

A Unified View of Attention and Residual Sinks: Outlier-Driven Rescaling is Essential for Transformer Training

Zihan Qiu^{*1}, Zeyu Huang^{*2}, Kaiyue Wen^{*3}, Peng Jin^{*1}, Bo Zheng^{*1},
Yuxin Zhou¹, Haofeng Huang^{1,4}, Zekun Wang¹, Xiao Li¹, Huaqing Zhang^{1,4},
Yang Xu¹, Haoran Lian¹, Siqi Zhang¹, Rui Men¹, Jianwei Zhang¹,
Ivan Titov², Dayiheng Liu^{†1}, Jingren Zhou¹, Junyang Lin^{†1}

¹Qwen Team ²University of Edinburgh ³Stanford University ⁴Tsinghua University

*Equal contribution. †Corresponding authors.

Abstract

We investigate the functional role of emergent outliers in large language models, specifically attention sinks (a few tokens that consistently receive large attention logits) and residual sinks (a few fixed dimensions with persistently large activations across most tokens). We hypothesize that these outliers, in conjunction with the corresponding normalizations (e.g., softmax attention and RMSNorm), effectively rescale other non-outlier components. We term this phenomenon *outlier-driven rescaling* and validate this hypothesis across different model architectures and training token counts. This view unifies the origin and mitigation of both sink types. Our main conclusions and observations include: (1) Outliers function jointly with normalization: removing normalization eliminates the corresponding outliers but degrades training stability and performance; directly clipping outliers while retaining normalization leads to degradation, indicating that outlier-driven rescaling contributes to training stability. (2) Outliers serve more as rescale factors rather than contributors, as the final contributions of attention and residual sinks are significantly smaller than those of non-outliers. (3) Outliers can be absorbed into learnable parameters or mitigated via explicit gated rescaling, leading to improved training performance (average gain of 2 points) and enhanced quantization robustness (1.2 points degradation under W4A4 quantization).

1 Introduction

Transformer-based Large Language Models (LLMs) exhibit outliers. These extreme values, exceeding regular activations or logits by orders of magnitude, pose practical challenges: they dominate the dynamic range of the representations during model quantization (Yao et al., 2022; Xiao et al., 2023b;a; Wei et al., 2023; Abecassis et al., 2025), and could lead to larger numerical errors in floating-point arithmetic (Budzinskiy et al., 2025). However, simply removing them through clipping severely degrades model performance (Kovaleva et al., 2021; Sun et al., 2024), suggesting that they play an essential functional role in transformers.

A prominent instance of outliers is the *attention sink* (Xiao et al., 2023b), where a small subset of attention logits becomes larger than the rest, causing a few special tokens (sink tokens) to consistently receive high attention scores. Recent work reveals that their formation is intrinsically linked to softmax normalization (Bondarenko et al., 2023; An et al., 2025), and their corresponding value vectors exhibit significantly smaller norm than those of non-sink tokens (Sun et al., 2024; An et al., 2025), indicating that these sink tokens do not dominate the attention output with their abnormally large attention score, but leverage it as the scaling factor within softmax normalization in attention. This view is further supported by the introduction of GatedAttention (GA) (Bondarenko et al., 2023; Qiu et al., 2025b; An et al., 2025), where an explicit gating mechanism enables the model to perform such rescaling, thereby mitigating the reliance on attention sinks. Another notable outlier phenomenon is massive activation (MA) (Sun et al., 2024) in the residual stream: tokens associated with attention sinks often exhibit extremely large activations in specific dimensions, which, after passing through normalization layers, can promote the formation of attention sinks (An et al., 2025).

Both types of outliers above share a key characteristic: they exert their effects via normalization. We unify this behavior under the term *outlier-driven rescaling*, in which outliers interact with normalization to rescale the non-outlier components after normalization. We validate the outlier-driven rescaling hypothesis on a distinct type of outliers within the residual (Dettmers et al., 2022; Bondarenko et al., 2023). These outliers appear in a fixed set of dimensions across the vast majority of tokens, exhibiting activations that are orders of magnitude larger than typical values, as shown in Fig. 1. Our experiments show that these outliers share many properties with attention sinks, which motivates us to name them *residual sinks*. Notably, residual

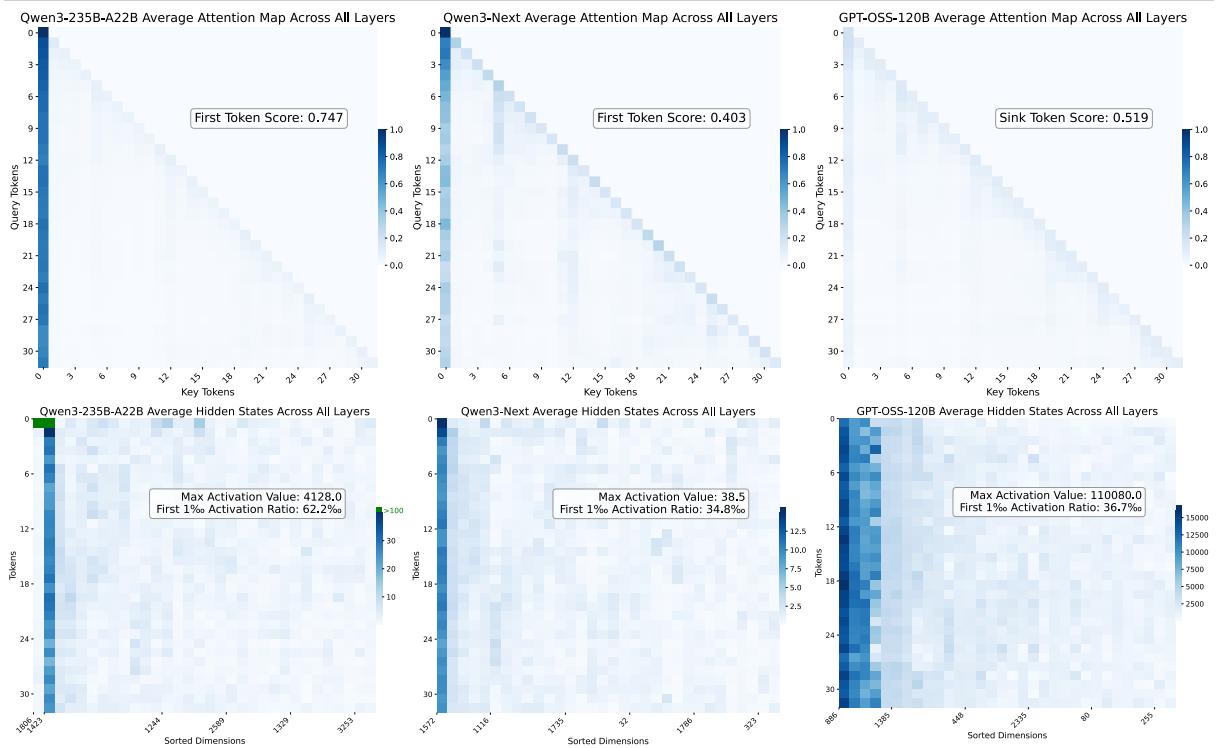


Figure 1: In the first row, all models exhibit varying degrees of attention sinks: the first token produces attention logits significantly larger than those of other tokens, dominating the attention scores. In the second row, Qwen3-235B-A22B shows massive activations, with dimensions 1806 and 1423 of the first token exceeding 1000. Beyond these extreme values, all models display consistent residual sinks: certain fixed dimensions yield persistently higher activations across all tokens compared to others.

sinks are not tied to specific inputs, suggesting they are not passing data-specific features. We further verify that they interact with the RMSNorm layers (Zhang & Sennrich, 2019) to perform outlier-driven rescaling. Extensive empirical evidence supports this hypothesis, as detailed below.

(1) Removing normalization (Zhu et al., 2025) reduces residual sinks but hurts model performance and training stability (Sec. 3.1). Suppressing outliers via clipping or architectural modifications also yields degradation (Sec. 3.2).

(2) We observe that in models that rely on outlier-driven rescaling, the RMSNorm weights for outlier-prone dimensions are consistently much smaller than the mean (e.g., 0.006 vs. 1), further suggesting that these outlier dimensions act primarily as rescale factors rather than direct contributors to the normalized output. *We prove the upper bound on the feature norm after RMSNorm decreases as the outlier magnitude increases given this property* (App. A.1).

(3) Residual sink can be absorbed in parameters, analogous to the absorption of attention sinks into learnable biases (Sun et al., 2024; Agarwal et al., 2025), eliminating the need for explicit outliers in the residual stream (Sec. 3.3).

(4) If rescaling is the underlying purpose, explicitly introducing it is expected to reduce outliers. Consistent with prior work showing that gating-based rescaling in attention reduces attention sinks (Bondarenko et al., 2023; An et al., 2025; Qiu et al., 2025b), our experiments demonstrate that inserting a lightweight gating after RMSNorm (GatedNorm) effectively mitigates residual sinks while preserving or even enhancing model performance (Sec. 3.4). With fewer outliers and smoother activations, the model trained with GatedNorm exhibits better quantization performance.

(5) Gating-based rescaling reduces the model’s reliance on outliers, thereby weakening sensitivity to outlier-inducing architectural choices. For example, while SwiGLU (Shazeer, 2020) typically outperforms sigmoid-based GLU in FFNs by generating larger activations that support outlier-driven rescaling, adding gating-based rescaling enables GLU to match or exceed SwiGLU (Fig. 3).

We validate our hypothesis across a wide range of models, including standard softmax attention (Vaswani et al., 2017), linear attention (Yang et al., 2024), linear-full attention hybrid architectures with 1B, 7B, and 24B parameters, trained on datasets ranging from 120B to 1T tokens. In the App. A.2, we summarize a comparison of attention sinks and residual sinks under the outlier-driven rescaling perspective, covering their patterns, functional roles, and mitigations. Our results consistently indicate that outliers at normalizations in softmax attention and RMSNorm are not pathological artifacts but rather rescaling-driven factors. By understanding the functional role of the residual sink, we propose simple, effective alternatives

that reduce residual sink while preserving their functional benefits, thereby improving training and quantization performance.

2 Outliers in Large Language Models

In this section, we present outliers in open-source LLMs and examine their interrelations. We focus on pre-norm transformers. For an input sequence of length L , the model first embeds tokens into hidden states $\mathbf{H}_0 \in \mathbb{R}^{L \times d}$, where d is the hidden dimension. The states are then processed through D layers. Denoting the i -th layer function as F_i , the hidden states are updated as:

$$\mathbf{H}_{i+1} = \mathbf{H}_i + F_i(\mathbf{H}_i), \quad \text{for } i = 0, 1, \dots, D - 1.$$

The hidden states \mathbf{H}_i , also referred to as the residual stream, constitute the primary focus of our study.

We analyze four models with distinct outlier patterns: Qwen3-235B-A22B (Yang et al., 2025), Qwen3-Next (Yang et al., 2025), GPT-OSS (Agarwal et al., 2025), and DeepSeek-V3 (Liu et al., 2024) in Fig. 1. For each model, we use the same set of input sequences and record both the attention maps and the residual activations across all layers. In Fig. 1, we average attention maps and hidden states over all layers for clarity. Because the hidden dimension d is large, visualizing all dimensions is impractical. To highlight outlier patterns, we reorder the feature dimensions by their overall activation magnitude. Specifically, for each dimension j , we compute its average absolute activation across all tokens, layers, and inputs:

$$\mathbf{H}_{\text{avg}}^j = \frac{1}{N(D+1)} \sum_{n=0}^{N-1} \sum_{i=0}^D |\mathbf{H}_{i,n}^j|,$$

where $\mathbf{H}_{i,n}^j$ is the activation of the j -th dimension for the n -th token at the i -th layer, and N is the number of tokens. We then sort dimensions in descending order of \mathbf{H}_{avg} . After this reordering, dimensions with consistently large activations appear on the left side of our visualizations, forming a structure that closely resembles the attention sink pattern.

In Qwen3-235B-A22B, the first token consistently receives high attention scores from nearly all other tokens, exhibiting *attention sink*. Correspondingly, this token shows MA in two dimensions (e.g., dimension 1806 and 1423). Beyond token-specific MA, we observe that *most* tokens exhibit consistently large activations in dimensions 1423. The pattern is stable across inputs, appearing as a dark vertical stripe in activation visualization, similar to shape of the attention sinks. We later find that its rescaling effect during normalization closely resembles that of the attention sink, leading us to term this *residual sink*. A similar pattern is observed in Deepseek-V3 in App. 8: attention sinks, MA, and residual sinks all exist.

Qwen3-Next introduces a gating in the attention to perform gating-based rescaling, thereby reducing its reliance on attention logit outliers. As a result, attention sinks are weaker than others. Moreover, the maximum activation magnitude in its residual stream is only 38.5 and no prominent MA are observed. Nevertheless, dimension 1572 consistently exhibits large activations higher than all other dimensions across tokens, clearly manifesting a residual sink.

GPT-OSS incorporates learnable sinks, effectively removing attention sinks from real input tokens. MA also disappear: the hidden states of real tokens no longer exhibit extreme values in specific dimensions. This aligns with prior interpretation: attention sinks act as an input-independent bias in softmax attention; MA enable this by producing near one-hot vectors after normalization, which activate only a few fixed matrix columns when projected into the key space. When a learnable bias (e.g., a dedicated sink key) is provided explicitly, the model no longer needs to generate MA from real tokens. However, despite the absence of attention sinks and MA, residual sinks persist in GPT-OSS.

3 Outlier-Driven Rescaling

In this section, we provide a detailed discussion and empirical evidence on the roles of these outliers. We conduct our experiments on the pre-norm transformer, closely following the design of dense models in Llama3 (Dubey et al., 2024) and Qwen3 Yang et al. (2025). Due to the large number of ablation and comparison conditions, we primarily evaluate all variants under a consistent setting: a 2B-parameter model trained on 120B tokens. When structural changes affect the model parameters, we adjust the FFN width accordingly to maintain a constant total parameter count. Full experimental setting details are provided in the App. A.4. Our analysis is organized into five parts.

In Sec. 3.1, we show that replacing normalization layers with point-wise functions such as Dynamic Tanh (DyT) (Zhu et al., 2025; Chen et al., 2025) significantly reduces outliers. However, as DyT cannot provide outlier-driven rescaling, both training stability and final performance degrade.

Table 1: Performance comparison under different rescaling strategies. ‘GA’ denotes Gated Attention. **Attn R** and **Norm R** refer to the rescaling mechanisms in softmax attention and RMSNorm, respectively. ‘Gating’ indicates learned gating-based rescaling; ‘Outlier-Driven’ denotes outlier-driven rescaling; ‘Restrict’ means rescaling is constrained due to activation clipping. ‘DyT’ refers to the pointwise Dynamic Tanh function. ‘GLU’ is a SwiGLU variant with sigmoid activation; all other configs use SwiGLU. LR denotes peak learning rates. **ID_s** denotes row ID, **C IDs** denotes the compared row ID. For **Outliers**, we retain the first two significant digits in the table. Deeper blue means larger activation magnitude; for **Final Loss**, deeper green means lower loss. **Gap**, denotes the relative loss gap between **ID_s** and **C IDs**, positive values indicate degradation and negative values indicate improvement. ‘-’ denotes divergence.

IDs	Basic Config	Additional Config	Attn R	Norm R	LR	Outliers	Final Loss	ID _s	C IDs	Gap
Full Attention										
(1)	Full Attention	-	Outlier-Driven	Outlier-Driven	4.3×10^{-3}	6,000	1.964	(1)	-	-
(2)	Full Attention	GA	Gating	Outlier-Driven	4.3×10^{-3}	2,800	1.957	(2)	(1)	-0.007
Linear & Hybrid Attention (Remove token mixing normalizations.)										
(3)	Linear Attention	-	None	Outlier-Driven	4.3×10^{-3}	510	1.933	(3)	-	-
(4)	Hybrid	-	Outlier-Driven	Outlier-Driven	4.3×10^{-3}	1,800	1.926	(4)	(3)	-0.007
(5)	Hybrid	GA	Gating	Outlier-Driven	4.3×10^{-3}	1,100	1.921	(5)	(4)	-0.005
Dynamic Tanh (Replace normalizations with pointwise function.)										
(6)	DyT	-	None	None	1.0×10^{-3}	-	-	(6)	-	-
(7)	DyT	-	None	None	5.0×10^{-4}	73	2.216	(7)	(1)	+0.259
(8)	DyT	GA	Gating	Outlier-Driven	2.0×10^{-3}	32	2.041	(8)	(2)	+0.084
(9)	DyT	GA, GateDyT	Gating	Gating	2.0×10^{-3}	53	1.969	(9)	(20)	+0.018
Clipping (Directly constrain outliers.)										
(10)	Full Attention	clip 10	Restrict	Restrict	4.3×10^{-3}	-	-	(10)	(1)	-
(11)	Full Attention	clip 100	Restrict	Restrict	4.3×10^{-3}	-	-	(11)	(1)	-
(12)	Full Attention	clip 1000	Restrict	Restrict	4.3×10^{-3}	1,000	1.970	(12)	(1)	+0.006
(13)	Full Attention	GA, clip 10	Restrict	Restrict	4.3×10^{-3}	10	1.960	(13)	(2)	+0.003
(14)	Full Attention	GA, clip 1000	Restrict	Restrict	4.3×10^{-3}	1,000	1.958	(14)	(2)	+0.001
GLU Variants (Constrain outliers through architecture modifications.)										
(15)	Full Attention	GLU	Restrict	Restrict	4.3×10^{-3}	1,300	1.975	(15)	(1)	+0.011
(16)	Full Attention	GA, GLU	Gating	Restrict	4.3×10^{-3}	800	1.955	(16)	(2)	-0.002
(17)	Full Attention	GA, PreAffine, GLU	Gating	Outlier-Driven	4.3×10^{-3}	150	1.952	(17)	(19)	-0.002
(18)	Full Attention	GA, GatedNorm, GLU	Gating	Gating	4.3×10^{-3}	280	1.948	(18)	(20)	-0.003
PreAffine & GatedNorm (Residual sink reduction methods.)										
(19)	Full Attention	GA, PreAffine	Gating	Outlier-Driven	4.3×10^{-3}	640	1.954	(19)	(2)	-0.003
(20)	Full Attention	GA, GatedNorm	Gating	Gating	4.3×10^{-3}	430	1.951	(20)	(2)	-0.006
(21)	Linear Attention	GatedNorm	None	Gating	4.3×10^{-3}	110	1.929	(21)	(3)	-0.004
(22)	Hybrid Attention	GA, GatedNorm	Gating	Gating	4.3×10^{-3}	780	1.918	(22)	(5)	-0.003

In Sec. 3.2, we demonstrate that even when normalization is retained, directly constraining outliers (via activation clipping) breaks the outlier-driven rescaling mechanism and harms model performance, sometimes causing training divergence. This also explains why architectural changes that suppress outlier generation, such as using sigmoid-based GLU variants, tend to underperform (Shazeer, 2020).

In Sec. 3.3, we show that outliers can be losslessly transferred from activations into learnable parameters: by introducing a lightweight learnable vector before normalization, the model can still perform outlier-driven rescaling without requiring large values in the residual stream, but use the amplified projections after the learnable vector.

In Sec. 3.4, we show that enabling rescaling via gating reduces residual sinks without performance degradation.

In Sec. 3.5, we find that once gating-based rescaling is introduced and the model’s reliance on outliers is reduced, its sensitivity to architectural choices diminishes. Specifically, DyT achieves stable convergence, and sigmoid-based GLU matches or even surpasses SwiGLU regarding performance.

3.1 Removing Normalizations Reduces Outliers with Degraded Stability and Performance

Normalizations in transformer layers are in two places: softmax in attention and normalization layers (e.g., RMSNorm). In attention, prior work shows that softmax normalization is a primary cause of attention sinks (Gu et al., 2024b; An et al., 2025). When softmax is replaced with sigmoid attention (Gu et al., 2024b), or when the denominator is combined with a learnable bias (Sun et al., 2024; Dong et al., 2024; Agarwal et al., 2025), attention sinks disappear.

Existing work finds that tokens exhibiting attention sinks produce value vectors with smaller norms than other tokens (Sun et al., 2024; An et al., 2025). From the perspective of outlier-driven rescaling, the presence of the attention sink allows the model to adjust the relative contribution of near-zero contributions (from sink tokens) versus others in the attention output, controlling the norm of the attention result. This interpretation also explains the training instability observed in sigmoid attention: without the normalization-induced rescaling, initial attention outputs have large norms (Ramapuram et al., 2024).

Our experiments find that replacing softmax-based attention with linear attention (Yang et al., 2024)

(without normalization in the token-mixing step) also reduces MA. As shown in rows (3)–(5) of Tab. 1, the peak activation drops to 510 for the linear attention model and 1100 for a hybrid model combining full and linear attention in a 1:3 ratio, compared to 6000 for the full attention baseline. Notably, while linear attention eliminates MA, residual sinks persist.

For normalization layers in the residual, prior work shows that replacing RMSNorm with DyT—defined as $DyT(x) = \gamma \cdot \tanh(\alpha x) + \beta$ —significantly reduces outliers (He et al., 2024; Owen et al., 2025a). The key difference lies in rescaling: in RMSNorm, each dimension is scaled based on statistics of the entire hidden state (e.g., root-mean-square), allowing the outlier in one dimension to influence all other dimensions. In contrast, DyT uses only pointwise operations, so no dimension directly influences another.

Our experiments confirm this limitation. When training DyT models with the same learning rate as the baseline, optimization diverges rapidly. To address this, we conduct hyperparameter sweeps for all DyT inclusive settings, evaluating learning rates in $\{5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 4 \times 10^{-3}\}$ and reporting the best performing configuration. The original DyT variant only converges at the smallest learning rate 5×10^{-4} , yielding a peak activation magnitude of 73 but suffering a significant performance drop of +0.259 in loss compared to the baseline (Tab 1 row (7) versus row (1)). This suggests that the outlier-driven rescaling effect is essential for both training stability and final performance.

Taken together, these results show that removing normalization breaks the outlier-driven rescaling mechanism. Consequently, the model stops generating outliers, but usually at the cost of degraded training stability and performance.

3.2 Directly Clipping or Constraining Outliers Hurts Stability and Performance

We now examine the effect of suppressing outliers while preserving normalization. To isolate the impact of different outliers, we consider two settings: (1) models exhibiting attention sinks, MA, and residual sinks; (2) models where attention sinks and MA are reduced via Gated Attention (GA) (Bondarenko et al., 2023; An et al., 2025; Qiu et al., 2025b), exhibiting primarily residual sinks.

First, we apply activation clipping to the residual of the full attention baseline (Tab. 1, row (1)), capping activations above a threshold at that value. When the clipping threshold is set to 100 or lower, training diverges early. At a threshold of 1000, the loss curve shows frequent spikes and converges to a higher final loss (+0.006, row (12)). This aligns with prior observations that clipping MA or attention sinks in the models after training severely degrades performance, often yielding near-random outputs.

Second, we combine GA, which reduces attention sinks and MA, with residual clipping. We find that once explicit rescaling is reintroduced via GA in the softmax attention, even aggressive clipping (at 10) permits convergence but still harms performance (+0.003, row (13)). Moreover, under the same setting, clipping at 1000 incurs a much smaller performance drop (+0.001, row (14)).

These results imply two key points: (i) The outlier-driven rescaling mechanism in attention is the primary source of instability when disrupted; (ii) Residual sinks also contribute meaningfully to performance, and directly constraining them without compensation leads to degradation.

Beyond clipping, we explore a less intrusive method to limit outliers by modifying a seemingly unrelated component, the activation function, thereby further validating the above conclusions. Prior work observes that outliers in transformers predominantly originate from FFN (Oh et al., 2024; Yona et al., 2025). Most modern models employ Gated Linear Units variants, particularly SwiGLU (Shazeer, 2020):

$$\text{SwiGLU}(\mathbf{x}) = \mathbf{x}_{\text{down}}((\mathbf{W}_{\text{up}}\mathbf{x}) \odot \text{swish}(\mathbf{W}_{\text{gate}}\mathbf{x})).$$

In SwiGLU, outliers emerge in the terms $\mathbf{W}_{\text{up}}\mathbf{x} \odot \text{swish}(\mathbf{W}_{\text{gate}}\mathbf{x})$, and are amplified by \mathbf{W}_{down} .

Conceptually, replacing the swish activation with sigmoid—which has a bounded range of (0, 1)—naturally constrains the magnitude of FFN outputs and effectively suppresses outlier generation. This modification reduces SwiGLU to standard GLU. As shown in Tab. 1, GLU exhibits significantly smaller outlier magnitudes at convergence compared to SwiGLU (1300 vs. 6000), but incurs a performance drop of +0.011. This aligns with prior findings that GLU variants using sigmoid underperform those using swish or GELU (Shazeer, 2020). In Sec. 3.5, we further observe that when the model’s reliance on outlier-driven rescaling is reduced through explicit rescaling like gating, GLU can even slightly outperform SwiGLU.

3.3 Fusing Outliers into Parameters

The previous sections establish that outliers serve a functional role in normalizations, and removing them without compensation harms performance. Theoretically, we proof the upper bound on the feature norm after RMSNorm decreases as the outlier increases (App. A.1), allowing outliers to rescale the feature norm. We now ask: *can we preserve this functionality while reducing explicit outliers?*

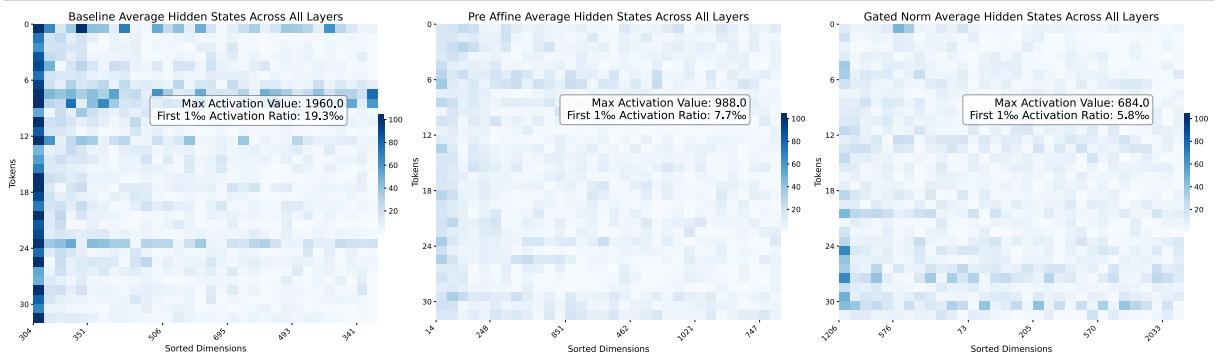


Figure 2: Reduced residual sinks. For the same input, the baseline (left) has dimension 304 consistently produces large activations across all tokens. This phenomenon is substantially mitigated in both the PreAffine (middle) and GatedNorm (right) variants.

Prior work shows that learnable sink tokens can shift attention sinks from input logits to fixed parameters (Sun et al., 2024; Dong et al., 2024; Agarwal et al., 2025), removing outliers from real tokens while preserving the outlier-driven rescaling. Inspired by this, we introduce a learnable element-wise scaling vector (termed *PreAffine*) before RMSNorm. Specifically:

$$\text{PreAffineRMSNorm}(\mathbf{x}) = \text{RMSNorm}(\lambda_1 \odot \mathbf{x}),$$

where $\lambda_1 \in \mathbb{R}^d$ is a trainable vector. Since residual sinks appear consistently in the same dimensions across nearly all tokens, λ_1 can learn to amplify those specific dimensions. Thus, even if the input activation \mathbf{x} contains no outliers, the scaled input $\lambda_1 \odot \mathbf{x}$ can still contain large values in selected dimensions, enabling outlier-driven rescaling as usual.

After adding PreAffine, the maximum activation in the network decreases from 2800 to 640, and the final loss slightly improves (-0.003, Tab. 1, row (19)). We also evaluate various outlier reduction methods on a 24.6B-A1.7B hybrid MoE model (configuration detailed in Section 4) and report the corresponding activation statistics in Fig. 2. The left panel shows the baseline with GA suppressing MA; here, dimension 304 consistently exhibits higher activation than other dimensions, indicating a residual sink. In the middle panel, we apply PreAffine to the same model. The persistent high activation in any single dimension disappears, and the peak activation drops from 1960 to 988.

Importantly, the rescaling capability of λ_1 differs from that of the standard RMSNorm parameter λ . λ_1 interacts with RMSNorm: large values in a few dimensions of λ_1 control the RMS of the scaled input, thereby rescaling the strength of non-outlier dimensions. We analyze the parameters λ (the standard RMSNorm weight) and λ_1 (the PreAffine) across models; full results are provided in App. A.5.1. We identify the dimensions with the largest deviations from 1, as these affine parameters exert the strongest influence. Our observations are as follows:

- (1) In the baseline model, most dimensions of λ remain close to 1. However, dimension 304 consistently deviates from 1 across all layers, reaching a minimum value of 0.004. Notably, the residual sink in the baseline is also in dimension 304. This indicates that the large activation caused by the residual sink is immediately scaled down after normalization. This suggests that once a dimension fulfills its role in outlier-driven rescaling, its downstream influence is intentionally dampened. This mirrors the observation that the attention sink token’s value vectors exhibit smaller norms.
- (2) In the model equipped with PreAffine, the deviation of λ_1 from 1 is significantly larger than that of λ . For example, λ_1 in dimension 1326 reaches 7.19, while the corresponding λ in the same dimension is only 0.06. This further supports the view that outliers are used to shape representations with normalization, and their direct contribution is suppressed.

3.4 GatedNorm: Explicitly Enabling Rescaling

Although PreAffine reduces outliers in the residual stream, outliers still appear within the normalization computation (i.e., in $\lambda_1 \odot \mathbf{x}$). To address this, we draw inspiration from GA and introduce GatedNorm: an element-wise low-rank self-gating mechanism applied after every normalization layer. Formally, given $\mathbf{y} = \text{RMSNorm}(\mathbf{x})$, we compute:

$$\mathbf{y}_g = \sigma(\mathbf{W}_{\text{up}}(\text{swish}(\mathbf{W}_{\text{down}}(\mathbf{y})))), \quad \mathbf{y}' = \mathbf{y}_g \odot \mathbf{y},$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times r}$, $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r \times d}$, $r \ll d$ (e.g., $r = 16$), σ is sigmoid activation.

Notably, GatedNorm adds only 3.7M parameters, which is approximately 2% of the total in a 2B model. To maintain parameter parity, we slightly reduce FFN capacity. In the 2B dense model GatedNorm incurs

about 5% latency overhead, and this overhead further decreases as model size increases, especially in MoEs. More detailed performance analysis is provided in the App. A.3. We examine GatedNorm on top of several settings in Tab. 1 that still exhibit residual sinks. As shown in rows (20)–(22), GatedNorm further reduces both loss and outlier magnitude of models with GatedAttention (including full attention, hybrid attention, and linear attention).

Figure 2 (right) shows that, on a 24.6B-A1.7B hybrid MoE model, GatedNorm also suppresses residual sinks. We analyze the learned scaling parameters λ in models using GatedNorm (App. A.5.1, Fig. 7). The maximum deviation of λ from 1 is only 0.73, compared with 0.004 in the baseline and 0.06 in the PreAffine model, indicating that when the network no longer relies on outlier-driven rescaling, the need to suppress any particular dimension after normalization disappears. Consequently, normalization outputs become smoother and quantization-friendly.

We also compare different gating variants, focusing on two design choices: the gating granularity (elementwise (score shape d) versus tensorwise (score shape 1)) and the activation function (sigmoid, tanh, silu, identity). Our findings are as follows.

First, elementwise gating with a sigmoid activation yields the most significant performance improvement over the baseline. Tensorwise gating with sigmoid reduces residual sinks to a similar extent as elementwise gating, but its final performance is close to the baseline and consistently inferior to the elementwise variant. This suggests that while both granularities can mitigate outliers, which supports the outlier-driven rescaling hypothesis, finer-grained (elementwise) rescaling enables more effective modulation and thus better performance.

Second, when using elementwise gating, replacing sigmoid with tanh, SiLU, or no activation (similar to adaLN (Perez et al., 2018; Xu et al., 2019; Peebles & Xie, 2023; Karras et al., 2024) in DiT) leads to unstable outlier dynamics during training. This implies that the bounded nature of sigmoid and its fine-grained control near zero are beneficial for stable rescaling, consistent with Chen et al. (2025). Moreover, under tensorwise gating, all non-sigmoid activations, including tanh, SiLU, or none, cause training divergence, further highlighting the necessity of a well-behaved, bounded activation like sigmoid for stable gating.

3.5 GatedNorm Improve Robustness to Architecture Choice

Sec. 3.1 and 3.2 show that both DyT and sigmoid-based GLU underperform the baseline. One possible explanation is their architectures inherently restrict the outlier-driven rescaling mechanism. In this section, we investigate whether explicitly providing rescaling, via GatedNorm, can recover their performance by reducing reliance on outliers.

We first equip DyT with GA (row 8 in Tab. 1) to provide explicit rescaling in the attention module. This enables the model to train stably at the baseline’s learning rate ($4.3e-3$), though its optimal learning is $2e-3$. This further confirms the critical role of attention rescaling in training stability. We then apply a low-rank self-gating mechanism after the DyT layer, analogous to GatedNorm, resulting in GatedDyT (row 9 in Tab. 1). With this addition, the performance gap between DyT and the RMSNorm baseline narrows from 0.084 to 0.018, highlighting the importance of explicit rescaling not only in attention but also at the normalization layer itself.

We further compare SwiGLU and GLU before and after incorporating GatedNorm. As shown in Fig. 3 (bottom), vanilla SwiGLU generates maximum activations exceeding 4×10^4 during training, while vanilla GLU peaks around 1×10^4 . Lacking outlier-driven rescaling, GLU converges to a higher loss (+0.011). After employing GatedNorm, GLU slightly outperforms SwiGLU under the same setup (-0.002 in row 16; -0.003 in row 17).

Fig. 3 (top) shows that, with gating, the GLU loss curve improves from the worst (blue) to the best (green) among all variants.

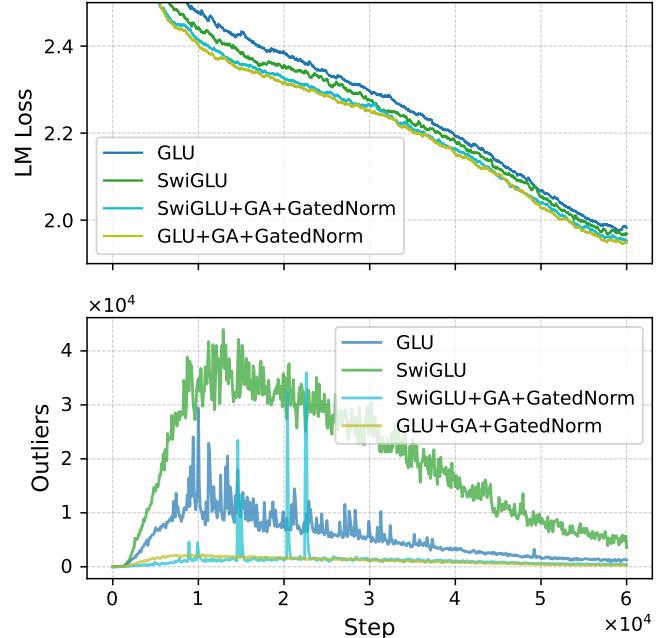


Figure 3: SwiGLU and GLU with different rescaling method.

Table 2: Performance of different configurations. ‘GA’ denotes Gated Attention. ‘MMLU-R’ denotes MMLU-Redux; ‘MMLU-P’ denotes MMLU-Pro; ‘GPQA-D’ denotes GPQA-Diamond; ‘S-GPQA’ denotes SuperGPQA. ‘W4A4’ indicates 4-bit weight and activation quantization; ‘SQ’ denotes smooth quantization. All models are based on the 7B-A2B or 24B-A3B architecture unless otherwise noted.

Type	Add Config	Knowledge			STEM			Code			Multilingual			
		MMLU-R	MMLU-P	S-GPQA	GPQA-D	GSM8k	Math	Crux	MultiPL_E	MBPP	MMMLU	MGSM	Avg	
MoE-7B-A2B, 1.2T tokens														
(1)	BF16	GA	59.59	31.91	16.39	26.93	66.22	45.36	44.38	34.63	50.04	46.16	37.40	41.73
(2)	BF16	GA, PreAffine	61.59	32.00	18.75	27.40	67.10	44.26	46.75	37.99	49.20	46.42	39.90	42.85
(3)	BF16	GA, GatedNorm	61.71	33.46	19.05	29.90	68.04	46.66	45.56	34.81	51.00	45.76	38.27	43.11
24B-A3B, 500B tokens														
(4)	BF16	GA	67.49	46.02	25.64	31.66	79.45	53.92	58.00	45.88	56.80	55.67	59.27	52.71
(5)	BF16	GA, PreAffine	67.96	48.48	24.34	32.64	82.07	52.62	59.13	46.90	54.80	55.46	58.46	52.99
(6)	BF16	GA, GatedNorm	69.70	47.13	25.81	33.87	82.26	62.70	60.12	47.52	58.40	55.47	59.72	54.79
24B-A3B, 500B tokens, FP4 Quantization														
(7)	W4A4	GA	65.77	45.71	24.73	31.98	76.38	52.66	56.50	44.41	56.60	53.21	51.85	50.89
(8)	W4A4+SQ	GA	66.63	44.86	24.68	32.77	77.45	52.82	55.19	46.83	56.60	53.20	52.55	51.23
(9)	W4A4	GA, PreAffine	66.05	43.61	24.14	31.60	78.70	49.50	56.69	44.01	53.00	51.35	49.58	49.84
(10)	W4A4+SQ	GA, PreAffine	67.17	43.44	23.77	31.66	79.42	49.05	57.56	42.76	55.00	51.98	50.72	50.23
(11)	W4A4	GA, GatedNorm	66.92	46.36	25.59	32.70	81.35	59.40	59.13	46.97	58.80	53.70	56.78	53.43
(12)	W4A4+SQ	GA, GatedNorm	67.35	47.15	25.11	33.62	80.59	61.88	58.06	47.18	58.80	53.39	56.00	53.56

This suggests that the performance gap between GLU and SwiGLU primarily stems from their differing capacities to support outlier-driven rescaling. This also explains why higher-order activation functions that more readily produce outliers, such as ReLU² (Zhang et al., 2024) or PolyNorm (Zhuo et al., 2024), can be advantageous when such rescaling is required. When rescaling is explicitly provided via GatedNorm, the model becomes robust to architectural choices that affect outlier generation.

4 Scaling Outlier Mitigations and Deployment-Level Quantization

In this section, we evaluate different combinations of GatedAttention and GatedNorm, or PreAffine, in large-scale settings. As Tab. 1 row (5) shows, hybrid models exhibit advantages, we conduct experiments on efficient hybrid MoE models following Qwen3-Next, on two settings: (1) a 7.4B-parameter model with 1.7B activated parameters (MoE-7B-A-2B), trained on 1.2T tokens; (2) a 24.6B-parameter model with 2.7B activated parameters (MoE-24B-A3B), trained on 500B tokens. In this setting, GatedNorm incurs less than 3% latency overhead. Full details are provided in App. A.4.1.

Fig 4 shows the training loss curves and outliers for the MoE-24B-A3B model. Our observations include: (1) Baseline’s outliers rise rapidly in early training and gradually decay as the learning rate decreases. The PreAffine model also exhibits an initial outlier surge (within the first 10% of training), but quickly diminishes thereafter. One possible explanation is that the learnable scaler λ_1 initially lacks sufficient magnitude to fully support outlier-driven rescaling, forcing the model to rely temporarily on activation outliers; as training progresses and specific dimensions of λ_1 grow large, outliers are ‘absorbed’ into the parameters. GatedNorm shows no early surge of outliers and maintains consistently low activation magnitudes throughout training, indicating reduced dependence on outlier-driven rescaling. (2) A clear performance gap emerges in the mid-to-late stages, with GatedNorm achieving a lower final loss.

Tab. 2 compares different outlier mitigation strategies across BF16 and quantized settings.

Detailed quantization configs are in App. A.4.2 Since FP8 quantization yields relatively minor changes, we focus on aggressive FP4 W4A4 quantization. Results show consistent gains from PreAffine and GatedNorm across both model sizes, with GatedNorm delivering larger improvements. Notably, GatedNorm achieves +1.0 point gain over the baseline on knowledge tasks and exceeds +2.0 points on STEM and Code tasks.

Under FP4 quantization, we observe: (1) for models using only GA and PreAffine, SmoothQuant (Xiao et al., 2023a) provides an average +0.5 point gain, greater than the gain for GatedNorm, suggesting that

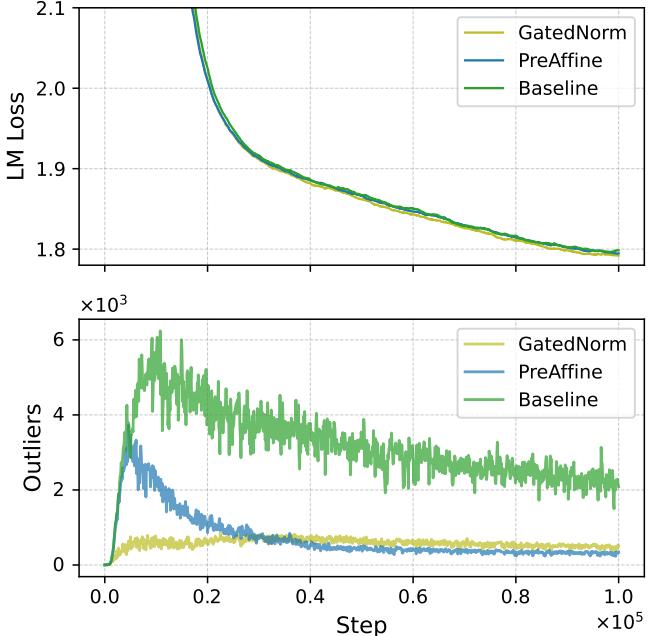


Figure 4: Loss and outliers of MoE-24B-A3B models.

GatedNorm’s rescaling largely mitigates outlier sensitivity; (2) GatedNorm demonstrates the smallest performance drop after FP4 quantization (-1.23 points), compared to GA (-1.50) and PreAffine (-2.76). On the MGSM benchmark, only GatedNorm maintains degradation within 5 points; all other methods incur nearly 10-point losses.

This superior quantization robustness aligns with our observation of a property of gating-based rescaling: it exhibits outlier-suppressing behavior, where dimensions with large $|y|$ tend to receive smaller gating scores y_g , yielding smoother final activations. Overall, PreAffine relocates outliers, similar in spirit to learnable sink and SmoothQuant (Xiao et al., 2023a), but still relies on them to perform outlier-driven rescaling. In contrast, GatedNorm provides explicit gating, producing inherently smoother activations throughout the network and achieving superior quantization robustness.

5 Related Works

Outliers in transformers have been widely studied. In BERT models, outliers in fixed-dimensions are primarily attributed to the weight and bias parameters in LayerNorm (Bondarenko et al., 2021; Kovaleva et al., 2021; Wei et al., 2022), and are closely associated with the attention patterns of special tokens (Puccetti et al., 2022). This phenomenon resembles the attention sink observed in GPT-style models and significantly impacts model performance (Kovaleva et al., 2021; Xiao et al., 2023b). The outliers discussed in BERT largely correspond to MA (Sun et al., 2024; Gu et al., 2024b; Yu et al., 2024) in autoregressive models. These MA typically originate in semantically sparse special tokens, emerge early in FFNs, and propagate through the residual stream, continuously influencing attention distributions in subsequent layers (Sun et al., 2024; Oh et al., 2024; Gu et al., 2024b; Yona et al., 2025).

Several works identify input-independent outliers that consistently appear in fixed dimensions of GPT models (Dettmers et al., 2022). These outliers are not directly related to attention sinks (He et al., 2024; An et al., 2025). He et al. (2024) further attributes these outliers to the normalizations themselves, showing that they persist even when the LayerNorm weights are removed. As outliers hurt both training and inference quantization, a number of approaches aim to mitigate their impact. Common techniques include row-wise, channel-wise, group-wise scaling to limit the quantization error caused by outliers (Yao et al., 2022; Xiao et al., 2023a; Wei et al., 2023; Abecassis et al., 2025), as well as Hadamard transformations to redistribute outlier across dimensions (Xi et al., 2023; Wang et al., 2025).

Another studies focus on reducing outliers during training. From the *optimization* perspective, strategies such as increasing weight decay, gradient clipping (Ahmadian et al., 2023), constraining weight variance (Owen et al., 2025b;a; Xie et al., 2026), or adding explicit regularization loss terms (Liang et al., 2025) can suppress outliers. Some studies also examine whether the Adam optimizer causes the outliers in pre-training (Kaul et al., 2024; He et al., 2024; Xie et al., 2026). From the *architectural* viewpoint, prior work has noted a strong connection between outliers and normalization, and proposed removing normalization to eliminate outliers (He et al., 2024; Owen et al., 2025a;b). Our work shows that, through architectural interventions that explicitly replace outlier-driven rescaling, models can be trained with much smaller activation magnitudes while still using Adam, standard training recipes and normalizations.

Most closely related to our work are studies that investigate the functional role of outliers. Bondarenko et al. (2023); An et al. (2025) propose that attention outliers act as context-aware scaling factors, and demonstrate that introducing gating-based scaling in the attention reduces attention outliers. Karras et al. (2020) identify normalization as the source of outliers in intermediate feature maps of StyleGAN, where these outliers serve to scale signals during normalization. It further shows that input-dependent convolutional weights generated via gating can replicate this scaling effect even without normalization. This gating-based scaling is also adopted in adaLN (Perez et al., 2018; Xu et al., 2019; Peebles & Xie, 2023; Karras et al., 2024). Our work extends this insight to residual sinks in LLMs, highlighting the widespread presence of outlier-driven rescaling across transformers and providing systematic evidence through a series of targeted architectural interventions.

6 Conclusion

This paper argues that outliers in LLMs are not mere artifacts but have functional roles. They work together with normalization mechanisms (softmax and RMSNorm) to perform outlier-driven rescaling, which rescales the magnitude of non-outlier features. This mechanism is essential for stable training and strong performance. Removing outliers and breaking outlier-driven rescaling harms the model. By explicitly providing gating-based rescaling, we can reduce activation outliers while maintaining or even improving performance. Moreover, explicitly enabling rescaling reduces sensitivity to architecture choice. These approaches also yield smoother activations and significantly better quantization robustness, especially under aggressive low-bit settings.

Limitations

This work empirically demonstrates the importance of outlier-driven rescaling in network training and shows that models can leverage normalization, such as RMSNorm, to adjust feature norms. However, we do not investigate why such rescaling is necessary for effective training or representation learning. A deeper theoretical understanding of the role of rescaling remains an open question.

References

- Felix Abecassis, Anjulie Agrusa, Dong Ahn, Jonah Alben, Stefania Alborghetti, Michael Andersch, Sivakumar Arayandi, Alexis Bjorlin, Aaron Blakeman, Evan Briones, et al. Pretraining large language models with nvfp4. *arXiv preprint arXiv:2509.25149*, 2025.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
- Arash Ahmadian, Saurabh Dash, Hongyu Chen, Bharat Venkitesh, Zhen Stephen Gou, Phil Blunsom, Ahmet Üstün, and Sara Hooker. Intriguing properties of quantization at scale. *Advances in Neural Information Processing Systems*, 36:34278–34294, 2023.
- Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*, 2025.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. *arXiv preprint arXiv:2109.12948*, 2021.
- Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36: 75067–75096, 2023.
- Stanislav Budzinskiy, Wenyi Fang, Longbin Zeng, and Philipp Petersen. Numerical error analysis of large language models. *arXiv preprint arXiv:2503.10251*, 2025.
- Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.
- Mingzhi Chen, Taiming Lu, Jiachen Zhu, Mingjie Sun, and Zhuang Liu. Stronger normalization-free transformers. *arXiv preprint arXiv:2512.10938*, 2025.
- 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.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.
- Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, et al. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024.
- Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, 2025.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024a.

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. *arXiv preprint arXiv:2410.10781*, 2024b.

Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. Understanding and minimising outlier features in transformer training. *Advances in Neural Information Processing Systems*, 37:83786–83846, 2024.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unravelling the enigmas of large language models. *arXiv preprint arXiv:2410.17174*, 2024.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. Bert busters: Outlier dimensions that disrupt transformers. *arXiv preprint arXiv:2105.06990*, 2021.

Guang Liang, Jie Shao, Ningyuan Tang, Xinyao Liu, and Jianxin Wu. Tweo: Transformers without extreme outliers enables fp8 training and quantization for dummies. *arXiv preprint arXiv:2511.23225*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Jaehoon Oh, Seungjun Shin, and Dokwan Oh. House of cards: Massive weights in llms. *arXiv preprint arXiv:2410.01866*, 2024.

OpenAI. Multilingual massive multitask language understanding (mmmlu), 2024. Dataset available at Hugging Face.

Louis Owen, Nilabhra Roy Chowdhury, Abhay Kumar, and Fabian Güra. A refined analysis of massive activations in llms. *arXiv preprint arXiv:2503.22329*, 2025a.

Louis Owen, Abhay Kumar, Nilabhra Roy Chowdhury, and Fabian Güra. Variance control via weight rescaling in llm pre-training. *arXiv preprint arXiv:2503.17500*, 2025b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell’Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1286–1304, 2022.

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models, 2025a. URL <https://arxiv.org/abs/2501.11873>.

-
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025b.
- Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, et al. Theory, analysis, and best practices for sigmoid self-attention. *arXiv preprint arXiv:2409.04431*, 2024.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.
- Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Hongyu Wang, Shuming Ma, and Furu Wei. Bitnet v2: Native 4-bit activations with hadamard transformation for 1-bit llms. *arXiv preprint arXiv:2504.18415*, 2025.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.
- 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.
- Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling. *arXiv preprint arXiv:2304.09145*, 2023.
- Haocheng Xi, Changhao Li, Jianfei Chen, and Jun Zhu. Training transformers with 4-bit integers. *Advances in Neural Information Processing Systems*, 36:49146–49168, 2023.
- 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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023a.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023b.
- Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, et al. Controlled llm training on spectral sphere. *arXiv preprint arXiv:2601.08393*, 2026.
- Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.
- 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.
- Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zerquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183, 2022.

Itay Yona, Ilia Shumailov, Jamie Hayes, Federico Barbero, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Zhengyan Zhang, Yixin Song, Guanghui Yu, Xu Han, Yankai Lin, Chaojun Xiao, Chenyang Song, Zhiyuan Liu, Zeyu Mi, and Maosong Sun. Relu2 wins: Discovering efficient activation functions for sparse llms. *arXiv preprint arXiv:2402.03804*, 2024.

Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14901–14911, 2025.

Zhijian Zhuo, Ya Wang, Yutao Zeng, Xiaoqing Li, Xun Zhou, and Jinwen Ma. Polynomial composition activations: Unleashing the dynamics of large language models. *arXiv preprint arXiv:2411.03884*, 2024.

A Appendix

A.1 A Brief Calculation On How Outlier Interacts With the Normalization Layer

We will illustrate below on how residual sinks can rescale the norm of the features after the LayerNorm transformation. Denote the input feature as $\mathbf{h} \in \mathbb{R}^D$. Denote the rescaling parameters of the LN as λ and suppose there is only one outlier dimension d .

As we observe the outlier corresponds to very small affine parameter, assume $|\lambda_d| \leq \epsilon \|\lambda\|_\infty$. Further assume outlier corresponds to r ratio of the norm of the features, that is $r = |\mathbf{h}_d| / \|\mathbf{h}\|_2$. We then have the following inequality:

$$\|\text{LN}(\mathbf{h})\|_{\text{rms}} \leq \|\lambda\|_\infty \sqrt{(1 - r^2) + \epsilon^2 r^2}. \quad (1)$$

This shows that the upper bound on the feature norm after LayerNorm decreases as the outlier becomes larger, allowing the network to rescale the feature norm by changing the magnitude of outliers.

To prove this inequality, let

$$\mathbf{u} := \frac{\mathbf{h}}{\|\mathbf{h}\|_{\text{rms}}}, \quad \|\mathbf{u}\|_{\text{rms}} = 1, \quad |\mathbf{u}_d| = r, \quad \sum_{i \neq d} \mathbf{u}_i^2 = 1 - r^2.$$

Then

$$\text{LN}(\mathbf{h}) = \frac{\lambda \odot \mathbf{h}}{\|\mathbf{h}\|_{\text{rms}}}, \quad (2)$$

$$\|\text{LN}(\mathbf{h})\|_2 = \frac{\|\lambda \odot \mathbf{h}\|_2}{\|\mathbf{h}\|_{\text{rms}}} = \sqrt{D} \frac{\|\lambda \odot \mathbf{h}\|_2}{\|\mathbf{h}\|_2} = \sqrt{D} \|\lambda \odot \mathbf{u}\|_2, \quad (3)$$

and

$$\|\lambda \odot \mathbf{u}\|_2^2 = \lambda_d^2 r^2 + \sum_{i \neq d} \lambda_i^2 \mathbf{u}_i^2. \quad (4)$$

Using $\sum_{i \neq d} \lambda_i^2 \mathbf{u}_i^2 \leq \|\lambda_{-d}\|_\infty^2 \sum_{i \neq d} \mathbf{u}_i^2 = \|\lambda_{-d}\|_\infty^2 (1 - r^2)$, we get

$$\|\text{LN}(\mathbf{h})\|_2 \leq \sqrt{D} \sqrt{\|\lambda_{-d}\|_\infty^2 (1 - r^2) + \lambda_d^2 r^2}. \quad (5)$$

Table 3: Comparison of attention sinks and residual sinks under the outlier-driven rescaling perspective.

Aspect	Attention Sink	Residual Sink
Occurrence	Special tokens (e.g., first token)	Most tokens, fixed dimensions
Associated Normalization	Softmax in attention	RMSNorm
Functional Role	Rescales attention output norm (An et al., 2025)	Rescales RMSNorm output norm
Effect of Removal	Collapse to randomness (Xiao et al., 2023b; Sun et al., 2024)	Leads to performance degradation (Sec. 3.2)
Downstream Suppression	Corresponding value vectors are small (Sun et al., 2024; An et al., 2025)	Corresponding RMSNorm affine weights are small (Sec. 3.3)
Reduction	Diminishes with sigmoid (Ramapuram et al., 2024) or linear attention (Sec. 3.1)	Diminishes with pointwise functions (Sec. 3.1)
Absorption into Parameters	Learnable sink tokens/biases (Sun et al., 2024)	PreAffine (Sec. 3.3)
Explicit Rescaling Alternative	Gated Attention (Bondarenko et al., 2023; An et al., 2025; Qiu et al., 2025b)	GatedNorm (Sec. 3.4)

A.2 Comparison Between Different Outliers

Tab. 3 summarizes the parallels between attention sinks and residual sinks from the perspective of outlier-driven rescaling. Both phenomena arise at normalization layers: softmax for attention sinks and RMSNorm for residual sinks. They serve as modulators that control the scale of non-outlier components. Although they manifest differently—attention sinks are token-specific while residual sinks are dimension-specific—their functional roles are analogous. Both enable rescaling of downstream representations. Critically, both types of outliers are actively suppressed after fulfilling their rescaling function. This is evidenced by small value vector norms for attention sinks and small RMSNorm affine weights for residual sinks. Architectural modifications that remove normalization eliminate these outliers but degrade performance or stability. Conversely, explicitly providing alternative rescaling pathways, such as Gated Attention or GatedNorm, effectively reduces reliance on outliers while preserving or even enhancing model performance.

A.3 Efficiency Analysis of GatedNorm

We evaluate the end-to-end training overhead of our method on an 8-layer dense transformer using the ZeRO-1 optimizer configuration in Megatron-LM (Shoeybi et al., 2019). The hidden dimension is varied while the low-rank dimension of the gating mechanism is fixed to 16. For the PreAffine variant, we implement a custom fused kernel in Triton to ensure efficient execution.

Table 4: End-to-end training overhead of GatedNorm as a function of hidden dimension (rank fixed at 16).

Hidden Size	Relative Overhead
2048	8.1%
4096	5.9%
8192	3.6%

As shown, the relative overhead decreases rapidly with increasing model scale. This trend is explained by two factors. First, the computational cost of GatedNorm scales linearly with the hidden dimension and depends only weakly (via a constant factor) on the low-rank dimension, whereas the dominant GEMM operations in attention and FFN layers scale quadratically with the hidden size. Consequently, the fraction of total compute attributable to GatedNorm becomes increasingly negligible at larger scales.

Second, kernel launch overhead is more significant for smaller hidden dimensions. In such cases, GatedNorm consists of multiple lightweight kernels whose launch latency can create execution bubbles and reduce hardware utilization. As the hidden dimension grows, the per-kernel workload increases sufficiently to amortize launch costs, improving pipeline efficiency and further reducing relative overhead.

Third, in MoE settings, the relative overhead of GatedNorm becomes even smaller. MoE models incur substantial communication and routing costs that dominate the training step time. Since GatedNorm introduces only lightweight, local computation, its contribution to the total step time is further diluted in this regime, resulting in lower relative overhead compared to dense models of similar scale.

A.4 Experimental Setup

A.4.1 Scaling Setups

Model Architecture We evaluate our methods on a suite of large language models with diverse architectures, including both dense and MoE variants. The architectural specifications are summarized in Table 5. All models use a head dimension of 256 for softmax attention and share the same hidden size of 2048. The 2B model is a standard dense Transformer. In contrast, the 7B-A2B and 24B-A3B models are MoE architectures that combine linear and softmax attention in a hybrid configuration: softmax attention is

Table 5: Architectural specifications of the target LLMs used in our experiments. The 7B-A2B and 24B-A3B model are MoEs. All models use a head dimension of 256. Embedding weights are tied in the dense models but not in the MoE model.

Model	2B	7B-A2B	24B-A3B
Layers	28	48	24
Softmax Attention Interval	-	4	4
Softmax Attention Query Heads	8	8	16
Softmax Attention Key / Value Heads	2	1	2
Softmax Attention Head Dimension	256	256	256
Linear Attention Head Dimension	-	128	128
Linear Attention Value Head	-	16	32
Linear Attention Query / Key Head	-	8	16
Tie Embedding	Yes	No	No
Hidden Size	2048	2048	2048
FFN Size	6144	384	512
Number of Experts	-	128	256
Number of Shared Experts	-	1	1
Top- k	-	6	8

applied every 4 layers, while linear attention is used in the remaining layers.

The MoE models employ a large number of experts (128 for 7B-A2B and 256 for 24B-A3B), with top- k routing ($k = 6$ and $k = 8$, respectively) and one shared expert to ensure baseline capacity. They are trained with global load balance loss (Qiu et al., 2025a). Notably, embedding weights are tied in the dense 2B model but untied in the MoE models, following common practice for large sparse architectures. The FFN expansion ratios differ significantly: the dense model uses a wide FFN (6144-dimensional), whereas the MoE models use much smaller per-expert FFNs (384 and 512, respectively), compensated by expert parallelism. Linear attention heads use a reduced dimensionality (128) and separate query/key and value projections, as detailed in the table.

Evaluation We evaluate model performance across a broad set of benchmarks spanning knowledge, reasoning, STEM, code generation, and multilingual capabilities. Specifically, we report results on MMLU-Redux and MMLU-Pro for general knowledge, SuperGPQA and GPQA-Diamond for expert-level scientific reasoning, GSM8K and MATH for mathematical problem solving, CruxEval, MultiPL-E, and MBPP for code generation, and MMMLU and MGSM for multilingual understanding.

MMLU-Redux Gema et al. (2025) is a refined version of the original MMLU benchmark with improved question quality and reduced ambiguity. MMLU-Pro Wang et al. (2024) extends this with more challenging, multi-hop questions requiring deeper reasoning. SuperGPQA Du et al. (2025) and GPQA-Diamond Rein et al. (2024) consist of expert-written, graduate-level scientific questions designed to assess advanced domain knowledge. GSM8K Cobbe et al. (2021) and MATH Hendrycks et al. (2021) evaluate grade-school and advanced mathematical reasoning, respectively. CruxEval Gu et al. (2024a) tests code generation via input-output specification completion, while MultiPL-E Cassano et al. (2023) and MBPP Austin et al. (2021) assess cross-language and beginner-level Python programming. MMMLU OpenAI (2024) and MGSM Shi et al. (2022) are multilingual extensions of MMLU and GSM8K, covering dozens of languages to evaluate cross-lingual transfer.

A.4.2 Quantization Settings

Unified Optimization Strategy. To mitigate activation outliers, we apply *SmoothQuant* (Xiao et al., 2023a) as a universal pre-processing step. A calibration set of 4096 sequences is used solely to compute per-channel smoothing factors, explicitly migrating quantization difficulty from activations to weights.

Quantization Configurations. Building on this smoothed baseline, we evaluate two hardware-aligned formats:

- **FP8 (W8A8):** We employ the E4M3 format. Weights are quantized using a 128×128 *per-block* scaling strategy, while activations utilize dynamic *per-token* quantization.
- **FP4 (W4A4):** We utilize the NVIDIA FP4 (Abecassis et al., 2025) format with *hierarchical two-stage scaling*. Weights are grouped into blocks of 16, where a shared floating-point scale (1st stage) normalizes the range before 4-bit mapping (2nd stage). For activations, we modify the 1st-stage scaling from static to *dynamic per-token* to preserve fidelity.

A.5 More Visualization Results

A.5.1 RMSNorm Visualization

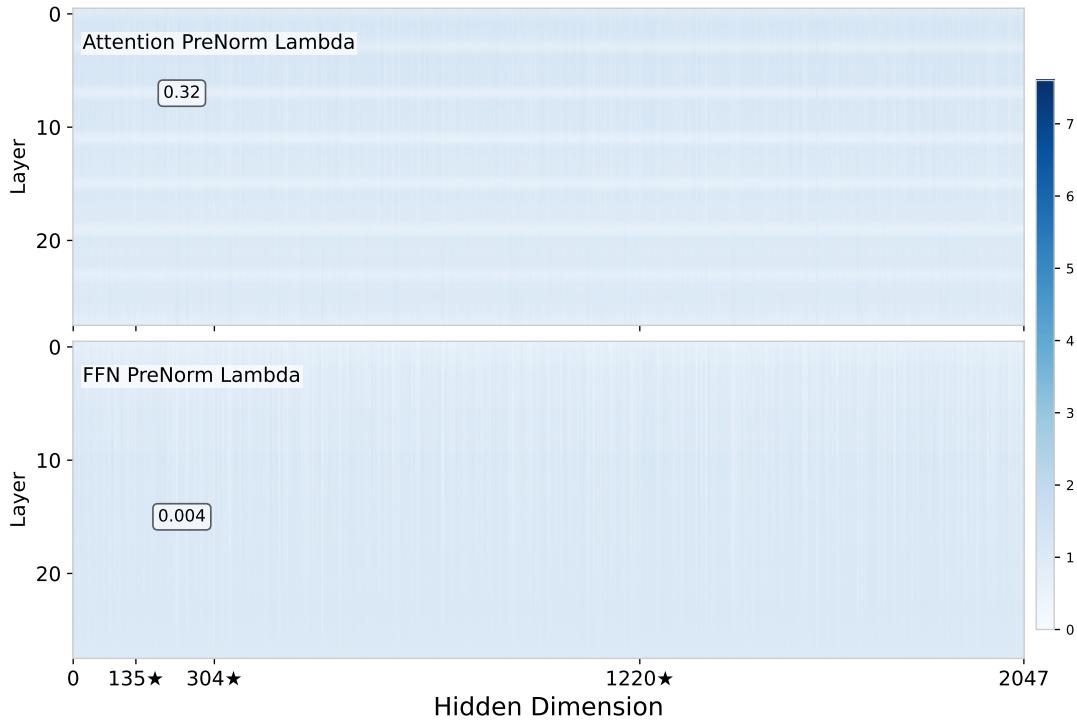


Figure 5: RMSNorm weights for baseline.

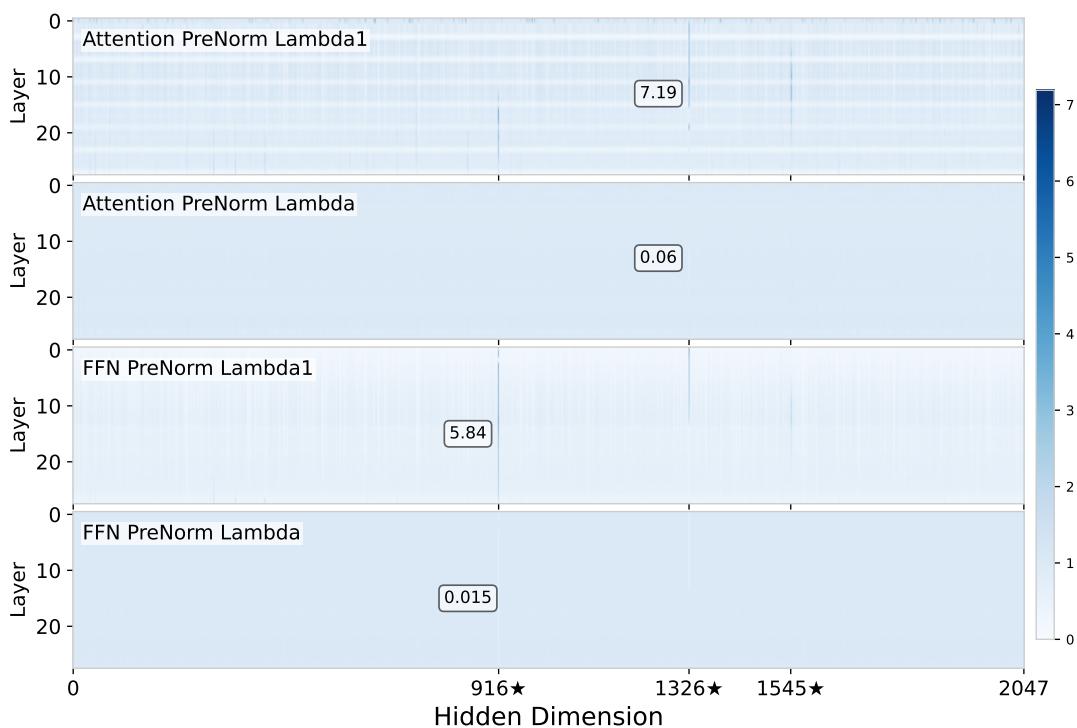


Figure 6: RMSNorm weights for PreAffineRMSNorm.

A.5.2 Outliers for Other Models

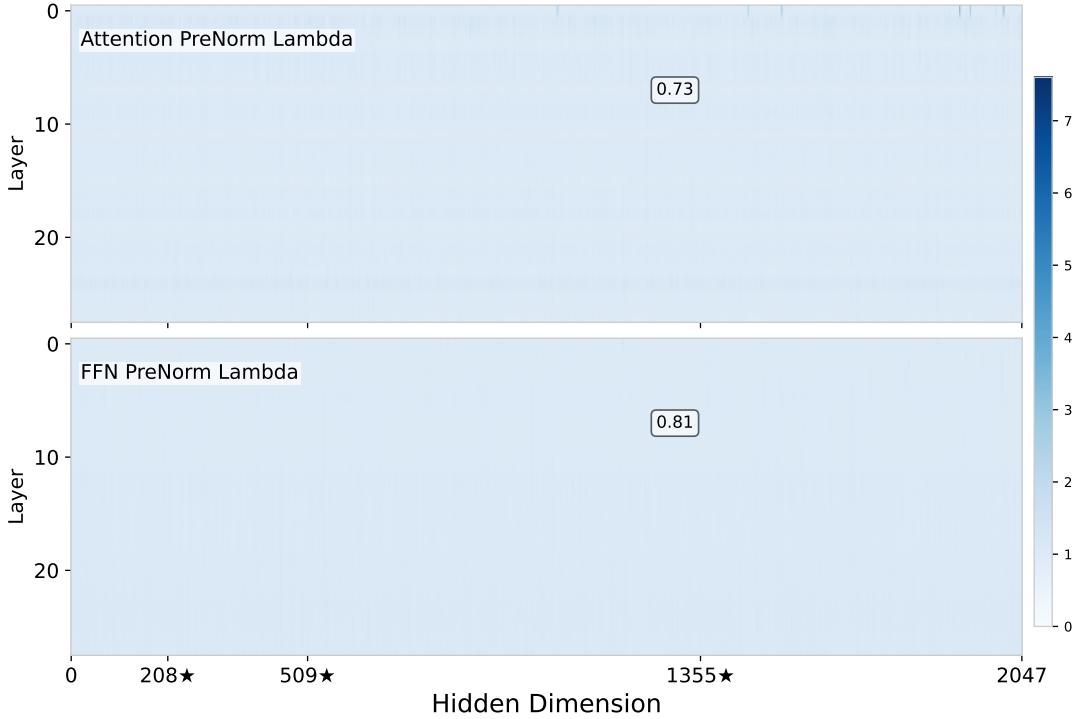


Figure 7: RMSNorm weights for GatedRMSNorm. Weights in most dimensions are near 1, while the largest deviation to 1 is 0.73.

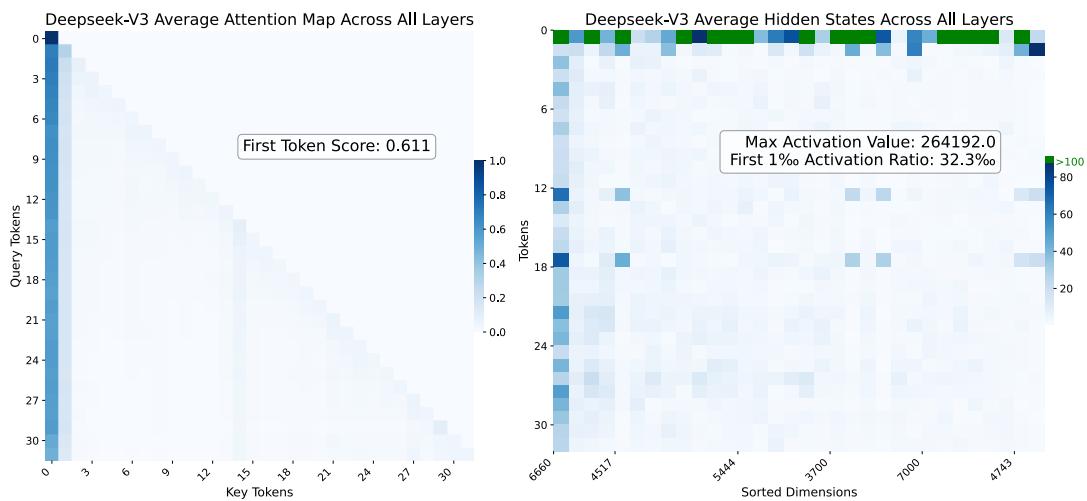


Figure 8: Outliers for Deepseek-V3. The activation pattern of <begin of sentence> token is different from other tokens. Attention sink and residual sink both exists.