

NEURIPS 2026

Attention Sinks and Outliers in Attention Residuals

OASIS · outlier- and sink-aware null routing for AttnResidual

Haozheng Luo*, Haoran Dai*, Ching-Yuen Huang*, Shaoyang Zhang, Xi Chen, Eric Hanchen Jiang, Yijiang Li, Chenghao Qiu, Chenwei Xu, Zhenyu Pan, Haotian Zhang, Binghui Wang, Yan Chen

* Equal contribution · openreview.net/forum?id=yjVSLVS0Dh · github.com/robinzixuan/OASIS



OVERVIEW

Problem and proposal

Problem

A second Softmax over residual branches (dual normalization) intensifies attention sinks and activation outliers, so low-bit quantization breaks.

Proposal · OASIS

Softmax₁ null routes at the token and depth levels (AoS), plus token-to-depth null coupling. The training objective is unchanged.

−95.90%

Avg. kurtosis

−81.75%

Max $\|x\|_\infty$

−82.00%

W8A8 perplexity

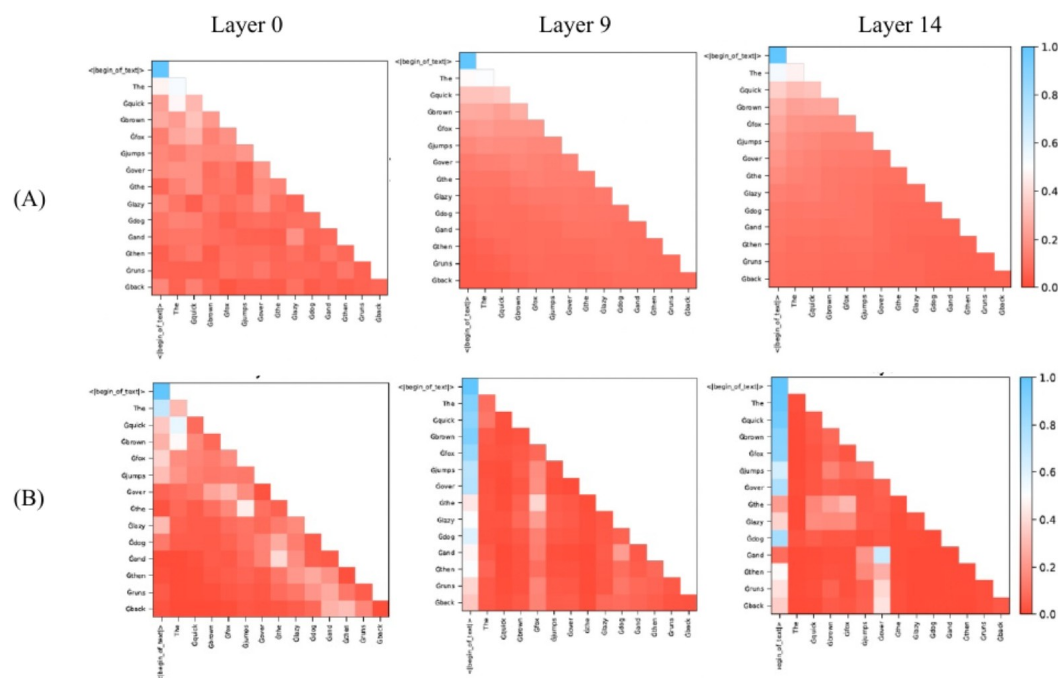
+42.11%

W4A4 GSM8K Pass@1

Relative to Vanilla AttnResidual, averaged over LLaMA-3.2-1B, Qwen3-0.6B and Phi-4.

FAILURE MODE 1

Intensified attention sinks



(A) vanilla Transformer · (B) AttnResidual — LLaMA-3.2-1B, layers 0 / 9 / 14

BOS ATTENTION MASS · LLAMA-3.2-1B

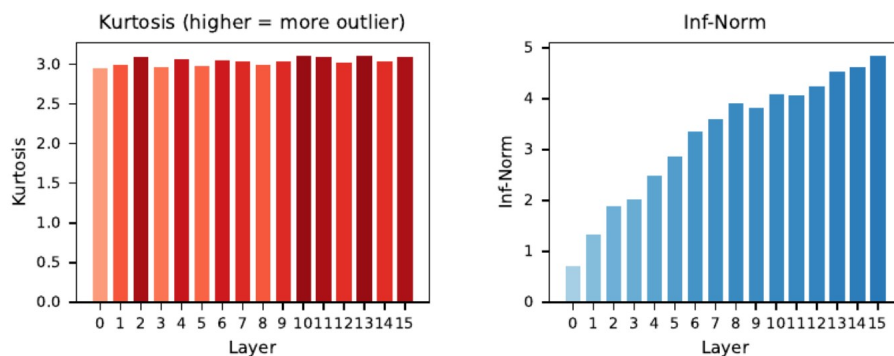
0.109 → 0.563

original backbone → Vanilla AttnResidual

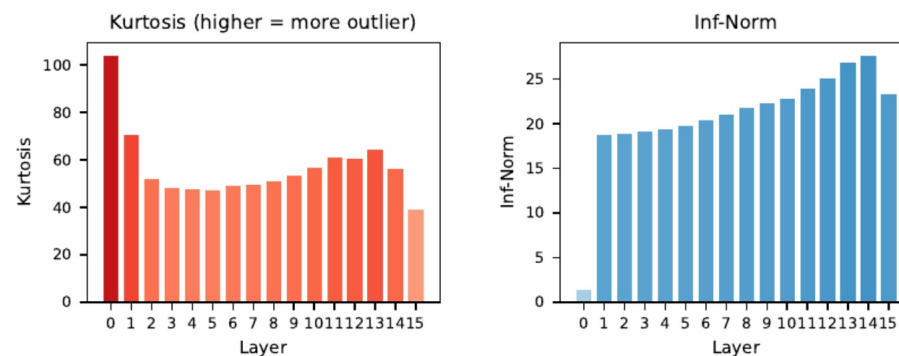
- The `<|begin_of_text|>` sink deepens with depth (layers 0 → 9 → 14).
- The extra depth Softmax reinforces sink accumulation instead of mitigating it.
- Sink rate reaches 78.6% (LLaMA) and 66.4% (Qwen); attention entropy falls to 1.86 / 1.83.

FAILURE MODE 2

Amplified activation outliers



(a) Original Transformers



(b) Attention Residual

≈ 3

Baseline kurtosis, with $\|x\|_\infty < 5$

> 100

AttnResidual kurtosis at layer 0
(50–70 after)

28

Peak $\|x\|_\infty$ under AttnResidual

+206.5%

W8A8 PPL vs. CPT control (CPT
alone: +3.6%)

ROOT CAUSE

No null pathway in standard Softmax



Every update must spend its full probability mass on real tokens and real branches.

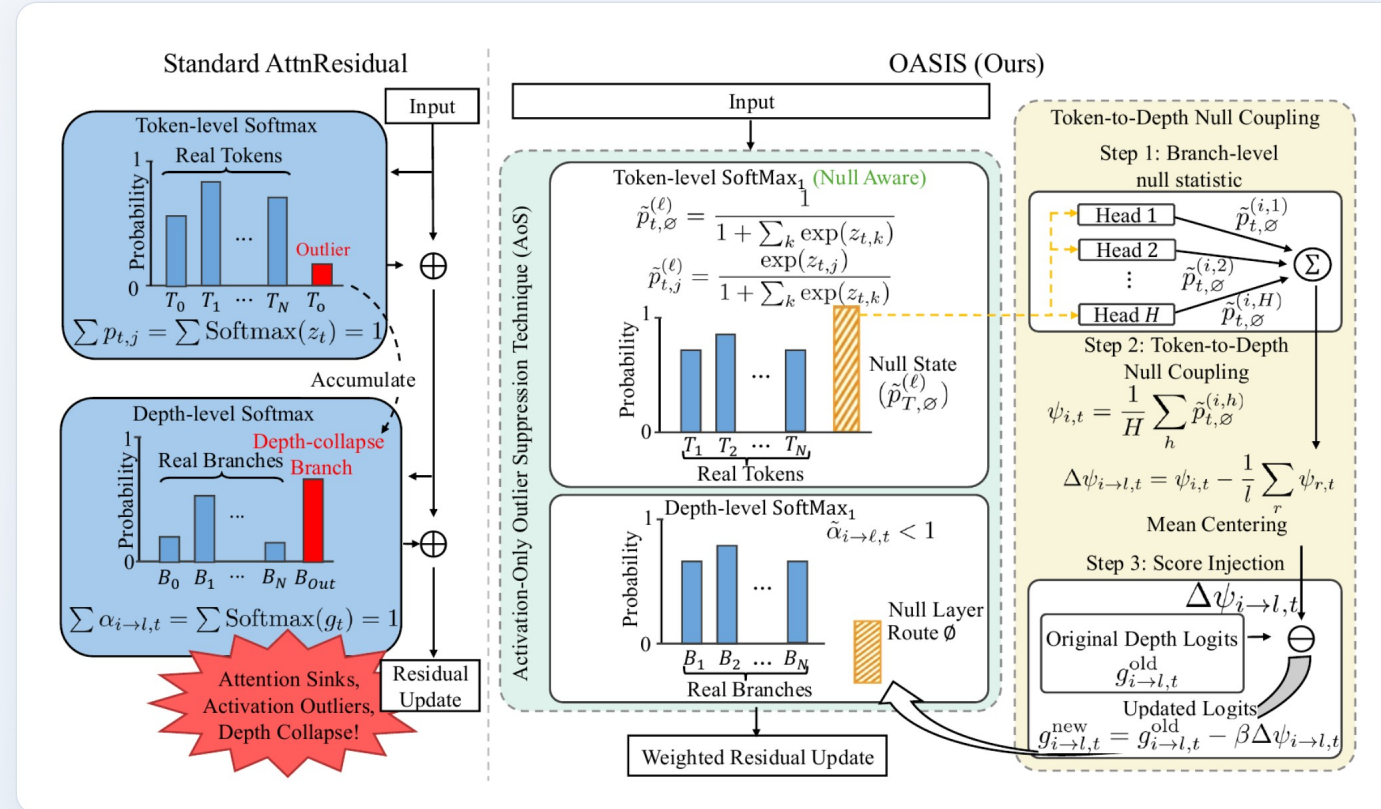
$$\sum_j p_{t,j} = 1 \qquad \sum_i \alpha_{i \rightarrow \ell, t} = 1$$

Token-level and depth-level Softmax both live on the probability simplex.

Under value separation, a near no-op output then requires concentrated routing: sinks at the token level, collapse pressure across depth, and heavy-tailed activations that inflate quantization ranges.

METHOD

OASIS at a glance



AoS · null routes

Swap token- and depth-level Softmax for Softmax₁: explicit null states at both levels.

OASIS · null coupling

Average each branch's token null mass over heads; above-average branches get a lower depth logit.

Routing only — residual structure and training objective stay the same.

METHOD

Null routes and token-to-depth coupling

1 · Null routes (AoS)

Softmax₁ at both levels; real mass < 1, ratios preserved.

$$\tilde{p}_{t,j} = \frac{e^{z_{t,j}}}{1 + \sum_k e^{z_{t,k}}}, \quad \tilde{p}_{t,\emptyset} = \frac{1}{1 + \sum_k e^{z_{t,k}}}$$
$$\tilde{\alpha}_{i \rightarrow \ell} = \frac{e^{g_{i \rightarrow \ell}}}{1 + \sum_k e^{g_{k \rightarrow \ell}}}, \quad \tilde{\alpha}_{\emptyset \rightarrow \ell} = \frac{1}{1 + \sum_k e^{g_{k \rightarrow \ell}}}$$

2 · Branch null statistic

Mean token-null mass over H heads, centered over branches $i \geq 1$.

$$\psi_{i,t} = \frac{1}{H} \sum_{h=1}^H \tilde{p}_{t,\emptyset}^{(i,h)} \quad \Delta\psi_{i \rightarrow \ell,t} = \psi_{i,t} - \bar{\psi}_{\ell,t}$$

3 · Score injection

Above-average null mass lowers the depth logit; then depth Softmax₁.

$$g_{i \rightarrow \ell,t}^{\text{new}} = g_{i \rightarrow \ell,t}^{\text{old}} - \beta \Delta\psi_{i \rightarrow \ell,t} \quad \beta = \text{softplus}(\cdot) \geq 0, \text{ learnable; initialized near } 0.007 \text{ (near no-op).}$$

THEORY

Theoretical guarantees

LEMMA 5.1

Token concentration

$$\|A_t\| \leq \delta \Rightarrow p_{t,o} \geq \frac{c_1 - \delta}{c_0 + c_1}$$

With separated values, a near-zero output forces mass onto a low-norm token.

THEOREM 5.1

Depth concentration

$$\|\Delta h_t^{(\ell)}\| \leq \delta_t \Rightarrow \alpha_{i^* \rightarrow \ell, t} \geq \frac{b_1 - \delta_t}{b_0 + b_1}$$

A near-reference mixture concentrates on the minimal-deviation branch; $\rightarrow 1$ as $b_0/b_1, \delta_t/b_1 \rightarrow 0$.

THEOREM 5.2

Softmax1 attenuation

$$\|\tilde{A}_t\| \leq V_{\max} (1 - \tilde{p}_{t,\emptyset})$$

Null mass attenuates real contributions without forcing concentration; same bound at depth with $B_{\max}$.

Proposition 5.1 (conditional): when leakage, peak concentration and entropy collapse each dominate vanilla attention, $S(\text{AttnResidual}) \geq S(\text{Vanilla})$.

RESULTS

Outlier suppression and low-bit robustness

Model	Method	Avg. kurtosis ↓	Max $\ x\ _\infty$ ↓	FP16 PPL ↓	W8A8 PPL ↓	W4A4 GSM8K ↑
LLaMA-3.2-1B	Vanilla	91.71	84.43	8.98	30.19	12.97
	OutEffHop	3.08	13.78	9.07	10.42	13.16
	OASIS (ours)	3.02	12.25	8.01	8.16	13.52
Qwen3-0.6B	Vanilla	93.04	74.78	10.57	36.64	45.67
	OutEffHop	3.04	14.16	9.99	10.60	52.36
	OASIS (ours)	2.95	13.00	9.20	9.52	58.88
Phi-4	Vanilla	38.83	37.96	11.21	1176.73	43.16
	OutEffHop	3.95	18.70	10.93	16.98	71.42
	OASIS (ours)	2.27	8.68	10.94	11.72	83.37

Averaged over three backbones: W8A8 PPL **-82.00%** and W4A4 GSM8K Pass@1 **+42.11%** vs. Vanilla, with no FP16 regression.

RESULTS

Long-context recovery at 12K tokens

LLaMA-3.2-1B		Qwen3-0.6B			
Method	Role of absorbed mass	RULER ↑	LongBench ↑	RULER ↑	LongBench ↑
Original backbone	—	0.786	0.225	0.891	0.194
Vanilla AttnResidual	—	0.500	0.115	0.592	0.133
S1: Learnable sink logit	Absorbed, not propagated to depth	0.597	0.131	0.811	0.184
S2: Prefix sink token	Adds a learned value to output	0.586	0.128	0.800	0.181
StreamingLLM	Kept in KV cache, not propagated	0.578	0.124	0.792	0.177
AttnResidual + AoS	Not propagated across levels	0.567	0.121	0.781	0.174
AttnResidual + OASIS	Propagated to depth routing	0.721	0.213	0.832	0.189

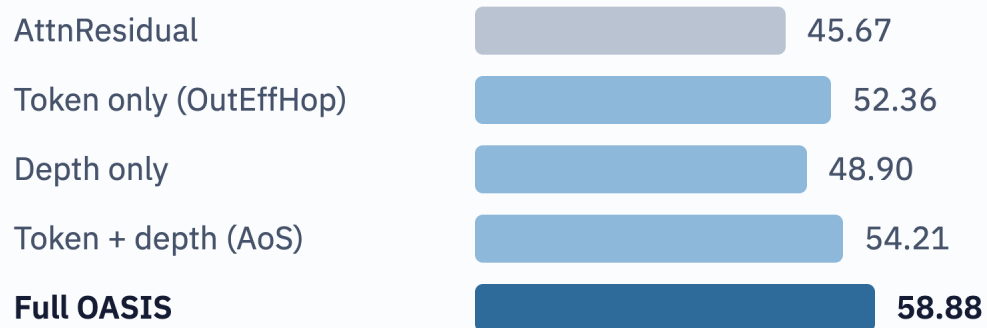
OASIS recovers **77.3% / 80.3%** of the RULER drop and **89.1% / 91.8%** of the LongBench drop (LLaMA / Qwen) — the only variant that feeds absorbed mass into depth routing.

ABLATION

Which parts matter

Component ablation

Qwen3-0.6B · W4A4 GSM8K Pass@1 (%), axis from 0



Coupling adds +4.67 points on top of both independent null routes.

Alternative normalizers

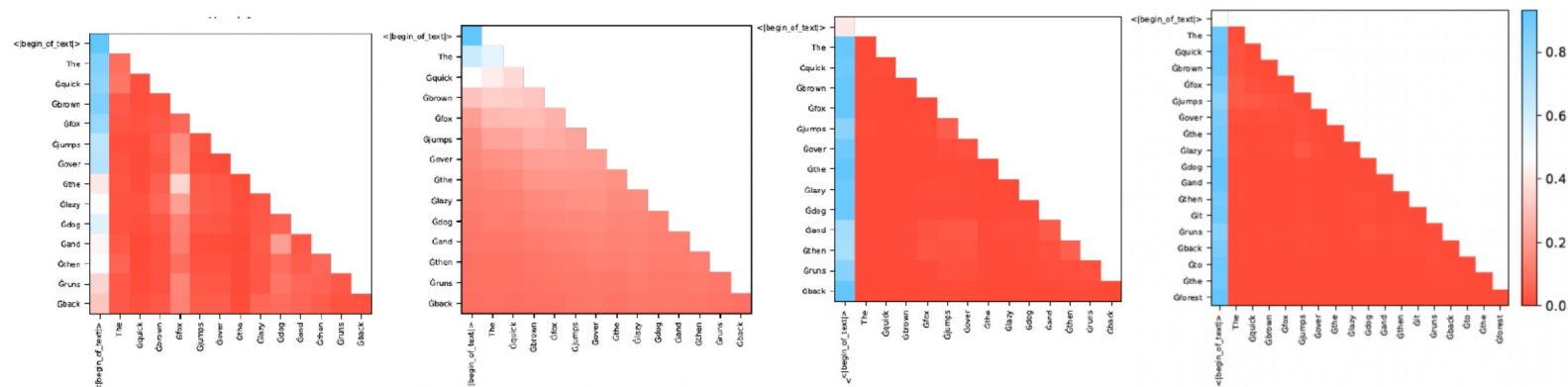
LLaMA-3.2-1B

Normalizer	Kurtosis ↓	W8A8 PPL ↓	W4A4 GSM8K ↑
Softmax	91.71	30.19	12.97
Sparsemax	87.32	19.81	12.98
Entmax15	13.13	19.60	13.16
OASIS	3.02	8.16	13.52

Sparse normalizers trim outliers; explicit null routing removes them.

RESULTS

Global attention-sink statistics



Trained attention-residual variants (paper Fig. 4): a blue first column is the `<|begin_of_text|>` sink.

BOS MASS ↓

LLaMA **0.563 → 0.010**

Qwen **0.499 → 0.001**

SINK RATE ↓

LLaMA **78.6% → 0.46%**

Qwen **66.4% → 0.01%**

ENTROPY ↑

LLaMA **1.86 → 4.55**

Qwen **1.83 → 4.44**

2ND-SINK MASS ↓

LLaMA **0.046 → 0.010**

Qwen **0.042 → 0.008**

CONCLUSION

Explicit null capacity, stable residual routing



PROBLEM

AttnResidual's dual normalization amplifies attention sinks, activation outliers and depth-collapse pressure.

INSIGHT

Standard Softmax has no null pathway, so near no-op updates force concentrated routing.

METHOD

OASIS: Softmax, null routes at token and depth levels, plus token-to-depth null coupling; objective unchanged.

RESULTS

−95.90% kurtosis, −81.75% max $\|x\|_\infty$, −82.00% W8A8 PPL, +42.11% W4A4 GSM8K; long-context performance largely recovered.

QUESTIONS WELCOME

Thank you

Code · github.com/robinzixuan/OASIS

Paper · openreview.net/forum?id=yjVSLVS0Dh

Webpage: <https://oasis-research.github.io/>

hluo@u.northwestern.edu · hdai10@hawk.illinoistech.edu · chy.huang@rutgers.edu

bwang70@illinoistech.edu · ychen@northwestern.edu

