Attention Sinks and Outliers in Attention Residuals

OASIS: null routing across tokens and depth

Paper by Haozheng Luo, Haoran Dai, Ching-Yuen Huang et al.

Research group reading presentation

arXiv:2605.17887v2 | revised 24 September 2026

Reading: 7 October 2026

Explicit null capacity attenuates contributions; token null evidence guides depth routing.

Two routing decisions share the same constraint

Token attention: which context?

$$A_t^{(i)} = \sum_j p_{t,j}^{(i)} v_j^{(i)}, \qquad \sum_j p_{t,j}^{(i)} = 1$$

All probability goes to real tokens.

Depth routing: which source?

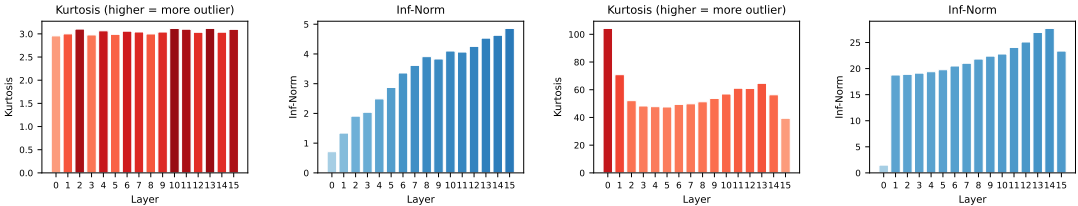
$$h_t^{(\ell)} = \sum_{i=0}^{\ell-1} \alpha_{i\to\ell,t} s_{i,t}, \qquad \sum_i \alpha_i = 1$$

All probability goes to real branches.

- $s_{0,t}$ is the initial state; $s_{i,t} = f_i(h_t^{(i)})$ is a transformed source.
- A small desired contribution can create pressure to find a surrogate no-op.
- The paper studies this pressure after converting pretrained backbones.

Paper, Secs. 3–4.1, PDF pp. 3–4. Concentration requires the geometric conditions discussed later.

The observed activations become much more extreme

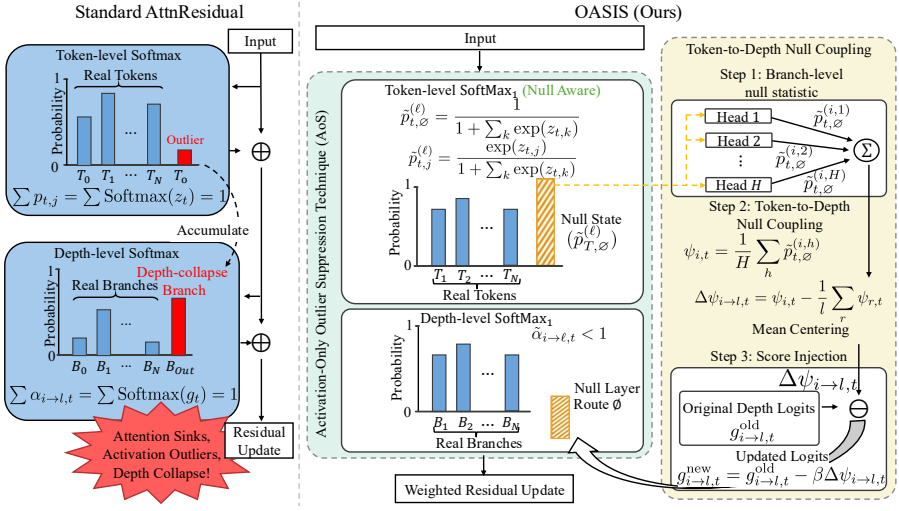


(a) Original Transformers **(b) Attention Residual**

- Read layer index horizontally; compare the different vertical scales.
- Original: kurtosis near 3. AttnResidual: about 50–70 after the first layer.
- Larger ranges make low-bit activation quantization more difficult.

Original Fig. 3, PDF p. 3. These layer-wise plots differ from the stack-wide statistics in Tables 1–2.

OASIS connects two null-aware routers



Original Fig. 1, PDF p. 2. Middle: AoS. Right: the token-to-depth coupling added by OASIS.

A zero-valued null state controls total contribution

$$\tilde{p}_j = \frac{e^{z_j}}{1 + \sum_k e^{z_k}}, \qquad \tilde{p}_\varnothing = \frac{1}{1 + \sum_k e^{z_k}}$$

Equivalent to one additional softmax category

Null logit $= 0$; null value $= 0$. The real-token probabilities sum to $1 - \tilde{p}_\varnothing$.

$$\tilde{p}_j = (1 - \tilde{p}_\varnothing) p_j, \qquad \tilde{A} = (1 - \tilde{p}_\varnothing) A$$

- At fixed logits, relative allocation among real tokens is preserved.
- A common downward logit shift now reduces the total real contribution.

Paper, Eq. (4.1), PDF p. 4; Appendix C.5, PDF pp. 20–21.

AoS applies the same mechanism across depth

$$\tilde{\alpha}_{i\to\ell,t} = \frac{e^{g_{i\to\ell,t}}}{1+\sum_{r=0}^{\ell-1} e^{g_{r\to\ell,t}}}, \qquad \tilde{h}_t^{(\ell)} = \sum_{i=0}^{\ell-1} \tilde{\alpha}_{i\to\ell,t} s_{i,t}$$

- **AoS**: token $\mathrm{Softmax}_1$ + depth $\mathrm{Softmax}_1$.
- The real source mixture can now be attenuated.
- The initial-state branch $i = 0$ remains a real candidate.

Depth-null semantics

The null value is zero. Attenuating the source mixture does not automatically preserve a previous hidden state.

Paper, Sec. 4.2 and Eq. (4.2), PDF p. 4; Appendix C.5, PDF pp. 21–22.

Each branch supplies a token-level null statistic

$$\psi_{i,t} = \frac{1}{H} \sum_{h=1}^{H} \tilde{p}_{t,\varnothing}^{(i,h)}, \qquad i \geq 1$$

$$\overline{\psi}_{\ell,t} = \frac{1}{\ell - 1} \sum_{r=1}^{\ell-1} \psi_{r,t}, \qquad \Delta\psi_{i\to\ell,t} = \begin{cases} \psi_{i,t} - \overline{\psi}_{\ell,t}, & i \geq 1, \\ 0, & i = 0. \end{cases}$$

- Average over H attention heads at the same token position.
- Center over transformed source branches available to target layer ℓ.
- The initial-state branch has no attention heads; its logit is left unchanged.

Paper, Sec. 4.3, PDF p. 5. No coupling is applied at $\ell = 1$.

Null evidence changes relative depth preferences

$$g^{\text{new}}_{i\to\ell,t} = g^{\text{old}}_{i\to\ell,t} - \beta\Delta\psi_{i\to\ell,t}, \qquad \beta \geq 0$$

$$\tilde{\alpha}_{i\to\ell,t} = \frac{e^{g^{\text{new}}_{i\to\ell,t}}}{1 + \sum_r e^{g^{\text{new}}_{r\to\ell,t}}}$$

- Above-average null mass receives a negative logit adjustment.
- Below-average null mass receives a positive adjustment.
- $\beta = \mathrm{softplus}(\theta)$ is learned; $\theta = -5$ gives $\beta \approx 0.0067$ initially.

Design hypothesis

A branch's tendency to route to null is useful evidence for depth selection.

Paper, Sec. 4.3, PDF p. 5; Appendix E, PDF p. 23. The training objective is unchanged.

A small example separates coupling from attenuation

Teaching example: two transformed branches, $\psi_1 = 0.8$, $\psi_2 = 0.2$; all old logits $= 0$; $\beta = 1$.

Candidate	New logit	AoS weight	OASIS weight
Initial state	0	0.2500	0.2445
Branch 1: high null	-0.3	0.2500	0.1811
Branch 2: low null	$+0.3$	0.2500	0.3300
Null state	0	0.2500	0.2445

- Higher-null branch 1 loses relative weight to branch 2.
- Null mass need not increase; the initial state's weight can also change.

Reading analysis derived from Sec. 4.3, PDF p. 5; illustrative numbers, not an experiment.

Concentration follows from a separation assumption

Token geometry: $\|v_o\| \le c_0 < c_1$; every other value has $\langle v_j, u\rangle \ge c_1$ along a shared unit direction u.

$$\|A\| \le \delta < c_1 \quad \Longrightarrow \quad p_o \ge \frac{c_1 - \delta}{c_0 + c_1}$$

Depth geometry: for $u_i = s_i - h^{\mathrm{ref}}$, assume $\|u_{i^*}\| \le b_0$ and $\langle u_i, w\rangle \ge b_1$ for all $i \ne i^*$, with unit w.

$$\|\Delta h\| \le \delta_t < b_1 \quad \Longrightarrow \quad \alpha_{i^*} \ge \frac{b_1 - \delta_t}{b_0 + b_1}$$

Why the condition matters

Values u and $-u$ average to zero with equal weights: separation is essential.

Lemma 5.1 and Theorem 5.1, PDF pp. 5–6; assumptions/proofs, pp. 17–19, 22. Counterexample: reading analysis.

Null mass gives a direct bound on contribution size

Assume $\|v_j\| \le V_{\max}$ and $\|s_i\| \le B_{\max}$ for all real candidates:

$$\left\|\tilde{A}_t^{(\ell)}\right\| \le V_{\max}(1 - \tilde{p}_{t,\varnothing}^{(\ell)})$$

$$\left\|\tilde{h}_t^{(\ell)}\right\| \le B_{\max}(1 - \tilde{\alpha}_{\varnothing,t}^{(\ell)})$$

- Sufficient null mass allows a small output without a dominant real candidate.
- The depth bound concerns the mixed state itself, relative to zero.
- Outlier suppression and quantized accuracy are assessed empirically.

Theorem 5.2, PDF p. 7; Appendix C.5, PDF pp. 20–22.

Evaluation combines architecture and quantization controls

- **Backbones:** LLaMA-3.2-1B, Qwen3-0.6B, Phi-4 (the linked checkpoint is Phi-4-reasoning).
- **Adaptation:** continued pretraining on BookCorpus + Wiki40B/en; 500 steps in the standard setup.
- **Precision:** FP16, W8A8; OmniQuant for W6A6/W4A4.
- **Tasks:** WikiText-2/Pile PPL; GSM8K and knowledge/math accuracy; 12K RULER/LongBench.
- **Controls:** original backbone, matched CPT, Vanilla AttnResidual, token-only, AoS, clipping/gating and sink baselines.

The authors report three seeds. Subsequent compact tables show means unless $\pm$ is displayed.

Paper, Sec. 6, PDF pp. 7–10; Appendix E, PDF pp. 22–23. The 12K experiment is a separate setup.

AoS improves quantization beyond token-only null routing

Backbone	Method	W8A8 PPL ↓	W4A4 PPL ↓
LLaMA-3.2-1B	Vanilla AttnResidual	30.19	74.97
	Token only / OutEffHop	10.42	26.18
	AoS	9.36	22.63
Qwen3-0.6B	Vanilla AttnResidual	36.64	50.82
	Token only / OutEffHop	10.60	32.36
	AoS	9.62	17.52

Depth-null routing adds value on top of token-null routing. Clipped and gated combinations also benefit (backup).

Paper, Table 1, PDF p. 8. WikiText-2 perplexity; W4A4 uses OmniQuant. Selected mean values.

Coupling adds 4.67 percentage points over AoS

Qwen3-0.6B, W4A4 GSM8K Pass@1

Configuration	Token null	Depth null	Coupling	Accuracy
Vanilla AttnResidual	–	–	–	45.67%
Token only	Yes	–	–	52.36%
Depth only	–	Yes	–	48.90%
AoS	Yes	Yes	–	54.21%
OASIS	Yes	Yes	Yes	58.88%

The key component comparison

$58.88 - 54.21 = 4.67$ points; relative gain over AoS $= 8.61\%$.

Paper, Sec. 6.2, PDF p. 8; Table 8, PDF p. 24. Higher is better.

W8A8 perplexity approaches the FP16 result

Backbone	Method	FP16 PPL ↓	W8A8 PPL ↓
LLaMA-3.2-1B	Vanilla AttnResidual	8.98	30.19
	OASIS	8.01	8.16
Qwen3-0.6B	Vanilla AttnResidual	10.57	36.64
	OASIS	9.20	9.52
Phi-4	Vanilla AttnResidual	11.21	1176.73
	OASIS	10.94	11.72

Per-model relative W8A8 reductions: **72.97%, 74.02%, 99.00%**. Their arithmetic mean is the headline **82.00%**.

Paper, Table 2, PDF p. 8. WikiText-2; mean values. Relative reductions recalculated from the table.

GSM8K recovery differs sharply across backbones

W4A4 Pass@1 (%)	LLaMA-3.2-1B	Qwen3-0.6B	Phi-4
Vanilla AttnResidual	12.97 ± 0.24	45.67 ± 0.86	43.16 ± 0.25
AoS	13.34 ± 0.18	54.21 ± 0.66	76.85 ± 0.21
OASIS	13.52 ± 0.17	58.88 ± 0.59	83.37 ± 0.21
Gain vs. Vanilla (points)	+0.55	+13.21	+40.21
Relative gain (%)	+4.24	+28.92	+93.16

- The headline **42.11%** averages the three relative gains.
- FP16 gains are only 0.11, 0.10 and 0.24 points, respectively.

Paper, Table 2, PDF p. 8. Mean $\pm$ standard deviation over three seeds. Gains: reading calculations.

Global statistics support model-wide sink suppression

Backbone	Method	BOS mass ↓	Entropy ↑	Null mass
Qwen3-0.6B	Original backbone	0.1064	3.74	0
	AttnResidual	0.4990	1.83	0
	OASIS	0.0013	4.44	0.0004
LLaMA-3.2-1B	Original backbone	0.1091	3.34	0
	AttnResidual	0.5630	1.86	0
	OASIS	0.0099	4.55	0.2185

Interpretation: Qwen suppresses sinks with very little average null mass. Learned redistribution is relevant beyond directly absorbing BOS mass.

Paper, Table 7, PDF p. 24. Aggregated across layers, heads, queries and samples. Interpretation: reading analysis.

At 12K, OASIS recovers most of the lost performance

	LLaMA-3.2-1B		Qwen3-0.6B	
Method	RULER	LongBench	RULER	LongBench
Original backbone	0.786	0.225	0.891	0.194
Vanilla AttnResidual	0.500	0.115	0.592	0.133
Learnable sink logit	0.597	0.131	0.811	0.184
AoS	0.567	0.121	0.781	0.174
OASIS	0.721	0.213	0.832	0.189

- OASIS exceeds the tested stabilization variants; the original backbone remains higher in every column.
- RULER loss recovered: 77.3% / 80.3%; LongBench: 89.1% / 91.8%.

Paper, Table 3, PDF p. 10. Higher scores are better. Recovery = (OASIS–Vanilla)/(Original–Vanilla).

The low-bit gains extend beyond WikiText-2 and GSM8K

Model	Method	Pile PPL ↓	MMLU % ↑	TriviaQA % ↑	MATH500 % ↑
LLaMA	AttnResidual	38.24	26.42	19.13	2.81
	AoS	11.62	29.27	24.04	3.43
	OASIS	10.17	30.19	25.56	3.77
Qwen	AttnResidual	44.52	33.21	12.09	15.41
	AoS	11.73	39.37	15.07	19.50
	OASIS	11.58	41.00	16.14	21.51

Pile uses **W8A8**; accuracy tasks use **W4A4**. Full-precision accuracy changes are much smaller.

Paper, Table 4, PDF p. 10. LLaMA-3.2-1B and Qwen3-0.6B; selected mean values.

The evidence leaves concrete implementation questions

- **Branch quality proxy:** head-averaged null probability can hide a useful minority of heads or non-attention branch content.
- **Attenuation tradeoff:** excessive null allocation may suppress weak useful signals; null usage varies strongly across models.
- **Reproducibility:** W8A8 calibration, the 12K training budget and global sink metric definitions need more detail.
- **Deployment:** low-bit accuracy is measured; optimized-kernel latency, throughput and memory gains are not reported.

Paper, Appendix E–F, PDF pp. 22–23; Tables 3 and 7. Proxy and reporting questions: reading analysis.

The contribution is a feedback path across routing levels

1. **AoS supplies null capacity.** Both token and depth routes can attenuate their real contributions.
2. **OASIS uses null evidence.** Above-average token null behavior lowers a branch's relative depth preference.
3. **The ablation isolates coupling.** Qwen W4A4 GSM8K rises from 54.21% (AoS) to 58.88% (OASIS).

Discussion question

When does a branch's decision to contribute little reliably predict its value to later layers?

Paper, Sec. 4.3, PDF p. 5; Table 8, PDF p. 24.

Paper and supporting material

Haozheng Luo, Haoran Dai, Ching-Yuen Huang et al.
Attention Sinks and Outliers in Attention Residuals.
arXiv:2605.17887v2, revised 24 September 2026.

- Paper: arxiv.org/abs/2605.17887v2
- Code linked by the authors: github.com/robinzixuan/OASIS
- Companion Chinese notes: `reading-notes.zh.md`
- Original figures reproduced with paper/page attribution.
- Numerical examples and explicitly labeled interpretations are reading analysis.

All experimental numbers in this deck are from the versioned paper. Main talk: slides 1–21; details follow.

Backup: matched CPT supports an architectural effect

Backbone	Method	Kurtosis ↓	Max norm ↓	W8A8 PPL ↓
LLaMA-3.2-1B	Backbone + CPT	3.07	5.20	9.85
	AttnResidual	91.71	84.43	30.19
	OASIS	3.02	12.25	8.16
Qwen3-0.6B	Backbone + CPT	3.16	4.77	11.50
	AttnResidual	93.04	74.78	36.64
	OASIS	2.95	13.00	9.52

CPT alone keeps the activation statistics close to the original backbone. OASIS improves low-bit PPL even though maximum magnitude remains higher than the CPT control.

Paper, Table 5, PDF p. 23. Matched continued-pretraining setup; selected mean values.

Backup: clipping and gating remain complementary

W4A4 PPL ↓	LLaMA-3.2-1B	Qwen3-0.6B	Phi-4
Clipped Softmax	26.81	39.89	112.18
Clipped AoS	22.57	16.95	37.62
Gated Attention	24.27	28.78	102.45
Gated AoS	21.45	15.43	31.89

- AoS improves the low-bit results of these existing stabilization designs.
- FP16 rankings can differ: LLaMA Gated Attention is 8.26 PPL, versus 8.68 for Gated AoS.

Paper, Table 1, PDF p. 8. WikiText-2; OmniQuant W4A4. Selected mean values.

Backup: sparsity and null capacity are distinct freedoms

LLaMA-3.2-1B	Kurtosis ↓	Max norm ↓	W8A8 PPL ↓
Softmax	91.706	84.433	30.189
Sparsemax	87.322	79.991	19.814
Entmax15	13.134	29.381	19.600
OASIS	3.024	12.250	8.162

- Sparse alternatives improve some outlier and quantization metrics.
- An explicit zero-valued null state also permits the total real mass to shrink.

Paper, Table 6, PDF p. 23. Selected mean values.

Backup: centering changes competition with the null state

For transformed branches $i, k \geq 1$:

$$\frac{\tilde{\alpha}_i}{\tilde{\alpha}_k} = \exp(g_i^{\mathrm{old}} - g_k^{\mathrm{old}}) \exp[-\beta(\psi_i - \psi_k)].$$

- The mean cancels when comparing two transformed branches.
- It does not cancel relative to the fixed null logit or the unchanged initial-state logit.
- With only one transformed branch, its centered statistic is zero.

Stable implementation

Include the null logit in log-sum-exp. Shifting only real logits changes the null probability.

Reading derivation from Sec. 4.3, PDF p. 5.

Backup: the pathology score has assumed dominance

$$\mathcal{S}_t = L_t + \lambda_1 C_t + \lambda_2 E_t, \qquad \lambda_1, \lambda_2 \geq 0$$

- L_t: irrelevant-token leakage; C_t: peak joint routing probability.
- $E_t = -H(\alpha_t) - \sum_\ell \alpha_t^{(\ell)} H(p_t^{(\ell)})$: negative joint entropy.
- Proposition 5.1 assumes all three AttnResidual terms are at least their vanilla counterparts, then sums them.

Sink aggregation is a convex average:

$$\Sigma_t = \sum_\ell \alpha_t^{(\ell)} \sigma_t^{(\ell)}, \qquad \min_\ell \sigma_t^{(\ell)} \leq \Sigma_t \leq \max_\ell \sigma_t^{(\ell)}.$$

It approaches a selected branch's sink mass if depth routing concentrates there.

Paper, Definitions 5.1–5.2, Proposition 5.1 and Lemma 5.2, PDF pp. 5–6; Assumption D.4, p. 22.

Backup: standard adaptation and inference settings

Setting	Reported value
Continued pretraining	500 steps; sequence 2048; block size 512
Learning rate	5×10^{-5}; linear schedule; 100 warmup steps
Batch / accumulation	1 / 64
Gradient clipping / weight decay	1.0 / 0.1
Coupling parameter	$\beta = \mathrm{softplus}(\theta)$; initialize $\theta = -5$
Hardware	4× H100 80GB
Generation	Temperature 0.6; top-p 0.95; max 4096 tokens

Long-context results use models trained with maximum length 12K. The paper gives less detail on the budget for that separate experiment.

Paper, Appendix E, PDF pp. 22–23; Sec. 6.4, PDF pp. 9–10.