healthy and suppressed states.

Table 17: Qwen2.5-3B sham adequacy. Displacement and row-coverage ranges across the five sham seeds, with the registered 90/90 gate outcome.

Task	Sham family	displacement	row coverage	gate
CulturalBench Easy	healthy	80.19–80.23%	68.51–68.55%	
CulturalBench Easy	suppressed	97.16–97.17%	90.29–90.32%	✓
CulturalBench Hard	healthy	74.55–74.57%	62.91–62.92%	
CulturalBench Hard	suppressed	96.44–96.45%	89.79–89.80%	
Global-MMLU-Lite	healthy	68.47–68.49%	57.83–57.84%	
Global-MMLU-Lite	suppressed	96.31–96.32%	88.08–88.09%	

All sham conditions pass the registered sink-control check, confirming that the failures are capacity-clipping failures of the displacement target rather than implementation errors. Recovery restricted to examples lost under suppression is 21.43%, 24.69%, 42.15%, and 28.00% for the four Qwen2.5-1.5B tasks and 27.5%, 44.0%, and 26.9% for the three Qwen2.5-3B tasks. Reverse imposition of the suppressed allocation on the intact model changes accuracy by -0.326, -0.244, $+0.043$, and $+0.256$ pp in Qwen2.5-1.5B, whereas in Qwen2.5-3B it leaves Global-MMLU-Lite at 56.68% and CulturalBench Easy at 67.07% but collapses CulturalBench Hard to 2.04%.

L Outliers and Quantization

Corpora and conditions. The outlier study uses a fresh corpus from window $[5100, 5500)$ with 400 passages of 128 tokens in 100/100/200 discovery, validation, and test splits; the quantization study uses window $[5500, 5884)$ with 384 passages of 512 tokens in 128 calibration and 256 evaluation passages. Both windows are disjoint from every earlier corpus under the three overlap rules of App. A. The outlier stage evaluates 43 conditions per passage and the quantization stage 78.

Direct-edit diagnostics. A direct-edit setter-rate diagnostic confirms that editing a coordinate changes its own tensor almost tautologically: rates are 22.2222%, 49.0278%, and 83.4500% for the selected population across the three registered operator families, 5.7500% for a same-layer masked variant, and a random median of 0.0278% (range 0–0.0556%). Basic scale statistics are 4.008046, 3.621468, and 4.101701 with pairwise differences 0.386578 and 0.093655. These quantities are descriptive only: because they include the edited coordinates themselves, they cannot separate propagation from direct editing, which is precisely why the primary bridge score of Eq. 12 admits only strictly downstream operators and masks same-layer selected coordinates identically in both conditions.

Bridge validation and endpoints. On the validation split the propagated bridge gives 0.003296 [0.002560, 0.004039] against a random P95 of 0.000183, while the amplification endpoint is unresolved at 0.000112 [−0.000296, 0.000520] with $p = 0.6001$. On the test split the suppression endpoint is 0.001103 [0.000538, 0.001670] with $p_{\mathrm{Holm}} = 0.0004$ and the amplification endpoint remains unresolved ($p_{\mathrm{Holm}} = 0.1346$). Restricting to down-projection inputs strictly downstream of the population gives 0.0478 [0.0454, 0.0503] for suppression against a random P95 of 0.00106, and 0.011170 [0.009945, 0.012402] for amplification against a random P95 of -0.0000533.

Massive-activation sweep. The registered residual sweep tests every post-block residual coordinate for the conjunction $|r| > 100$, $|r|/\mathrm{median} \geq 10^3$, at position 0, on at least 50% of passages. No coordinate qualifies anywhere in this checkpoint, so the canonical massive-activation phenotype is absent even though the sink and the SCN population are both present.

Structural super-neuron proxy. Because Li et al. (2026) characterize super neurons through several structural properties but do not provide a unique portable selection rule for Qwen, we use a preregistered clean-room proxy rather than claim an exact reproduction of their population. The proxy is constructed independently of SCN membership and sink outcomes on 256 frozen 40-token neutral passages.

For neuron i in layer ℓ, let $W_g^\ell[i,:]$ and $W_u^\ell[i,:]$ denote its gate- and up-projection rows, and let $w_{d,i}^\ell = W_d^\ell[:,i]$ be its down-projection column. We compute five quantities:

$$G_{\ell i} = \left\| W_g^\ell[i,:] \right\|_2, \qquad U_{\ell i} = \left\| W_u^\ell[i,:] \right\|_2, \tag{34}$$

$$S_{\ell i} = \frac{\mathbb{E}_x |a_{x,0,i}^\ell|}{\mathbb{E}_{x,t>0} |a_{x,t,i}^\ell| + \varepsilon}, \qquad D_{\ell i} = \frac{\max_j |W_d^\ell[j,i]|}{\mathrm{mean}_j |W_d^\ell[j,i]| + \varepsilon}, \tag{35}$$

with $\varepsilon = 10^{-12}$. Here $S_{\ell i}$ measures first-token activation selectivity and measures how strongly the neuron's residual write is concentrated in its largest output dimension.

We additionally measure the sparsity of each down-projection column using its Gini coefficient. For $v_j = |W_d^\ell[j,i]|$ and ascending order statistics $v_{(1)} \leq \cdots \leq v_{(d)}$,

$$H_{\ell i} = \frac{\sum_{j=1}^d (2j - d - 1) v_{(j)}}{d \sum_{j=1}^d v_{(j)}}. \tag{36}$$

Because these quantities have different units and scales, we combine their *within-layer ranks* rather than their raw values. For any quantity $q_{\ell i}$, let

$$R_\ell(q_{\ell i}) = \frac{\mathrm{rank}_\ell(q_{\ell i}) - 1}{d_{\mathrm{MLP}} - 1}, \tag{37}$$

where ranks are ascending percentile ranks in $[0,1]$ and ties receive their average rank. The final structural proxy score is

$$L_{\ell i} = \frac{1}{5} \left[R_\ell(G_{\ell i}) + R_\ell(U_{\ell i}) + R_\ell(S_{\ell i}) + R_\ell(D_{\ell i}) + R_\ell(H_{\ell i}) \right]. \tag{38}$$

We rank neurons by $L_{\ell i}$ separately within each layer and select the top k_ℓ, where k_ℓ is the frozen number of SCNs in that layer. Thus the proxy contains exactly 116 coordinates and exactly matches the SCN layer histogram, controlling for both population size and depth. Ties are broken by the smaller neuron index. No SCN identities, sink scores, or H3 quantization data enter the ranking. Because Li et al. (2026) do not supply a corresponding native Qwen population or universal threshold, we refer to this set throughout as a *structural super-neuron proxy*, not as the canonical Li et al. super-neuron population.

Quantization conditions. The reference is FP32 (CE 2.838966, PPL 17.10) and its registered drift control is null. Activation-only and weight-only quantization give 3.615326 and 3.630764 CE, showing that the W8A8 damage is not attributable to either factor alone. Protection budgets are 0.0227930% of activation-input coordinates and 0.0135992% of weight elements for the SCN condition, and 0.0462372% and 0.0719246% for the two magnitude conditions at their registered allocations.

Table 18: Coordinate overlap between protected sets, with the shared fraction of the SCN population.

Pair	shared coordinates	Jaccard	shared fraction
SCN vs layer-matched magnitude	9	0.0404	7.76%
SCN vs super-neuron proxy	3	0.0131	2.59%
Layer-matched magnitude vs super-neuron proxy	5	0.0220	4.31%
SCN vs global magnitude	6	0.0265	5.17%

Boundaries. Under W8A4 the super-neuron proxy gains $+0.092626$ against a random P95 of $+0.008928$ while the magnitude condition loses -0.072426, so no sparse protection succeeds at that precision. Restricting quantization to down-projection modules in SCN-bearing layers gives CEs of 2.839134 (reference), 4.616973 (naive), 3.267617 (SCN), 3.058746 (magnitude), and 4.380695 (super-neuron proxy), a gap of 1.777839 [1.747985, 1.807693] and an SCN-minus-magnitude difference of -0.208871 [–0.223156, –0.195281], replicating the main-text pattern within the population's own operator family.

Execution record. The stage ran in 9 hours 52 minutes 31 seconds with a peak memory of 8.49 GiB, writing 100, 12,900, 12,900, and 30 rows in the outlier tables and 7,296 and 19,968 rows in the quantization tables, across five registered condition families. Twenty-six focused tests and eighteen audit checks were executed, all passing. A calibration-sensitivity map over 57 conditions is retained as a diagnostic and is not used for any claim.

M Claim Boundaries

Table 19: What each experiment establishes and what it does not. Registered scope only; no claim in the paper exceeds the right-hand column's exclusions.

Section	Established	Not established
Sparse control (§3)	Bidirectional, dose-monotonic, matched-random-specific, held-out control of the sink by a frozen sparse population	Uniqueness or minimality of the population; sink exclusivity; behavioral harmlessness
Differentiation (App. F)	Non-identity with six re-implemented prior definitions under joint membership and ranking criteria	Exhaustiveness over all possible prior definitions; family-wise significant specificity, given the $p \geq 1/21$ floor
Pathways (§7)	An exact decomposition of the sink loss into query and key routes, validated by live restoration	Ranking of pathway magnitudes across models; any claim about value-side routes beyond the null audit
Behavior (§4)	Checkpoint-dependent downstream effects at both full and sink-matched suppression	A causal role for the sink in those effects; transfer to models outside the four studied
Dissociation (§5)	Near-disjoint membership and doubly dissociated crossed effects; substantial answer flipping above matched-random rates	Causal analysis of the shared arms (underpowered by rule); the completed Llama extension arm
Rescue (§6)	Exact restoration of the registered sink vector without corresponding behavioral restoration	Position-0-specific mediation, since the sham-adequacy gate fails; absence of any mediation
Quantization (§8)	Propagated downstream outlier reduction and 65% W8A8 loss recovery at a 0.02%-scale budget	Optimality of SCN-based protection; hardware speedup; benefit at W8A4