Attention Sinks and Outliers in Attention Residuals

Haozheng Luo[1,*,†] **Haoran Dai**[2,8,*,†] **Ching-Yuen Huang**[3,*,†] **Shaoyang Zhang**[4]
Xi Chen[1] **Eric Hanchen Jiang**[5] **Yijiang Li**[6] **Chenghao Qiu**[7]
Chenwei Xu[1] **Zhenyu Pan**[1] **Haotian Zhang**[8] **Binghui Wang**[2,†] **Yan Chen**[1,†]

[1]Northwestern University [2]Illinois Institute of Technology [3]Rutgers University
[4]University of Michigan–Ann Arbor [5]University of California, Los Angeles
[6]University of California, San Diego [7]Texas A&M University [8]QuiverAI, Inc.

[*]Equal contribution. [†]Correspondence:
hluo@u.northwestern.edu, hdai10@hawk.illinoistech.edu, chy.huang@rutgers.edu
bwang70@illinoistech.edu, ychen@northwestern.edu

Abstract

We propose **OASIS**, an outlier- and sink-aware method that stabilizes dual-normalized attention-residual architectures through explicit null routing and token-to-depth null coupling. AttnResidual introduces an additional depth-wise normalization channel that improves inter-layer routing flexibility but can also amplify attention sinks, activation outliers, and low-bit quantization error. OASIS builds on explicit $\mathrm{Softmax}_1$-based null routes at both the token and depth levels and uses token-level null evidence to downweight depth branches exhibiting stronger null behavior. Theoretically, we characterize a conditional mechanism for sink-like attention concentration under dual normalization, offering insight into the low-bit sensitivity observed in AttnResidual. Experimentally, we compare OASIS against five baselines on three language-model backbones and multiple language-modeling, reasoning, and long-context benchmarks and observe consistent improvements in both attention sink mitigation and post-quantization performance. Notably, relative to Vanilla AttnResidual, OASIS reduces maximum infinity norm by **81.75%** and average kurtosis by **95.90%**, lowers W8A8 perplexity by **82.00%**, and improves W4A4 GSM8K Pass@1 by **42.11%** on average across LLaMA-3.2-1B, Qwen3-0.6B, and Phi-4. Code is available at: https://github.com/robinzixuan/OASIS.

1 Introduction

We propose **OASIS**, an outlier- and sink-aware method for improving the quantization robustness of attention-residual architectures. Attention sinks and activation outliers are two central obstacles to low-bit inference: concentrated attention can repeatedly allocate mass to a small set of tokens, while unusually large activation values expand the quantization range and increase quantization error. OASIS addresses these effects through an explicit no-op mechanism based on the $\mathrm{Softmax}_1$ function, following its use in Luo et al. [40, 39].

Modern foundation models are primarily based on the Transformer architecture introduced by Vaswani et al. [54], which has achieved strong performance across a broad range of tasks [41, 23, 22, 36]. To increase model capacity, Team et al. [53] propose AttnResidual, which extends standard token-level attention with an additional depth-wise routing mechanism. This design introduces a second Softmax normalization over residual branches, enabling learned information flow across layers. This added flexibility also creates a second simplex constraint. In the AttnResidual settings studied here, this constraint can exacerbate instability under compression and long-context inference, where attention

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

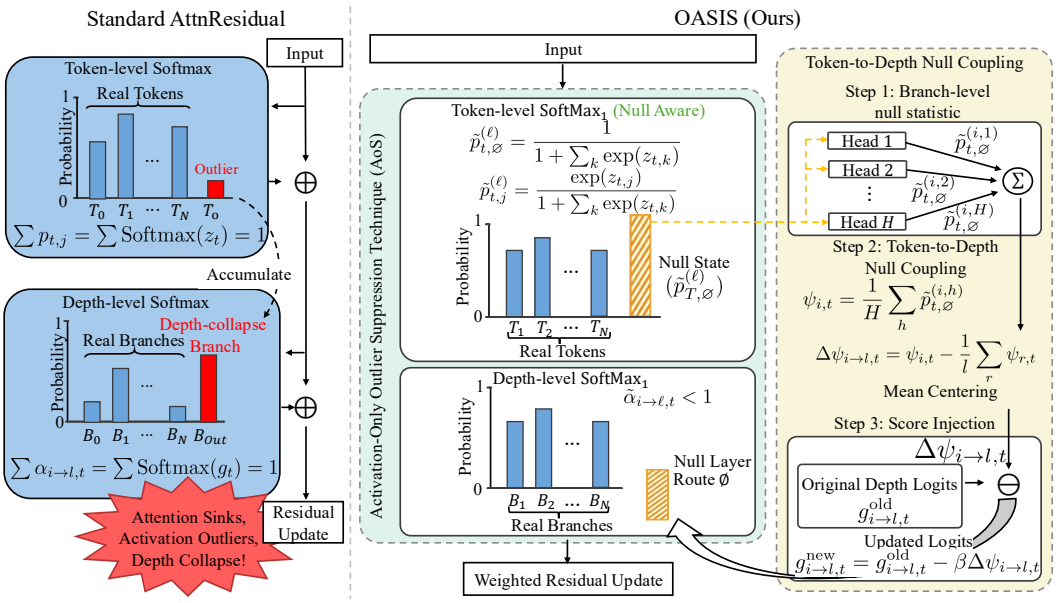


Figure 1: **Comparison of standard AttnResidual and OASIS.** OASIS adds null-aware token- and depth-level routing with token-to-depth null coupling, enabling explicit no-op allocation and reducing no-op outlier pressure, sink accumulation, and depth-collapse pressure.

outliers [27] and attention sinks [58] already impair quantization and attention allocation. In particular, when a token-level attention update or residual-branch update should contribute little information, standard Softmax must still assign its probability mass to real tokens and real branches, increasing routing concentration and pressure toward outlier formation.

To address this limitation, OASIS introduces an explicit *null route*, an additional probability state representing no useful contribution, through $\mathrm{Softmax}_1$ at both the token and depth levels. It then computes, for each residual branch, the average probability that its attention heads assign to the token-level null route. Branches with above-average null probability are treated as less informative and receive a lower depth-routing logit. Thus, OASIS connects token-level routing to depth-level routing, allowing the model to allocate no-op mass explicitly while making branch selection responsive to the information carried by each branch. The method modifies only the routing mechanism and leaves the training objective unchanged.

Contributions. Our main contributions are threefold:

- We show that the dual-normalization structure of AttnResidual can amplify activation outliers, attention sinks, and pressure toward depth collapse. We introduce **AoS**, which replaces the token- and depth-level $\mathrm{Softmax}$ operators with $\mathrm{Softmax}_1$ to provide explicit null routes at both levels.
- We propose **OASIS**, which extends AoS with token-to-depth null coupling. Token-level null probabilities are used to adjust depth routing, reducing the weight assigned to branches with stronger null behavior.
- Across LLaMA-3.2-1B, Qwen3-0.6B, and Phi-4, OASIS reduces average kurtosis by **95.90%** and maximum infinity norm by **81.75%** on average relative to Vanilla AttnResidual. It also lowers W8A8 perplexity by **82.00%** and improves W4A4 GSM8K Pass@1 by **42.11%** on average.

2 Related Work

Transformer outliers are associated with structured attention, special-token concentration, and uneven token influence [31, 8, 32], which hinder efficient and robust quantization [5, 49]. Standard softmax can further induce outliers when attention heads approximate partial or no-op updates [6], while first-token concentration has been linked to large activations [30]. Existing remedies include Gumbel-Softmax [10], gated attention [6], and $\mathrm{Softmax}_1$-based designs [27]. By adding a constant to the denominator, $\mathrm{Softmax}_1$ permits probability mass to enter an implicit null state instead of forcing it onto real tokens [39]. Appendix B shows additional related works.

3 Failure Modes of Dual Normalization in AttnResidual

In this section, we analyze the failure modes introduced by AttnResidual [18] in LLaMA-3.2-1B relative to a vanilla baseline. We show that its dual-normalization design gives rise to two coupled pathologies: intensified attention sinks and amplified activation outliers. We then trace both effects to a common architectural cause and motivate OASIS as a principled remedy. We compare AttnResidual with matched continued-pretraining controls in Table 5 to separate architectural effects from continued-pretraining dynamics.

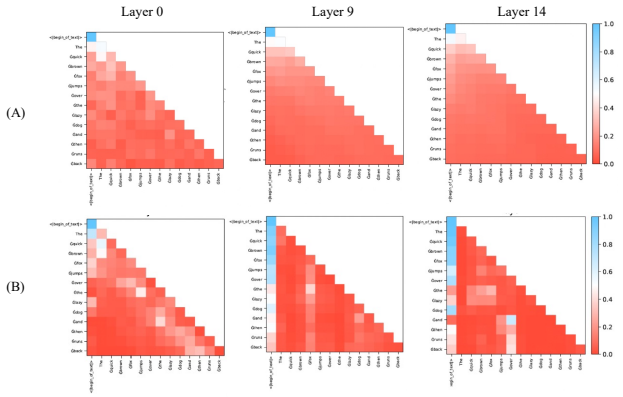
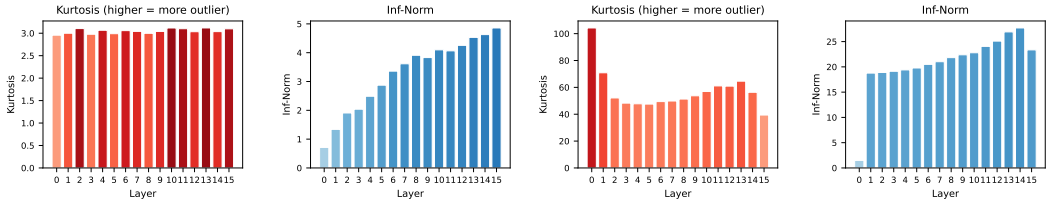


Figure 2: **Attention sink visualization.** We visualize token-level attention maps for the (A) vanilla Transformer and (B) AttnResidual at layers 0, 9, and 14. The <|begin_of_text|> token acts as a persistent attention sink whose concentration increases with depth, particularly under the dual-normalization design of AttnResidual.



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

Figure 3: **Outlier amplification in AttnResidual.** Layer-wise hidden-state kurtosis and ℓ_∞ norm for the original Transformer and AttnResidual. AttnResidual exhibits substantially heavier-tailed activations and larger activation ranges across depth.

Intensified attention sinks. We begin by examining token-level attention maps across depth. Although the <|begin_of_text|> token is already known to act as an attention sink in standard Transformers [58], we find that this effect becomes substantially more severe under AttnResidual (as shown in Figure 2). In particular, the additional $\mathrm{Softmax}$ in the depth-routing branch reinforces sink accumulation rather than mitigating it.

Outlier amplification. We next quantify the corresponding activation statistics across layers. Compared with the baseline, which maintains average kurtosis around 3 and $\|x\|_\infty < 5$, AttnResidual produces substantially heavier-tailed hidden states (as shown in Figure 3). Compared with the original Transformers, Kurtosis in AttnResidual exceeds 100 in the first layer and remains at 50–70 thereafter, while $\|x\|_\infty$ reaches as high as 28. Such amplification is known to impair post-training quantization [27, 12] and destabilize long-context inference.

Mechanistic implications. We find that both pathologies arise from the same architectural limitation: standard $\mathrm{Softmax}$ allocates all probability mass to active tokens and depth branches, leaving no null pathway. Consequently, under the separation conditions analyzed in Section 5, producing a small attention contribution or a depth mixture close to a reference state can require concentrated routing, thereby amplifying sink attention and activation outliers. Rather than suppressing these effects post hoc, OASIS (Section 4) introduces a $\mathrm{Softmax}_1$-based null-routing mechanism at both the token and depth levels, allowing the total contribution of real tokens and source branches to be attenuated without changing their relative allocations.

4 Methodology

Motivated by the sink concentration and activation outliers observed in Section 3, we first propose *Activation-only Outlier Suppression* (AoS), which replaces token- and depth-level $\mathrm{Softmax}$ with $\mathrm{Softmax}_1$ to introduce explicit null states. We further propose **OASIS** (Figure 1), which aggregates token-level null probabilities across attention heads within each transformed source branch and uses the centered branch statistic to adjust depth-routing logits. Branches with above-average null mass receive a negative logit adjustment, coupling the two routing levels while retaining the standard training objective and residual-branch structure.

4.1 Notation

Following Team et al. [53], the depth update at layer ℓ is: $h_\ell = \alpha_{0\to\ell} \cdot h_1 + \sum_{i=1}^{\ell-1} \alpha_{i\to\ell} \cdot f_i(h_i)$, with layer-mixing weights satisfying $\sum_{i=0}^{\ell-1} \alpha_{i\to\ell} = 1$. Here, $f_i(\cdot)$ is the i-th layer transformation and h_1 is the initial residual state. For convenience, we denote the source states mixed by depth routing as $s_{0,t} := h_t^{(1)}$ and $s_{i,t} := f_i(h_t^{(i)})$ for $i \geq 1$, so that $h_t^{(\ell)} = \sum_{i=0}^{\ell-1} \alpha_{i\to\ell,t} s_{i,t}$. For implementation, the router produces a depth-routing logit $g_{i\to\ell,t}$ for each source branch i, target layer ℓ, and token t, and the mixing weights are

$$\alpha_{i\to\ell,t} = \frac{\exp g_{i\to\ell,t}}{\sum_{r=0}^{\ell-1} \exp g_{r\to\ell,t}}, \qquad i \in \{0, 1, \ldots, \ell-1\}.$$

where g is a depth-routing *score* (logit). Inside each f_i, token-level attention is

$$p_{t,j}^{(\ell)} = \frac{\exp z_{t,j}^{(\ell)}}{\sum_{k=1}^{m} \exp z_{t,k}^{(\ell)}}, \qquad A_t^{(\ell)} = \sum_{j=1}^{m} p_{t,j}^{(\ell)} v_j^{(\ell)}.$$

Hence AttnResidual couples two normalizations: token $\mathrm{Softmax}$ (p) and depth $\mathrm{Softmax}$ (α). For theorem statements below, we write $\alpha_t^{(\ell)}$ as shorthand for depth weights when the index is implicit.

4.2 Activation-Only Outlier Suppression (AoS)

In autoregressive models, early tokens are visible to all subsequent positions, whereas later tokens are attended to only locally; this asymmetry can promote attention sinks [58]. Similar sink behavior also arises across depth: Wong et al. [56] identify persistent primary sinks in early layers and transient secondary sinks in middle layers. To mitigate sink-induced outliers at both routing levels, AoS replaces the standard Softmax with $\mathrm{Softmax}_1$ [44] in token-wise attention and depth routing. $\mathrm{Softmax}_1$ introduces an explicit null state that reduces the total mass assigned to real tokens or residual branches, while preserving their relative allocations. We formalize the token- and depth-level normalizations in Equations 4.1 and 4.2, respectively:

$$\tilde{p}_{t,j}^{(\ell)} = \frac{\exp z_{t,j}^{(\ell)}}{1 + \sum_{k=1}^{m} \exp z_{t,k}^{(\ell)}}, \quad \tilde{p}_{t,\varnothing}^{(\ell)} = \frac{1}{1 + \sum_{k=1}^{m} \exp z_{t,k}^{(\ell)}}, \tag{4.1}$$

$$\tilde{\alpha}_{i\to\ell} = \frac{\exp g_{i\to\ell}}{1 + \sum_{k=0}^{\ell-1} \exp g_{k\to\ell}}, \quad \tilde{\alpha}_{\varnothing\to\ell} = \frac{1}{1 + \sum_{k=0}^{\ell-1} \exp g_{k\to\ell}}. \tag{4.2}$$

Since the mass assigned to real tokens sums to less than one, token-level null routing attenuates the attention contribution by routing the remaining mass to a zero-valued null state. Analogously, replacing depth-$\mathrm{Softmax}$ with depth-$\mathrm{Softmax}_1$ attenuates the mixed contribution of real source branches. This depth-level null route reduces the probability budget available to real branches, and it does not by itself imply an identity mapping or preservation of a previous hidden state.

4.3 Token-to-Depth Null Coupling (OASIS)

While $\mathrm{Softmax}_1$ is null-aware, token- and depth-level routing remain decoupled, so depth may still weight uninformative layers. We introduce **OASIS**, a coupling mechanism that propagates token-level null evidence into depth routing. The key observation is that each source layer contributes a residual branch to depth routing. When most attention heads within such a branch route mass to the null space, the resulting branch output carries little informative content and should receive a lower depth weight.

Branch-level null statistic. For each transformed source branch $i \in \{1, \dots, \ell-1\}$, let $\tilde{p}_{t,\emptyset}^{(i,h)}$ denote the null posterior of attention head h for token t. We define $\psi_{i,t} := \frac{1}{H}\sum_{h=1}^{H} \tilde{p}_{t,\emptyset}^{(i,h)} \in [0,1]$. The initial-state branch $i = 0$ has no attention heads and therefore no token-level null posterior. We leave its routing logit unchanged and center the null statistic only over transformed branches:

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

For $\ell = 1$, no token-to-depth coupling is applied.

Score injection. Let $g_{i\to\ell,\,t}^{\text{old}}$ denote the pre-normalization depth routing logit produced by the router for source branch i, token t, and target layer ℓ. We subtract the centered null statistic, scaled by a non-negative coupling strength β:

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

When $\Delta\psi_{i\to\ell,t} > 0$ (above-average null mass), the logit decreases; when $\Delta\psi_{i\to\ell,t} < 0$ (below-average), it increases. The adjusted logits are then normalized with depth-$\mathrm{Softmax}_1$:

$$\tilde{\alpha}_{i\to\ell,\,t} = \frac{\exp\!\left(g_{i\to\ell,\,t}^{\text{new}}\right)}{1 + \sum_{r=0}^{\ell-1}\exp\!\left(g_{r\to\ell,\,t}^{\text{new}}\right)}.$$

The resulting weight $\tilde{\alpha}_{i\to\ell,t}$ replaces the original $\alpha_{i\to\ell}$ in the residual update of Section 4.1.

5 Theoretical Analysis

In this section, we analyze the routing constraints associated with dual normalization in AttnResidual [53], motivated by the routing mismatch discussed in Section 4. Under the stated separation conditions, we first derive a lower bound on the attention mass assigned to a low-norm token when a small attention output is required, and then examine the analogous concentration constraint for depth routing. We further characterize how depth mixing combines branchwise sink masses, including the case in which a single source branch dominates. Finally, we show that $\mathrm{Softmax}_1$, as used in AoS, provides an explicit null route for reducing contributions from real tokens and source branches, and characterize how OASIS uses token-level null evidence to adjust relative routing preferences among transformed source branches.

We use the notation introduced in Section 4. In particular, AttnResidual depth mixing (α), token attention (p), and the $\mathrm{Softmax}_1$ variables $(\tilde{p})$ follow Section 4 without redefinition. All assumptions and additional definitions used in this section are provided in Section D.

Lemma 5.1 (Attention concentration under separated value geometry)**.** Under Assumptions D.1 and D.3, fix a token position $t \in \mathcal{T}_{\text{noop}}$ and a branch ℓ. If the attention output satisfies

$$\|A_t^{(\ell)}\| = \left\|\sum_{j=1}^{m} p_{t,j}^{(\ell)} v_j^{(\ell)}\right\| \leq \delta,$$

with $\delta < c_1$, then the attention weight on token o satisfies

$$p_{t,o}^{(\ell)} \geq \frac{c_1 - \delta}{c_0 + c_1}.$$

Proof. See Section C.1 for a detailed proof. □

Lemma 5.1 shows that under the separation condition, a small attention output requires mass on the low-norm candidate, and the bound forces near-unit mass when both c_0/c_1 and δ/c_1 approach zero.

Definition 5.1 (Per-token pathology score)**.**

$$\mathcal{S}_t = \sum_{\ell} \alpha_t^{(\ell)} \sum_{j\in\mathcal{N}_t} p_{t,j}^{(\ell)} + \lambda_1 \max_{\ell,j} \alpha_t^{(\ell)} p_{t,j}^{(\ell)} + \lambda_2 \left(-H(\alpha_t) - \sum_{\ell} \alpha_t^{(\ell)} H(p_t^{(\ell)})\right).$$

Here, $H(\cdot)$ denotes Shannon entropy. The three terms measure irrelevant-token leakage, peak routing concentration, and joint depth–token entropy collapse, respectively. We use $\mathcal{S}_t$ as a surrogate for these routing quantities, rather than a direct measure of activation outliers.

Definition 5.2 (Vanilla pathology score)**.** For vanilla attention (single token-$\mathrm{Softmax}$), we use

$$\mathcal{S}_t^{\mathrm{Vanilla}} = \sum_{j \in \mathcal{N}_t} p_{t,j} + \lambda_1 \max_j p_{t,j} + \lambda_2 \big(- H(p_t) \big).$$

Proposition 5.1 (Conditional aggregation of routing pathologies)**.** Under matched no-op tolerance, suppose that AttnResidual satisfies Assumption D.4, i.e., its irrelevant-token leakage, peak routing concentration, and entropy-collapse terms are individually no smaller than those of vanilla attention. Then, for $\lambda_1, \lambda_2 \geq 0$,

$$S_t^{\mathrm{AttnResidual}} \geq S_t^{\mathrm{Vanilla}}.$$

Theorem 5.1 (Conditional concentration in depth routing)**.** Let the depth-composed deviation from a reference state be

$$\Delta h_t^{(\ell)} = \sum_{i=0}^{\ell-1} \alpha_{i \to \ell, t}, u_{i,t}, \qquad \alpha_{i \to \ell, t} \geq 0, \qquad \sum_{i=0}^{\ell-1} \alpha_{i \to \ell, t} = 1.$$

Suppose that the composed deviation is bounded by $\left\| \Delta h_t^{(\ell)} \right\| \leq \delta_t$, and that there exist a minimal-deviation branch i_t^* and a unit vector $w_t^{(\ell)}$ satisfying

$$\left\| u_{i_t^*, t} \right\| \leq b_0, \qquad \left\langle u_{i,t}, w_t^{(\ell)} \right\rangle \geq b_1 \quad \text{for all } i \neq i_t^*, \qquad b_1 > \delta_t.$$

Then the depth-routing weights satisfy

$$\alpha_{i_t^* \to \ell, t} \geq \frac{b_1 - \delta_t}{b_0 + b_1}, \qquad \sum_{i \neq i_t^*} \alpha_{i \to \ell, t} \leq \frac{b_0 + \delta_t}{b_0 + b_1}.$$

Consequently, in the high-separation regime

$$\frac{b_0}{b_1} \to 0, \qquad \frac{\delta_t}{b_1} \to 0,$$

the routing mass concentrates on the minimal-deviation branch:

$$\alpha_{i_t^* \to \ell, t} \to 1.$$

Proof. See Section C.3 for a detailed proof. □

Theorem 5.1 provides a conditional concentration bound: it does not imply that depth routing universally collapses, but characterizes the regime in which a near-zero composed deviation forces most routing mass onto a minimal-deviation branch.

Lemma 5.2 (Sink aggregation under depth routing)**.** Let the branchwise and depth-aggregated sink masses be

$$\sigma_t^{(\ell)} := \sum_{j \in \mathcal{K}} p_{t,j}^{(\ell)}, \qquad \Sigma_t := \sum_{\ell} \alpha_t^{(\ell)} \sigma_t^{(\ell)}.$$

Since the depth-routing weights lie on the simplex,

$$\min_{\ell} \sigma_t^{(\ell)} \leq \Sigma_t \leq \max_{\ell} \sigma_t^{(\ell)}.$$

If depth routing conditionally concentrates on a branch ℓ^*, such that $\alpha_t^{(\ell^*)} \to 1$, then

$$\Sigma_t \to \sigma_t^{(\ell^*)}.$$

Thus, depth routing forms a convex combination of branchwise sink masses and approaches the selected branch after concentration.

Proof. See Section C.4 for a detailed proof. □

Lemma 5.2 characterizes how existing branchwise sink mass is aggregated by depth routing; it does not by itself imply that depth routing creates or universally preserves an attention sink.

Theorem 5.2 ($\mathrm{Softmax}_1$ enables attenuation of real-token and branch contributions)**.** Introduce explicit null mass through $\mathrm{Softmax}_1$:

$$\tilde{p}^{(\ell)}_{t,j} = \frac{\exp z^{(\ell)}_{t,j}}{1 + \sum_k \exp z^{(\ell)}_{t,k}}, \qquad \tilde{p}^{(\ell)}_{t,\varnothing} = \frac{1}{1 + \sum_k \exp z^{(\ell)}_{t,k}}.$$

Assume that $\|v^{(\ell)}_j\| \le V_{\max}$ for every real token j and

$$B_{\max} := \max_{0 \le i \le \ell - 1} \|s_{i,t}\| < \infty.$$

Then $\mathrm{Softmax}_1$ attenuates real-token and branch contributions while preserving their relative allocations, without requiring concentration on any particular token or branch. Its effects on outliers, sinks, and depth routing are evaluated empirically.

Proof. See Section C.5 for a detailed proof. □

Theorem 5.2 provides the theoretical basis for AoS: explicit null routes allow token- and depth-level contributions to be attenuated without forcing the remaining probability mass to concentrate on real tokens or source branches.

6 Experimental Studies

We conduct a series of experiments to evaluate OASIS for quantization robustness and its ability to mitigate attention sink, benchmarking on Qwen3 [61] and LLaMA [19]. All experiments run with three independent random seeds, and we report mean and standard deviation for each metric.

Models. In our experiments, we use LLaMA-3.2 [19] and Qwen3 [61] as backbone models to evaluate quantization robustness and attention sink. Specifically, we continue training the LLaMA-3.2-1B[1], Qwen3-0.6B[2] and Phi-4[3] checkpoints with the attention residual architecture, following Gomez [18].

Data. Following the setup of Hu et al. [27], we use two real-world corpora for training: BookCorpus [64] and Wiki40B/en [21]. To comprehensively evaluate the impact of Softmax variants on attention sinks and quantization robustness, we consider a diverse suite of downstream benchmarks. Specifically, we use WikiText-2 [43] and The Pile to evaluate language-modeling performance and measure perplexity before and after quantization; GSM8K [9], MMLU [24], TriviaQA [28], and MATH500 [24] to assess reasoning and knowledge capabilities; and RULER [26] and LongBench [3] to evaluate long-context understanding. For accuracy-based benchmarks, higher values indicate better performance, whereas lower perplexity indicates better language-modeling and quantization performance.

6.1 Outlier Efficiency of AoS

Metrics. We report the *maximum infinity norm* $||x||_\infty$ and the *average kurtosis* of activation tensors $\mathbf{x}$ across all transformer layers as outlier metrics, averaging over all output components within the transformer stack. These statistics are strongly correlated with model quantizability and capture robustness to activation outliers [27, 5, 49]. Prior work [12, 55, 5] has shown that the presence of outliers can substantially degrade model performance after quantization, so we evaluate each model both before and after quantization. For pre-quantization performance, we report *perplexity* in **FP16** (16-bit floating point). For post-quantization performance, we report the same metric under **W8A8**, and further include **W6A6** and **W4A4** results using OmniQuant [48]. OmniQuant is a post-training quantization method for large language models that improves low-bit quantization by jointly optimizing weight and activation quantization with learnable transformations, without requiring full model retraining.

Results. In Table 1, AoS consistently outperforms the baselines. Unlike OutEffHop, which modifies only the token-level $\mathrm{Softmax}$, AoS normalizes both token- and residual-level pathways, leading to more consistent outlier suppression and stronger post-quantization robustness. Averaged across the

[1] https://huggingface.co/meta-llama/Llama-3.2-1B
[2] https://huggingface.co/Qwen/Qwen3-0.6B
[3] https://huggingface.co/microsoft/Phi-4-reasoning

Table 1: **Outlier reduction and quantization performance.** Comparison of AoS with Vanilla, OutEffHop, Clipped Softmax, and Gated Attention on LLaMA-3.2-1B , Qwen3-0.6B and Phi-4. We report average kurtosis, maximum activation magnitude $\|\mathbf{x}\|_\infty$, and perplexity under FP16, W8A8, and OmniQuant W6A6/W4A4. Best results are in **bold**, and the second-best results are underlined.

Model	Method	Avg. kurtosis	Max inf. norm	FP16	W8A8	OmniQuant	
						W6A6	W4A4
LLaMA-3.2-1B	Vanilla AttnResidual	91.71 ± 0.08	84.43 ± 0.15	**8.98** ± 0.01	30.19 ± 0.06	39.38 ± 0.01	74.97 ± 0.02
	OutEffHop	3.08 ± 0.02	13.78 ± 0.08	9.07 ± 0.01	10.42 ± 0.01	13.56 ± 0.00	26.18 ± 0.00
	AoS	**3.03** ± 0.01	**12.54** ± 0.07	9.00 ± 0.01	**9.36** ± 0.01	**11.28** ± 0.00	**22.63** ± 0.00
	Clipped Softmax	3.30 ± 0.00	14.24 ± 0.01	9.05 ± 0.00	10.26 ± 0.00	14.37 ± 0.00	26.81 ± 0.00
	Clipped OutEffHop	2.40 ± 0.00	10.85 ± 0.01	9.08 ± 0.00	10.19 ± 0.00	13.28 ± 0.00	25.62 ± 0.00
	Clipped AoS	**2.36** ± 0.00	**9.87** ± 0.01	**9.01** ± 0.00	**9.34** ± 0.00	**11.25** ± 0.00	**22.57** ± 0.00
	Gated Attention	4.39 ± 0.02	14.89 ± 0.01	**8.26** ± 0.00	9.21 ± 0.01	13.10 ± 0.02	24.27 ± 0.03
	Gated OutEffHop	2.14 ± 0.01	8.33 ± 0.00	8.75 ± 0.00	9.20 ± 0.03	12.08 ± 0.04	23.23 ± 0.08
	Gated AoS	**2.11** ± 0.01	**8.03** ± 0.00	8.68 ± 0.00	**8.88** ± 0.03	**10.70** ± 0.03	**21.45** ± 0.07
Qwen3-0.6B	Vanilla AttnResidual	93.04 ± 0.10	74.78 ± 0.13	10.57 ± 0.01	36.64 ± 0.08	44.78 ± 0.05	50.82 ± 0.59
	OutEffHop	3.04 ± 0.02	14.16 ± 0.06	9.99 ± 0.01	10.60 ± 0.01	13.36 ± 0.03	32.36 ± 0.37
	AoS	**3.03** ± 0.02	**13.90** ± 0.05	**9.35** ± 0.01	**9.62** ± 0.01	**10.48** ± 0.03	**17.52** ± 0.39
	Clipped Softmax	3.03 ± 0.00	14.16 ± 0.01	10.37 ± 0.00	21.38 ± 0.01	28.74 ± 0.06	39.89 ± 0.60
	Clipped OutEffHop	3.02 ± 0.00	12.69 ± 0.01	9.87 ± 0.00	10.44 ± 0.00	11.39 ± 0.03	30.74 ± 0.36
	Clipped AoS	**3.02** ± 0.00	**12.45** ± 0.01	**9.23** ± 0.00	**9.47** ± 0.00	**10.33** ± 0.03	**16.95** ± 0.36
	Gated Attention	3.03 ± 0.00	14.36 ± 0.02	10.56 ± 0.00	19.26 ± 0.01	23.43 ± 0.05	28.78 ± 0.53
	Gated OutEffHop	3.03 ± 0.00	13.81 ± 0.01	10.00 ± 0.00	10.57 ± 0.00	11.52 ± 0.00	21.26 ± 0.02
	Gated AoS	**3.03** ± 0.00	**13.55** ± 0.01	**9.35** ± 0.00	**9.59** ± 0.00	**10.46** ± 0.00	**15.43** ± 0.30
Phi-4	Vanilla	38.83 ± 0.02	37.96 ± 0.02	11.21 ± 0.00	1176.73 ± 0.01	1611.07 ± 0.01	2724.05 ± 0.02
	OutEffHop	3.95 ± 0.00	18.70 ± 0.01	**10.93** ± 0.00	16.98 ± 0.00	37.75 ± 0.00	160.70 ± 0.00
	AoS	**3.85** ± 0.00	**9.11** ± 0.01	11.09 ± 0.00	**12.18** ± 0.00	**16.85** ± 0.00	**48.26** ± 0.00
	Clipped Softmax	4.76 ± 0.01	24.53 ± 0.01	11.43 ± 0.00	23.08 ± 0.01	46.64 ± 0.03	112.18 ± 0.30
	Clipped OutEffHop	3.51 ± 0.01	15.75 ± 0.01	**10.86** ± 0.00	16.66 ± 0.00	29.59 ± 0.02	90.74 ± 0.36
	Clipped AoS	**3.42** ± 0.00	**7.69** ± 0.01	11.02 ± 0.00	**12.01** ± 0.00	**16.70** ± 0.01	**37.62** ± 0.38
	Gated Attention	4.05 ± 0.01	24.96 ± 0.02	11.07 ± 0.00	20.76 ± 0.01	42.82 ± 0.03	102.45 ± 0.28
	Gated OutEffHop	3.34 ± 0.01	14.81 ± 0.01	**10.75** ± 0.00	15.97 ± 0.03	37.16 ± 0.02	46.69 ± 0.05
	Gated AoS	**3.26** ± 0.01	**7.44** ± 0.01	10.90 ± 0.00	**11.36** ± 0.03	**14.34** ± 0.02	**31.89** ± 0.28

Table 2: **OASIS performance across common language models.** We conduct experiments on LLaMA-3.2-1B , Qwen3-0.6B and Phi-4 to evaluate activation stability and quantization robustness. Compared with Vanilla AttnResidual, OASIS reduces average kurtosis by **95.90%**, maximum infinity norm by **81.75%**, and W8A8 perplexity by **82.00%**, while improving W4A4 GSM8K accuracy by **42.11%** on average. Best results are in **bold** and second-best results are underlined.

Model	Method	Avg. kurtosis	Max inf. norm	Perplexity↓		GSM8K↑	
				FP16	W8A8	FP16	W4A4
LLaMA-3.2-1B	Vanilla	91.71 ± 0.08	84.43 ± 0.15	8.98 ± 0.01	30.19 ± 0.06	13.64 ± 0.17	12.97 ± 0.24
	OutEffHop	3.08 ± 0.02	13.78 ± 0.08	9.07 ± 0.01	10.42 ± 0.01	13.58 ± 0.15	13.16 ± 0.21
	AoS	3.03 ± 0.01	12.54 ± 0.07	9.00 ± 0.01	9.36 ± 0.01	13.62 ± 0.13	13.34 ± 0.18
	OASIS	**3.02** ± 0.02	**12.25** ± 0.07	**8.01** ± 0.01	**8.16** ± 0.01	**13.75** ± 0.12	**13.52** ± 0.17
Qwen3-0.6B	Vanilla	93.04 ± 0.10	74.78 ± 0.13	10.57 ± 0.01	36.64 ± 0.08	62.58 ± 0.24	45.67 ± 0.86
	OutEffHop	3.04 ± 0.02	14.16 ± 0.06	9.99 ± 0.01	10.60 ± 0.01	62.59 ± 0.19	52.36 ± 0.74
	AoS	3.03 ± 0.02	13.90 ± 0.05	9.35 ± 0.01	9.62 ± 0.01	62.65 ± 0.17	54.21 ± 0.66
	OASIS	**2.95** ± 0.02	**13.00** ± 0.05	**9.20** ± 0.01	**9.52** ± 0.01	**62.68** ± 0.15	**58.88** ± 0.59
Phi-4	Vanilla	38.83 ± 0.02	37.96 ± 0.02	11.21 ± 0.00	1176.73 ± 0.01	93.78 ± 0.17	43.16 ± 0.25
	OutEffHop	3.95 ± 0.00	18.70 ± 0.01	**10.93** ± 0.00	16.98 ± 0.00	93.91 ± 0.15	71.42 ± 0.23
	AoS	3.85 ± 0.00	9.11 ± 0.01	11.09 ± 0.00	12.18 ± 0.00	93.84 ± 0.16	76.85 ± 0.21
	OASIS	**2.27** ± 0.00	**8.68** ± 0.01	10.94 ± 0.01	**11.72** ± 0.00	**94.02** ± 0.15	**83.37** ± 0.21

two backbones in the plain attention-residual setting, AoS reduces average kurtosis and maximum infinity norm relative to Vanilla by **96.72%** and **83.40%**, respectively, and lowers perplexity under W8A8, W6A6, and W4A4 by **71.60%**, **72.36%**, and **68.42%**. These gains persist in clipped and gated variants, where the corresponding reductions are **10.95%** and **15.40%** for the outlier metrics, and **51.09%**, **60.78%**, and **57.79%** for perplexity, showing that AoS complements clipping- and gating-based designs.

6.2 Component Ablation of Null Routing and Token-to-Depth Coupling

To isolate the contributions of token-level null routing, depth-level null routing, and token-to-depth coupling, we evaluate their combinations on Qwen3-0.6B under W4A4 quantization. As shown in Table 8, token-only and depth-only routing improve GSM8K accuracy from **45.67%** to **52.36%** and **48.90%**, respectively. Combining the two independent null routes yields **54.21%**, while adding token-to-depth coupling further improves accuracy to **58.88%**. These results show that coupling contributes beyond the two independent null routes.

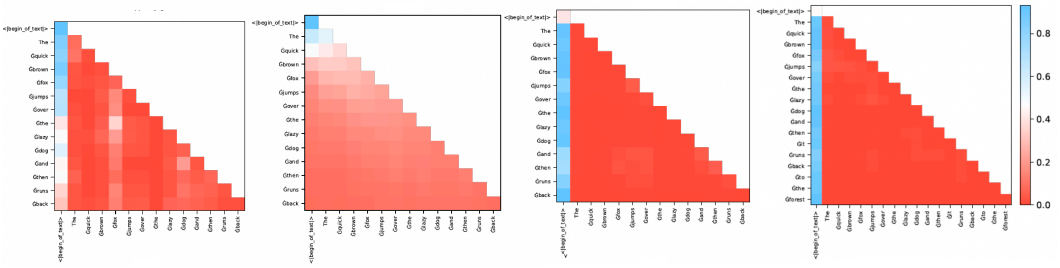


Figure 4: **Attention sink mitigation in representative attention maps.** We visualize representative head-0 token-level attention maps on a short causal prompt for trained attention-residual variants. Strong vertical concentration on the leading <|begin_of_text|> token indicates an attention sink. Relative to standard variants, null-aware routing weakens both the dominant first-token sink and secondary sink bands, yielding a smoother and more local causal attention pattern.

6.3 Outlier Efficiency of OASIS

Setup. To evaluate model capacity in a standard reasoning setting, we follow Pan et al. [45] and benchmark all models on GSM8K with chain-of-thought prompting. We report Pass@1 as the evaluation metric and additionally evaluate perplexity under FP16 and W8A8, as well as Pass@1 under FP16 and W4A4 with OmniQuant.

Baselines. We compare three representative methods spanning the main design paradigms for improving quantization robustness and mitigating attention sink: (1) **Clip Attention** [6], which clips $\mathrm{Softmax}$ logits to limit the extreme attention scores that drive activation outliers; (2) **Gated Attention** [6], which introduces an explicit gating mechanism that allows attention heads to perform partial or near no-op updates without inducing large $\mathrm{Softmax}$ magnitudes; and (3) **OutEffHop** [27], which replaces standard attention with an outlier-efficient Hopfield-style retrieval mechanism and provides a principled $\mathrm{Softmax}_1$-based alternative for suppressing outlier formation. This set of baselines covers clipping-based, gating-based, and $\mathrm{Softmax}_1$-based approaches, providing a broad comparison across prior strategies for outlier control and quantization-friendly attention design.

Results. As shown in Table 2, OASIS consistently improves numerical stability and robustness under quantization on both models. On LLaMA-3.2-1B and Qwen3-0.6B, it lowers mean kurtosis and maximum infinity norm to **2.987** and **12.624**, and this translates into a mean W8A8 perplexity of **8.840** and a mean W4A4 GSM8K Pass@1 of **36.20%**, corresponding to a **73.55%** reduction and a **23.47%** gain relative to Vanilla, respectively; the corresponding FP16 averages are **8.604** and **38.22%**. Relative to AoS, OASIS further reduces mean W8A8 perplexity by **6.85%** and improves mean W4A4 GSM8K Pass@1 by **7.18%**.

6.4 Improvement in Attention Sinks

We evaluate sink behavior directly by visualizing token-level attention maps on a short causal prompt containing the leading <|begin_of_text|> token, which is known to induce sink formation [58]. Using the trained models , we inspect the representative attention head from layer 14 and compare the resulting attention patterns qualitatively. In these maps, stronger vertical concentration on the initial token indicates a more severe sink, whereas a smoother lower-triangular pattern reflects healthier local attention. As shown in Figure 4, standard attention-residual variants exhibit a pronounced sink on the initial token and, in some cases, develop secondary vertical bands over early tokens, indicating that attention mass remains concentrated on a small subset of positions. AoS partially alleviates this effect, but the bias toward early tokens remains visible. OASIS yields the cleanest pattern: attention is redistributed away from <|begin_of_text|>, secondary sinks are suppressed, and the maps recover a smoother causal decay dominated by nearby content tokens. These qualitative changes are consistent with the reductions in outlier metrics and support the central claim that null-aware routing mitigates the structural cause of sink accumulation rather than merely masking its downstream effects.

We also evaluate models trained with a maximum sequence length of 12K on RULER and LongBench. As shown in Table 3, Vanilla AttnResidual substantially degrades long-context performance, whereas OASIS largely recovers the original backbone's performance on both benchmarks. On RULER, OASIS improves over Vanilla AttnResidual by 0.221 and 0.240 points for LLaMA-3.2-1B and

Table 3: **Long-context evaluation on RULER and LongBench at 12K tokens.** We conduct experiments on `LLaMA-3.2-1B` and `Qwen3-0.6B` to evaluate long-context performance. OASIS consistently outperforms Vanilla AttnResidual and all sink-oriented baselines, improving RULER/LongBench by **+0.221/+0.098** and **+0.240/+0.056** over Vanilla AttnResidual, respectively. Variances are omitted as they are consistently $\leq 0.2\%$. Best results are shown in **bold**.

Method	Role of absorbed mass	LLaMA-3.2-1B		Qwen3-0.6B	
		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 but not propagated to depth routing	0.597	0.131	0.811	0.184
S2: Prefix sink token	Contributes a learned value to the attention output	0.586	0.128	0.800	0.181
StreamingLLM	Retained in the KV cache but not propagated to depth routing	0.578	0.124	0.792	0.177
AttnResidual + AoS	Not propagated across routing levels	0.567	0.121	0.781	0.174
AttnResidual + OASIS	Propagated to depth routing	**0.721**	**0.213**	**0.832**	**0.189**

Table 4: **Generalization across language modeling and downstream tasks.** We conduct experiments on Pile, MMLU, TriviaQA, and MATH500 using `LLaMA-3.2-1B` and `Qwen3-0.6B` to evaluate generalization beyond the primary benchmarks. OASIS consistently outperforms Vanilla AttnResidual and all competing variants under both FP16 and low-bit quantization, reducing W8A8 perplexity from **38.24 to 10.17** and **44.52 to 11.58**, while improving W4A4 accuracy across all downstream tasks. Lower perplexity and higher accuracy are better. Best results are shown in **bold**.

		The Pile ↓		MMLU ↑		TriviaQA ↑		MATH500 ↑	
Model	**Method**	**FP16**	**W8A8**	**FP16**	**W4A4**	**FP16**	**W4A4**	**FP16**	**W4A4**
LLaMA-3.2-1B	Original backbone	11.32 ± 0.01	11.79 ± 0.01	31.30 ± 0.22	29.01 ± 0.33	28.14 ± 0.30	24.40 ± 0.43	4.16 ± 0.44	2.59 ± 0.42
	Vanilla AttnResidual	11.08 ± 0.01	38.24 ± 0.08	31.50 ± 0.21	26.42 ± 0.32	27.96 ± 0.32	19.13 ± 0.51	4.40 ± 0.43	2.81 ± 0.44
	AttnResidual + OutEffHop	11.15 ± 0.01	12.87 ± 0.01	31.38 ± 0.20	28.60 ± 0.32	27.97 ± 0.30	23.10 ± 0.40	4.42 ± 0.41	3.20 ± 0.42
	AttnResidual + AoS	11.10 ± 0.01	11.62 ± 0.01	31.58 ± 0.24	29.27 ± 0.32	28.17 ± 0.30	24.04 ± 0.42	4.46 ± 0.43	3.43 ± 0.44
	AttnResidual + OASIS	**9.98 ± 0.01**	**10.17 ± 0.01**	**31.76 ± 0.24**	**30.19 ± 0.30**	**28.88 ± 0.29**	**25.56 ± 0.43**	**4.56 ± 0.39**	**3.77 ± 0.44**
Qwen3-0.6B	Original backbone	13.04 ± 0.01	13.61 ± 0.01	43.38 ± 0.22	40.06 ± 0.30	18.02 ± 0.29	15.41 ± 0.43	24.54 ± 0.61	19.59 ± 0.72
	Vanilla AttnResidual	12.83 ± 0.01	44.52 ± 0.09	43.46 ± 0.23	33.21 ± 0.39	17.97 ± 0.33	12.09 ± 0.50	24.83 ± 0.64	15.41 ± 0.79
	AttnResidual + OutEffHop	12.12 ± 0.01	12.94 ± 0.01	43.49 ± 0.20	38.09 ± 0.31	18.08 ± 0.34	14.56 ± 0.39	24.80 ± 0.59	18.54 ± 0.73
	AttnResidual + AoS	11.41 ± 0.01	11.73 ± 0.01	43.73 ± 0.24	39.37 ± 0.31	18.22 ± 0.34	15.07 ± 0.43	24.97 ± 0.60	19.50 ± 0.71
	AttnResidual + OASIS	**11.19 ± 0.01**	**11.58 ± 0.01**	**43.77 ± 0.21**	**41.00 ± 0.32**	**18.49 ± 0.29**	**16.14 ± 0.43**	**25.18 ± 0.60**	**21.51 ± 0.74**

Qwen3-0.6B, respectively, recovering **77.3% and 80.3%** of the performance lost relative to the original backbone. On LongBench, OASIS improves by 0.098 and 0.056 points, recovering **89.1% and 91.8%** of the corresponding degradation. These results indicate that OASIS stabilizes inter-layer attention routing without sacrificing the benefits of AttnResidual.

6.5 Ablation Studies

Generalization across domains and tasks. We further evaluate whether the benefits of OASIS generalize beyond WikiText-2 and GSM8K. As shown in Table 4, OASIS achieves the lowest perplexity on The Pile under both FP16 and W8A8 across LLaMA-3.2-1B and Qwen3-0.6B, indicating that its improvements are not specific to the Wikipedia domain. The same trend extends to downstream tasks: under W4A4 quantization, OASIS consistently outperforms Vanilla AttnResidual and the alternative stabilization methods on MMLU, TriviaQA, and MATH500, while preserving or slightly improving FP16 performance. Notably, Vanilla AttnResidual retains competitive FP16 performance but degrades sharply after quantization, whereas OASIS substantially closes this gap across all evaluated tasks. These results demonstrate that OASIS improves quantization robustness across both language-modeling domains and downstream reasoning and knowledge tasks, rather than overfitting to a particular benchmark. An additional ablation study of null routing and token-to-depth coupling is presented in Appendix 6.2.

7 Discussion and Conclusion

We propose OASIS, a null-aware routing framework that stabilizes AttnResidual under dual normalization by introducing $\mathrm{Softmax}_1$-based null routes at both the token and depth levels and coupling token-level null evidence back into depth routing. Across evaluated backbones and settings, OASIS reduces the maximum infinity norm by **81.75%** and average kurtosis by **95.90%**, lowers W8A8 perplexity by **82.00%**, and improves W4A4 GSM8K Pass@1 by **42.11%** on average, without modifying the training objective or overall architectural skeleton. These results show that explicit null capacity provides a simple and effective mechanism for stabilizing dual-normalized residual routing while improving robustness to quantization and long-context inference.

Acknowledgments

The authors thank the anonymous reviewers and program chairs for their constructive feedback. Binghui Wang is partially supported by NSF ECCS-2216926, CCF-2331302, CNS-2241713, and CNS-2339686. Haozheng Luo is partially supported by the Lambda Researcher Grant, OpenAI Research Access Program, Thinking Machine Research Grant, and Adobe Fellowship. Ching-Yuen Huang is partially supported by NSF grants 2310966, 2235405, 2212301, 2003874, 1951890 and NIH grant 2R01HL127661 awarded to Dimitris N. Metaxas. This research was supported in part by the computational resources and staff contributions provided by Northwestern University's Quest High Performance Computing Facility, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. This work was also supported in part by computational resources provided through NSF CIS260386. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

[1] Anand Anand, Umberto Cappellazzo, Stavros Petridis, and Maja Pantic. Mitigating attention sinks and massive activations in audio-visual speech recognition with llms. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17942–17946. IEEE, 2026.

[2] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*, 2025.

[5] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969, 2021.

[6] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

[7] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[10] Alireza Dadgarnia, Soroush Tabesh, Mahdi Nikdan, Michael Helcig, Eldar Kurtic, and Dan Alistarh. Gsq: Highly-accurate low-precision scalar quantization for llms via gumbel-softmax sampling. *arXiv preprint arXiv:2604.18556*, 2026.

[11] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

[12] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc., 2022.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *The Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[14] Tim Dettmers, Ruslan A. Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations*, 2024.

[15] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

[16] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

[17] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[18] Kye Gomez. attn_res: Implementation of attention residuals. https://github.com/kyegomez/attn_res, 2026. GitHub repository. Unofficial PyTorch implementation of Attention Residuals. Accessed: 2026-04-08.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] 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 *International Conference on Learning Representations*, 2025.

[21] Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, 2020.

[22] Haoyu He, Haozheng Luo, and Qi R Wang. St-moe-bert: A spatial-temporal mixture-of-experts framework for long-term cross-city mobility prediction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge*, pages 10–15, 2024.

[23] Haoyu He, Haozheng Luo, Yan Chen, and Qi Wang. RHYTHM: Reasoning with hierarchical temporal tokenization for human mobility. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[24] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[25] Sayed Mohammadreza Tayaranian Hosseini, Amir Ardakani, and Warren J. Gross. Innerq: Hardware-aware tuning-free quantization of kv cache for large language models. *arXiv preprint arXiv:2602.23200*, 2026.

[26] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

[27] Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Robin Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. Outlier-efficient hopfield layers for large transformer-based models. In *The Forty-first International Conference on Machine Learning (ICML)*, 2024.

[28] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[29] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unraveling the enigmas of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7057–7075, 2020.

[32] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4365–4374, 2019.

[33] Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. OWQ: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13355–13364, 2024.

[34] Yixiao Li, Yuxin Yu, Qingru Zhang, Chao Liang, Pengcheng He, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024.

[35] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *Proceedings of Machine Learning and Systems*, 6: 87–100, 2024.

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[37] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Haotang Deng, and Qi Ju. FastBERT: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6035–6044, 2020.

[38] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[39] Haozheng Luo, Chenghao Qiu, Maojiang Su, Zhihan Zhou, Zoe Mehta, Guo Ye, Jerry Yao-Chieh Hu, and Han Liu. Fast and low-cost genomic foundation models via outlier removal. In *Forty-second International Conference on Machine Learning*, 2025.

[40] Haozheng Luo, Zhuolin Jiang, Md Zahid Hasan, Yan Chen, and Soumalya Sarkar. FROST: Filtering reasoning outliers with attention for efficient reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[41] Haozheng Luo, Yimin Wang, Jiahao Yu, Binghui Wang, and Yan Chen. Contrastive reasoning alignment: Reinforcement learning from hidden representations. In *Forty-third International Conference on Machine Learning*, 2026.

[42] Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[43] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *The Fifth Conference on International Conference on Learning Representations (ICLR)*, 2017.

[44] Evan Miller. Blog post: Attention is off by one, 2023. URL https://www.evanmiller.org/attention-is-off-by-one.html. Accessed: July 4, 2024.

[45] Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[46] Yuval Ran-Milo. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 70–89, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-391-3.

[47] David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.

[48] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] Moran Shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yury Nahshan, Alex Bronstein, and Uri Weiser. Robust quantization: One model to rule them all. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5308–5317. Curran Associates, Inc., 2020.

[50] 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*, 2024.

[51] Zunhai Su and Kehong Yuan. KVSink: Understanding and enhancing the preservation of attention sinks in KV cache quantization for LLMs. In *Second Conference on Language Modeling*, 2025.

[52] Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, JiaxinHu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. Flatquant: Flatness matters for LLM quantization. In *Forty-second International Conference on Machine Learning*, 2025.

[53] Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, et al. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *The Thirty-first Conference in Neural Information Processing Systems (NeurIPS)*, 2017.

[55] Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances in Neural Information Processing Systems*, 35:17402–17414, 2022.

[56] Jeffrey T. H. Wong, Cheng Zhang, Louis Mahon, Wayne Luk, Anton Isopoussu, and Yiren Zhao. On the existence and behavior of secondary attention sinks. In *ICLR 2026 Workshop on Unifying Concept Representation Learning*, 2026.

[57] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38087–38099, 2023.

[58] 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.

[59] Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations*, 2025.

[60] Bruce Changlong Xu, Adarsh Kumarappan, and Mu Zhou. Alignment collapse under KV cache quantization: A 35-minute audit for quantized LLM deployments. In *AdaptFM: Resource-Adaptive Foundation Model Inference*, 2026.

[61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[62] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[63] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. BERT loses patience: Fast and robust inference with early exit. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[64] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[65] Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 18179–18213, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1.

Appendix

A Impact Statement **16**

B Additional Related Work **16**

C Proofs of Main Text **17**

C.1 Lemma 5.1 17

C.2 Proposition 5.1 18

C.3 Theorem 5.1 18

C.4 Lemma 5.2 20

C.5 Theorem 5.2 20

D Assumptions of Main Text **22**

E Experimental System and Implementation Settings **22**

F Limitations **23**

G Additional Numerical Experiments **23**

G.1 Comparison with the original backbone 23

G.2 Comparison with Alternative Attention Normalizers 24

G.3 Global attention-sink behavior 24

H Disclosure of LLM Usage **24**

A Impact Statement

We believe that this methodology can strengthen the core robustness of foundation models, including large language models, by improving quantization and mitigating attention sinks. However, such modifications may also amplify biases present in the training data, potentially leading to unfair or discriminatory outcomes for underrepresented groups.

B Additional Related Work

Attention Sinks. Attention sinks [58] reflect a normalization-induced bias in attention allocation. They emerge during training, exhibit layer-wise hierarchy, for example, primary vs. secondary [56, 20], and are closely tied to first-token dominance, which stabilizes representations by limiting over-mixing [4]. Both empirical and theoretical work links sinks to the feasible set under softmax normalization, showing they can be functionally necessary [46]. Accordingly, StreamingLLM [58] and follow-ups [56, 20] leverage sink structure for long-context modeling. Similar sink-like concentration patterns are also observed in multimodal transformers, including vision [42, 29] and audio [1] models, revealing the fact that sink formation is a modality-agnostic consequence of attention normalization. At the system level, sinks motivate efficient designs that either preserve or suppress them. H_2O [62] exploits high-attention heavy hitters for KV-cache eviction, while DuoAttention [59] and KVSink [51] more explicitly preserve sink structure for efficient inference. Prefixing Attention Sinks [50] and Softpick [65] aim to mitigate them via architectural or normalization changes. In contrast, building on AttnResidual's softmax-based depth aggregation [53], we argue that coupled token- and depth-level normalization constraints force near no-op updates onto real tokens and branches, amplifying sink accumulation. We therefore treat sinks as a robustness issue and mitigate them via explicit null destinations and token-to-depth coupling.

Quantization. Quantization methods for transformer-based foundation models are commonly divided into weight-only quantization and weight-activation quantization [17]. In the weight-only setting, GPTQ [16] performs post-training quantization through block-wise second-order reconstruction. SpQR [14] combines sparsity with quantization to improve the compression–accuracy trade-off. OWQ [33] protects outlier-related weights through outlier-aware weight quantization. AWQ [35] emphasizes activation-aware scaling and avoids part of the hardware inefficiency of mixed-precision storage. QUIP [7] studies very low-bit quantization with theoretical guarantees. QLoRA [13] restores capability through low-rank adaptation on quantized backbones, while LoftQ [34] makes quantization explicitly aware of later LoRA fine-tuning.

For weight–activation quantization, prior work jointly compresses weights and activations, with several approaches proposed; for instance, LLM.int8() [12] uses mixed-precision decomposition to

address the problem. Outlier Suppression [55], on the other hand, relies on migration and clipping to reduce the imbalance in activation ranges. SmoothQuant [57] takes a different approach by migrating some of the activation difficulty into the weights through an equivalent transformation. More recently, OmniQuant [48] has made significant improvements to low-bit post-training quantization through omnidirectional calibration. QuaRot [2] is another technique that removes hidden-state outliers through rotation, enabling end-to-end low-bit quantization of weights, activations, and KV cache. Building on this, SpinQuant [38] replaces fixed rotations with learned rotations to further improve quantized accuracy. Meanwhile, FlatQuant [52] argues that flatness remains crucial even after prior transformations and introduces learnable affine transformations to flatten weights and activations. One approach that is particularly relevant to our setting is Prefixing Attention Sinks [50]. Instead of just calibrating activations, it shows that inserting sink-like prefix states can directly mitigate activation outliers and improve activation quantization. Other recent work on KV-cache has also highlighted the connection between sink preservation and quantization robustness. For example, KVSink [51] explores how attention sinks interact with KV-cache quantization and preserves predicted sink tokens during compression. InnerQ [25] takes this a step further by incorporating high-precision windows for both recent tokens and sink tokens in a hardware-aware KV-cache quantization scheme. What sets our work apart is that we target the structural origin of outlier-prone activations in AttnResidual [53]—namely, dual-normalized routing—rather than relying on post hoc calibration or correction.

Attention Residuals. A separate literature studies how computation should be allocated across depth. Universal Transformer [11] performs iterative refinement with shared parameters across steps. LayerDrop [15] enables structured layer removal and reduced-depth inference. FastBERT [37] uses self-distillation for adaptive inference depth. PABEE [63] introduces patience-based early exit for faster inference. More recently, Mixture-of-Depths [47] dynamically allocates computation across tokens and layers under an explicit compute budget. These methods all relax the use of fixed-depth computation, but they do so through halting, skipping, or conditional execution.

Recent work shows that quantization can corrupt low-dimensional and critical activation subspaces despite nearly unchanged perplexity [60]. When it comes to computation, AttnResidual [53] takes a distinct approach. Rather than deciding when to stop computing, it replaces the traditional method of accumulating residuals with a learned approach that aggregates the outputs of previous layers. This design allows for more flexibility in routing, but it also adds an extra layer of normalization. Our work builds on this foundation. Specifically, we explore how the dual-normalization design of AttnResidual can amplify outliers, affect sink accumulation, and create pressure on no-op operations. To address these challenges, we propose a null-aware routing mechanism that makes both token-level and depth-level aggregation more robust.

C Proofs of Main Text

C.1 Lemma 5.1

Proof of Lemma 5.1. Write

$$A_t^{(\ell)} = p_{t,o}^{(\ell)} v_o^{(\ell)} + \sum_{j \neq o} p_{t,j}^{(\ell)} v_j^{(\ell)}.$$

By the alignment assumption, for the unit vector u we have

$$\Big\langle \sum_{j \neq o} p_{t,j}^{(\ell)} v_j^{(\ell)}, u \Big\rangle = \sum_{j \neq o} p_{t,j}^{(\ell)} \langle v_j^{(\ell)}, u \rangle \geq c_1 \sum_{j \neq o} p_{t,j}^{(\ell)}.$$

Hence

$$\Big\| \sum_{j \neq o} p_{t,j}^{(\ell)} v_j^{(\ell)} \Big\| \geq c_1 \sum_{j \neq o} p_{t,j}^{(\ell)}.$$

Using

$$\|A_t^{(\ell)}\| = \Big\| \sum_{j \neq o} p_{t,j}^{(\ell)} v_j^{(\ell)} + p_{t,o}^{(\ell)} v_o^{(\ell)} \Big\| \geq \Big\| \sum_{j \neq o} p_{t,j}^{(\ell)} v_j^{(\ell)} \Big\| - p_{t,o}^{(\ell)} \|v_o^{(\ell)}\|,$$

we obtain

$$\|A_t^{(\ell)}\| \geq c_1 \sum_{j \neq o} p_{t,j}^{(\ell)} - c_0 p_{t,o}^{(\ell)} = c_1 - (c_0 + c_1) p_{t,o}^{(\ell)}.$$

Based on Assumption D.1, if $\|A_t^{(\ell)}\| \le \delta$, then

$$\delta \ge c_1 - (c_0 + c_1)p_{t,o}^{(\ell)} \iff p_{t,o}^{(\ell)} \ge \frac{c_1 - \delta}{c_0 + c_1}.$$

This establishes a necessary lower bound on the attention mass assigned to the low-norm token under the stated separation assumption. The bound approaches one as both c_0/c_1 and δ/c_1 approach zero, characterizing the attention-concentration mechanism associated with near-no-op behavior studied in OutEffHop [27].

C.2 Proposition 5.1

Proof of Proposition 5.1. By Assumption D.4, the AttnResidual leakage, concentration, and entropy-collapse terms satisfy

$$L_t^{\mathrm{AR}} \ge L_t^{\mathrm{V}}, \qquad C_t^{\mathrm{AR}} \ge C_t^{\mathrm{V}}, \qquad E_t^{\mathrm{AR}} \ge E_t^{\mathrm{V}}.$$

Since the pathology scores are defined by

$$\mathcal{S}_t^{\mathrm{AttnResidual}} = L_t^{\mathrm{AR}} + \lambda_1 C_t^{\mathrm{AR}} + \lambda_2 E_t^{\mathrm{AR}},$$

and

$$\mathcal{S}_t^{\mathrm{Vanilla}} = L_t^{\mathrm{V}} + \lambda_1 C_t^{\mathrm{V}} + \lambda_2 E_t^{\mathrm{V}},$$

with nonnegative coefficients λ_1, λ_2, it follows immediately that

$$\mathcal{S}_t^{\mathrm{AttnResidual}} \ge \mathcal{S}_t^{\mathrm{Vanilla}}.$$

This proves the theorem. Note that Proposition 5.1 is conditional: it aggregates the three routing pathologies when their component-wise dominance conditions hold, but does not assert that these conditions hold for every token or backbone. We find that entropy-collapse dominance is consistent across the evaluated models, whereas leakage and peak-concentration dominance are model dependent.

C.3 Theorem 5.1

Proof of Theorem 5.1. We follow the residual-view analysis in OutEffHop [27]: for tokens in $\mathcal{T}_{\mathrm{noop}}$, stable optimization favors a small deviation from a reference state. Let $s_{0,t} = h_t^{(1)}$ and $s_{i,t} = f_i(h_t^{(i)})$ for $i \ge 1$ denote the states supplied by the source branches, and define their deviations from a reference state h_t^{ref} as

$$u_{i,t} := s_{i,t} - h_t^{\mathrm{ref}}.$$

Since the depth-routing weights satisfy $\sum_i \alpha_{i\to\ell,t} = 1$, the depth-composed deviation is

$$\Delta h_t^{(\ell)} := h_t^{(\ell)} - h_t^{\mathrm{ref}} = \sum_{i=0}^{\ell-1} \alpha_{i\to\ell,t} u_{i,t}.$$

For no-op tokens, the near-reference condition requires

$$\|\Delta h_t^{(\ell)}\| \le \delta_t, \qquad t \in \mathcal{T}_{\mathrm{noop}},$$

with small δ_t. Because depth coefficients satisfy simplex constraints ($\alpha_{i\to\ell,t} \ge 0$, $\sum_i \alpha_{i\to\ell,t} = 1$), this becomes a constrained routing problem in which a small composed deviation may require substantial routing mass on source branches with small deviations from the reference state, i.e., no-op routes.

Let

$$i_t^* \in \arg\min_{i\in\{0,\dots,\ell-1\}} \|u_{i,t}\|.$$

By assumption, there exists a unit vector $w_t^{(\ell)} \in \mathbb{R}^d$ such that

$$\langle u_{i,t}, w_t^{(\ell)}\rangle \geq b_1, \qquad i \neq i_t^*,$$

while the minimal-update branch satisfies $\|u_{i_t^*,t}\| \leq b_0$. Projecting the residual update onto $w_t^{(\ell)}$ gives

$$\left\langle \Delta h_t^{(\ell)}, w_t^{(\ell)} \right\rangle = \sum_{i=0}^{\ell-1} \alpha_{i\to\ell,t} \langle u_{i,t}, w_t^{(\ell)}\rangle.$$

Since $\|w_t^{(\ell)}\| = 1$ and $\|\Delta h_t^{(\ell)}\| \leq \delta_t$, we have

$$\left|\left\langle \Delta h_t^{(\ell)}, w_t^{(\ell)} \right\rangle\right| \leq \delta_t.$$

On the other hand,

$$\left\langle \Delta h_t^{(\ell)}, w_t^{(\ell)} \right\rangle = \alpha_{i_t^*\to\ell,t} \langle u_{i_t^*,t}, w_t^{(\ell)}\rangle + \sum_{i\neq i_t^*} \alpha_{i\to\ell,t} \langle u_{i,t}, w_t^{(\ell)}\rangle.$$

Using $|\langle u_{i_t^*,t}, w_t^{(\ell)}\rangle| \leq \|u_{i_t^*,t}\| \leq b_0$ and $\langle u_{i,t}, w_t^{(\ell)}\rangle \geq b_1$ for $i \neq i_t^*$, we obtain

$$\left\langle \Delta h_t^{(\ell)}, w_t^{(\ell)} \right\rangle \geq -\alpha_{i_t^*\to\ell,t} b_0 + b_1 \sum_{i\neq i_t^*} \alpha_{i\to\ell,t}.$$

Combining this with $\langle \Delta h_t^{(\ell)}, w_t^{(\ell)}\rangle \leq \delta_t$ yields

$$b_1 \sum_{i\neq i_t^*} \alpha_{i\to\ell,t} \leq \delta_t + \alpha_{i_t^*\to\ell,t} b_0.$$

Using the simplex constraint $\sum_i \alpha_{i\to\ell,t} = 1$, this becomes

$$b_1(1 - \alpha_{i_t^*\to\ell,t}) \leq \delta_t + b_0 \alpha_{i_t^*\to\ell,t},$$

and hence

$$\alpha_{i_t^*\to\ell,t} \geq \frac{b_1 - \delta_t}{b_0 + b_1}.$$

Equivalently,

$$1 - \alpha_{i_t^*\to\ell,t} \leq \frac{b_0 + \delta_t}{b_0 + b_1}.$$

Therefore, routing concentration occurs in the high-separation regime where $b_0/b_1 \to 0$ and $\delta_t/b_1 \to 0$. Under this condition, $\alpha_{i_t^*\to\ell,t} \to 1$ and $\sum_{i\neq i_t^*} \alpha_{i\to\ell,t} \to 0$. Thus, the theorem establishes depth-routing concentration only under the stated high-separation regime, and it does not imply that such concentration occurs for every token or trained backbone.

C.4 Lemma 5.2

Proof of Lemma 5.2. Let $\mathcal{K}$ denote sink-prone positions from Assumption D.2. For a query token t, define the total sink mass at layer ℓ by

$$\sigma_t^{(\ell)} := \sum_{j \in \mathcal{K}} p_{t,j}^{(\ell)}.$$

In AttnResidual, effective sink contribution is depth-aggregated:

$$\Sigma_t := \sum_{\ell} \alpha_t^{(\ell)} \sigma_t^{(\ell)}.$$

Before collapse, several branches can satisfy $\sigma_t^{(\ell)} > 0$ on the same sink set $\mathcal{K}$, so nonnegative depth weights add these contributions through convex combination. Thus Σ_t is a depth-weighted average of branchwise sink masses, satisfying

$$\min_{\ell} \sigma_t^{(\ell)} \le \Sigma_t \le \max_{\ell} \sigma_t^{(\ell)}.$$

This establishes the aggregation bound.

After conditional depth collapse, there is a dominant branch ℓ^* with $\alpha_t^{(\ell^*)} \approx 1$, hence

$$\Sigma_t \approx \sigma_t^{(\ell^*)}.$$

Therefore, after conditional concentration, the effective sink mass approaches that of the selected branch. This result describes the aggregation of existing branchwise sink mass and does not assert that depth routing necessarily creates or preserves a sink.

C.5 Theorem 5.2

Proof of Theorem 5.2. For a fixed query token t and branch ℓ, define

$$S_t^{(\ell)} := \sum_{k=1}^{m} \exp z_{t,k}^{(\ell)}.$$

Under standard $\mathrm{Softmax}$, all probability mass is assigned to real tokens:

$$\sum_{j=1}^{m} p_{t,j}^{(\ell)} = 1.$$

Therefore, a near-zero attention contribution must be formed entirely through allocations over real tokens, which leads to the conditional concentration mechanism characterized in Lemma 5.1. Moreover, standard $\mathrm{Softmax}$ is invariant to a common shift of all real-token logits. For any scalar c,

$$\mathrm{Softmax}\left(z_t^{(\ell)} - c\mathbf{1}\right) = \mathrm{Softmax}\left(z_t^{(\ell)}\right).$$

Thus, jointly decreasing all real-token logits cannot reduce their total probability mass, which remains equal to one.

Under $\mathrm{Softmax}_1$, the real-token and null probabilities are

$$\tilde{p}_{t,j}^{(\ell)} = \frac{\exp z_{t,j}^{(\ell)}}{1 + S_t^{(\ell)}}, \qquad \tilde{p}_{t,\varnothing}^{(\ell)} = \frac{1}{1 + S_t^{(\ell)}}.$$

Their total real-token mass is

$$\sum_{j=1}^{m} \tilde{p}_{t,j}^{(\ell)} = \frac{S_t^{(\ell)}}{1 + S_t^{(\ell)}} = 1 - \tilde{p}_{t,\varnothing}^{(\ell)}.$$

For the same real-token logits, $\mathrm{Softmax}_1$ preserves their relative allocation while scaling their total mass:

$$\tilde{p}_{t,j}^{(\ell)} = \left(1 - \tilde{p}_{t,\varnothing}^{(\ell)}\right) p_{t,j}^{(\ell)}.$$

After shifting every real-token logit by $-c$, the probabilities become

$$\tilde{p}_{t,j}^{(\ell)}(c) = \frac{\exp\left(z_{t,j}^{(\ell)} - c\right)}{1 + e^{-c} S_t^{(\ell)}}, \qquad \tilde{p}_{t,\varnothing}^{(\ell)}(c) = \frac{1}{1 + e^{-c} S_t^{(\ell)}}.$$

As $c \to \infty$,

$$\tilde{p}_{t,j}^{(\ell)}(c) \to 0, \qquad \tilde{p}_{t,\varnothing}^{(\ell)}(c) \to 1.$$

Hence, unlike standard $\mathrm{Softmax}$, $\mathrm{Softmax}_1$ can continuously attenuate the total contribution of real tokens through its explicit null route.

To quantify this attenuation, define the real-token contribution as

$$\tilde{A}_t^{(\ell)} := \sum_{j=1}^{m} \tilde{p}_{t,j}^{(\ell)} v_j^{(\ell)}.$$

If $\|v_j^{(\ell)}\| \le V_{\max}$ for every real token j, then

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

Therefore, the sufficient condition

$$\tilde{p}_{t,\varnothing}^{(\ell)} \ge 1 - \frac{\delta}{V_{\max}}$$

guarantees

$$\left\|\tilde{A}_t^{(\ell)}\right\| \le \delta$$

without requiring any particular real token to dominate.

An analogous result holds at the depth level. Let $\tilde{\alpha}_{\varnothing,t}^{(\ell)}$ denote the depth-null mass under depth-$\mathrm{Softmax}_1$. Using the source states defined in Section 4.1, define the mixed source representation as

$$\tilde{h}_t^{(\ell)} := \sum_{i=0}^{\ell-1} \tilde{\alpha}_{i\to\ell,t} s_{i,t}.$$

If $\|s_{i,t}\| \le B_{\max}$ for every source branch i, then

$$\left\|\tilde{h}_t^{(\ell)}\right\| \le B_{\max} \sum_{i=0}^{\ell-1} \tilde{\alpha}_{i\to\ell,t} = B_{\max} \left(1 - \tilde{\alpha}_{\varnothing,t}^{(\ell)}\right).$$

Consequently, the sufficient condition

$$\tilde{\alpha}_{\varnothing,t}^{(\ell)} \geq 1 - \frac{\delta_t}{B_{\max}}$$

guarantees

$$\left\| \tilde{h}_t^{(\ell)} \right\| \leq \delta_t$$

without requiring the real depth-routing weights to concentrate on a single source branch.

These bounds show that $\mathrm{Softmax}_1$ introduces an attenuation mechanism at both routing levels. Increasing null mass reduces the probability budget available to real tokens or source branches, thereby relaxing the simplex constraints that can induce conditional concentration. The depth-level bound concerns attenuation of the mixed source representation itself; it does not, for a nonzero reference state, imply preservation of that reference state. The resulting changes in attention sinks, activation outliers, and quantization robustness are established empirically rather than implied directly by these bounds.

D Assumptions of Main Text

Assumption D.1 (No-op regime)**.** There exists a non-empty subset of token positions $\mathcal{T}_{\mathrm{noop}}$ such that for each $t \in \mathcal{T}_{\mathrm{noop}}$, the attention branch is near no-op:

$$\|A_t^{(\ell)}\| \leq \delta,$$

for small $\delta > 0$.

Assumption D.2 (Sink-prone positions)**.** There exists a set of weakly relevant positions $\mathcal{K}$ (e.g., prefix/BOS-like positions) that repeatedly receive non-trivial attention mass across queries.

Assumption D.3 (Token separation for no-op branches)**.** For each $t \in \mathcal{T}_{\mathrm{noop}}$ and branch ℓ, there exist a candidate no-op token o and a unit vector $u \in \mathbb{R}^d$ such that

$$\|v_o^{(\ell)}\| \leq c_0, \qquad \langle v_j^{(\ell)}, u \rangle \geq c_1 \quad \text{for all } j \neq o.$$

In particular, $\|v_j^{(\ell)}\| \geq c_1$ for $j \neq o$, with $0 \leq c_0 < c_1$.

Assumption D.4 (Matched pathology dominance conditions)**.** For the same token t, let

$$L_t^{\mathrm{AR}} := \sum_{\ell} \alpha_t^{(\ell)} \sum_{j \in \mathcal{N}_t} p_{t,j}^{(\ell)}, \qquad C_t^{\mathrm{AR}} := \max_{\ell,j} \alpha_t^{(\ell)} p_{t,j}^{(\ell)},$$

$$E_t^{\mathrm{AR}} := -H(\alpha_t) - \sum_{\ell} \alpha_t^{(\ell)} H(p_t^{(\ell)}),$$

and let

$$L_t^{\mathrm{V}} := \sum_{j \in \mathcal{N}_t} p_{t,j}, \qquad C_t^{\mathrm{V}} := \max_j p_{t,j}, \qquad E_t^{\mathrm{V}} := -H(p_t).$$

Assume that under matched no-op tolerance,

$$L_t^{\mathrm{AR}} \geq L_t^{\mathrm{V}}, \qquad C_t^{\mathrm{AR}} \geq C_t^{\mathrm{V}}, \qquad E_t^{\mathrm{AR}} \geq E_t^{\mathrm{V}}.$$

E Experimental System and Implementation Settings

All experiments are run on four NVIDIA H100 (80GB) GPUs with a 12-core Intel Xeon Gold 6338 CPU, using PyTorch and the Hugging Face Transformers library. For inference, we use the default system prompt with temperature 0.6, top-p 0.95, and a maximum generation length of 4096 tokens.

Table 5: **Direct comparison with matched continued-pretraining controls.** We compare each AttnResidual variant against both the original pretrained backbone and a matched Backbone+CPT control trained with the same continued-pretraining setup. Vanilla AttnResidual preserves full-precision performance but substantially amplifies activation outliers and quantization error, whereas OASIS restores stable activation statistics and improves low-bit robustness.

Model	Method	Avg. kurtosis ↓	Max inf. norm ↓	FP16 PPL ↓	W8A8 PPL ↓	GSM8K FP16 ↑	GSM8K W4A4 ↑
LLaMA-3.2-1B	Backbone (no CPT)	3.05 ± 0.02	4.72 ± 0.05	9.12 ± 0.01	9.51 ± 0.01	13.48 ± 0.15	12.65 ± 0.22
	Backbone + CPT	3.07 ± 0.08	5.20 ± 0.42	9.05 ± 0.16	9.85 ± 0.11	13.58 ± 0.16	12.76 ± 0.10
	Vanilla AttnResidual	91.71 ± 0.08	84.43 ± 0.15	8.98 ± 0.01	30.19 ± 0.06	13.64 ± 0.17	12.97 ± 0.24
	OutEffHop	3.08 ± 0.02	13.78 ± 0.08	9.07 ± 0.01	10.42 ± 0.01	13.58 ± 0.15	13.16 ± 0.21
	AoS	3.03 ± 0.01	12.54 ± 0.07	9.00 ± 0.01	9.36 ± 0.01	13.62 ± 0.13	13.34 ± 0.18
	OASIS	$\mathbf{3.02 \pm 0.02}$	$\mathbf{12.25 \pm 0.07}$	$\mathbf{8.01 \pm 0.01}$	$\mathbf{8.16 \pm 0.01}$	$\mathbf{13.75 \pm 0.12}$	$\mathbf{13.52 \pm 0.17}$
Qwen3-0.6B	Backbone (no CPT)	3.02 ± 0.02	4.58 ± 0.04	10.74 ± 0.01	11.18 ± 0.01	62.41 ± 0.20	55.83 ± 0.71
	Backbone + CPT	3.16 ± 0.13	4.77 ± 0.88	10.65 ± 0.30	11.50 ± 0.39	62.65 ± 0.09	55.79 ± 0.26
	Vanilla AttnResidual	93.04 ± 0.10	74.78 ± 0.13	10.57 ± 0.01	36.64 ± 0.08	62.58 ± 0.24	45.67 ± 0.86
	OutEffHop	3.04 ± 0.02	14.16 ± 0.06	9.99 ± 0.01	10.60 ± 0.01	62.59 ± 0.19	52.36 ± 0.74
	AoS	3.03 ± 0.02	13.90 ± 0.05	9.35 ± 0.01	9.62 ± 0.01	62.65 ± 0.17	54.21 ± 0.66
	OASIS	$\mathbf{2.95 \pm 0.02}$	$\mathbf{13.00 \pm 0.05}$	$\mathbf{9.20 \pm 0.01}$	$\mathbf{9.52 \pm 0.01}$	$\mathbf{62.68 \pm 0.15}$	$\mathbf{58.88 \pm 0.59}$

Table 6: **Comparison of Alternative Attention Normalization Mechanisms.** We compare standard Softmax with sparse alternatives, including Sparsemax and Entmax15, alongside our method on `LLaMA-3.2-1B`. We measure activation outliers using average kurtosis and maximum infinity norm $\|\mathbf{x}\|_\infty$, and assess quantization robustness using perplexity under FP16 and W8A8, together with GSM8K accuracy under FP16 and W4A4. The results indicate that while sparse variants help reduce outlier magnitude compared to standard `Softmax`, our approach achieves the strongest overall outlier suppression and the smallest degradation in low-bit quantization; best results are shown in **bold**, and second-best results are underlined.

Method	Average kurtosis	Maximum ℓ_∞ norm	Perplexity ↓		GSM8K ↑	
			FP16	W8A8	FP16	W4A4
Softmax	91.706 ± 0.083	84.433 ± 0.146	8.981 ± 0.010	30.189 ± 0.063	$\underline{13.64} \pm 0.17$	12.97 ± 0.24
Sparsemax	87.322 ± 0.088	79.991 ± 0.143	$\underline{8.552} \pm 0.013$	19.814 ± 0.061	13.61 ± 0.16	12.98 ± 0.22
Entmax15	$\underline{13.134} \pm 0.032$	$\underline{29.381} \pm 0.097$	8.706 ± 0.010	$\underline{19.600} \pm 0.056$	13.58 ± 0.14	$\underline{13.16} \pm 0.18$
OASIS	$\mathbf{3.024} \pm 0.015$	$\mathbf{12.250} \pm 0.067$	$\mathbf{8.005} \pm 0.007$	$\mathbf{8.162} \pm 0.009$	$\mathbf{13.75} \pm 0.12$	$\mathbf{13.52} \pm 0.17$

For continued pretraining, we train for 500 steps with sequence length 2048 and block size 512 via Hugging Face Accelerate, using a learning rate of 5×10^{-5} with a linear scheduler and 100 warmup steps, batch size 1 with gradient accumulation 64, gradient clipping at 1.0, and weight decay 0.1. The OASIS coupling parameter β is implemented as a learnable non-negative scalar via softplus, initialized at -5 (yielding $\beta \approx 0.007$) to ensure near no-op initialization and stable adaptation.

F Limitations

Empirically, our evaluation focuses on a limited set of backbone models. Although OASIS consistently improves outlier statistics and post-quantization performance in our experiments, broader validation on larger models and additional downstream tasks is needed in future work. Also, while the null channel can reduce no-op pressure, excessive null allocation may suppress useful weak signals or alter interpretability of attention maps. Future work should study adaptive control of the null strength, its interaction with calibration and fine-tuning, and its deployment cost in optimized inference kernels.

G Additional Numerical Experiments

G.1 Comparison with the original backbone

We isolate the effect of AttnResidual from continued pretraining (CPT) by comparing all variants with matched backbone controls in Table 5. CPT alone has little effect, changing FP16 perplexity by only **0.8%** and W8A8 perplexity by **3.6%** on LLaMA-3.2-1B and **2.9%** on Qwen3-0.6B. By contrast, Vanilla AttnResidual slightly improves FP16 perplexity but increases kurtosis by **2,891%** and **2,849%**, maximum activation magnitude by **1,525%** and **1,467%**, and W8A8 perplexity by **206.5%** and **218.6%**, respectively. OASIS largely removes this instability, reducing kurtosis by **96.7%** and **96.8%** and W8A8 perplexity by **73.0%** and **74.0%** relative to Vanilla AttnResidual, while further improving FP16 perplexity by **10.9%** and **12.9%**. It also improves GSM8K W4A4 accuracy

Table 7: **Global attention-sink statistics.** Statistics are aggregated across all layers, attention heads, query positions, and evaluation samples. OASIS consistently suppresses BOS and secondary attention sinks while increasing attention entropy and enabling explicit null routing.

Model	Method	BOS mass ↓	Sink rate ↓	Entropy ↑	Secondary-sink mass ↓	Null mass
Qwen3-0.6B	Original backbone	0.1064	0.0027	3.74	0.013	0.0000
	Vanilla AttnResidual	0.4990	0.6642	1.83	0.042	0.0000
	OASIS	**0.0013**	**0.0001**	**4.44**	**0.008**	0.0004
LLaMA-3.2-1B	Original backbone	0.1091	0.1846	3.34	0.013	0.0000
	Vanilla AttnResidual	0.5630	0.7861	1.86	0.046	0.0000
	OASIS	**0.0099**	**0.0046**	**4.55**	**0.010**	0.2185

Table 8: **Component ablation of null routing and token-to-depth coupling on Qwen3-0.6B.** We isolate the contributions of token-level $\mathrm{Softmax}_1$ null routing, depth-level $\mathrm{Softmax}_1$ null routing, and the OASIS token-to-depth coupling under W4A4 quantization. Checkmarks indicate the components enabled in each configuration: *Token only* corresponds to the OutEffHop-style token null route, *Depth only* introduces only the depth null route, and *Independent token + depth* corresponds to AoS without coupling. We report GSM8K Pass@1 (%), where higher is better.

Configuration	Token null routing	Depth null routing	Token-to-depth coupling	W4A4 GSM8K ↑
AttnResidual	×	×	×	45.67
Token only (OutEffHop)	✓	×	×	52.36
Depth only	×	✓	×	48.90
Independent token + depth (AoS)	✓	✓	×	54.21
Full OASIS	✓	✓	✓	**58.88**

by **4.2%** and **28.9%** over Vanilla AttnResidual. These results attribute the low-bit degradation to the AttnResidual architecture, rather than CPT, and show that OASIS preserves its full-precision benefit while restoring activation and quantization robustness.

G.2 Comparison with Alternative Attention Normalizers

To isolate whether the gains of OASIS can be attributed to alternative attention normalization schemes, we tested on LLaMA-3.2-1B. Specifically, we compare Softmax, Sparsemax, Entmax15, and OASIS using average kurtosis, maximum ℓ_∞ norm, perplexity, and GSM8K Pass@1 (%). As shown in Table 6, sparse normalizers reduce outliers relative to Softmax, with Entmax15 as the strongest sparse baseline. OASIS achieves a **96.70%** reduction in average kurtosis and an **85.49%** reduction in maximum ℓ_∞ norm compared with Softmax, indicating the strongest overall outlier suppression. Moreover, OASIS reduces perplexity degradation by **99.26%** and reduces the GSM8K quantization-induced accuracy drop by **65.67%**.

G.3 Global attention-sink behavior

We further quantify attention-sink behavior globally by aggregating statistics across all layers, heads, query positions, and evaluation samples. As shown in Table 7, Vanilla AttnResidual strongly amplifies BOS concentration: compared with the original backbone, BOS mass increases from 0.1064 to 0.4990 on Qwen3-0.6B and from 0.1091 to 0.5630 on LLaMA-3.2-1B, accompanied by sink rates of **66.4%** and **78.6%**, respectively. OASIS suppresses BOS mass by **99.7%** and **98.2%** relative to Vanilla AttnResidual and reduces the corresponding sink rates by **99.98%** and **99.4%**. It also reduces secondary-sink mass by **81.0%** and **78.3%**, while increasing attention entropy by **142.6%** and **144.6%**. These global statistics confirm that the sink suppression observed locally is a model-wide effect rather than an artifact of selected layers or heads. The nonzero null mass further indicates that OASIS can redirect otherwise concentrated attention into an explicit no-op route, with the extent of null routing varying across models.

H Disclosure of LLM Usage

In this work, we use large language models (LLMs) to improve the conciseness and precision of the writing. We also use LLMs to assist with theoretical derivations in Section C and to better understand related literature (e.g., via Claude).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims made in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide theory in Section 5.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the details necessary to reproduce the main experimental results in the paper. The details of the experiments are included in both the main paper and the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and data, including detailed instructions to reproduce the main experimental results, are publicly available at https://github.com/robinzixuan/OASIS.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We introduce the experimental setting in main paper and provide additional details in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report all table values and observe stable results across three experimental runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We introduce the compute resources in Section E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed and followed the NeurIPS Code of Ethics throughout our research process.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Section A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our model does not have such a risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly credited all assets used in our research and respected their licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release a new assets in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not include crowdsourcing experiments or research involving human subjects in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not include crowdsourcing experiments or research involving human subjects in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We disclose LLM usage in Section H.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.