

470 A Proofs of the Theorems of curse of depth

471 A.1 Proof of Lemma 1

472 *Proof.* Given Equation (2) from [40], we have:

$$\begin{aligned} y &= x_{\ell+1} = x'_\ell + \text{FFN}(\text{LN}(x'_\ell)), \\ x'_\ell &= x_\ell + \text{Attn}(\text{LN}(x_\ell)). \end{aligned} \quad (13)$$

473 Based on our Assumption 1, let $\text{Var}(\text{Attn}(\text{LN}(x_\ell))) = \sigma_{\text{Attn}}^2$. Then we can write:

$$\begin{aligned} \text{Var}(x'_\ell) &= \text{Var}(x_\ell) + \text{Var}(\text{Attn}(\text{LN}(x_\ell))) + \text{Cov}(\text{Attn}(\text{LN}(x_\ell)), \text{Var}(x_\ell)) \\ &= \sigma_{x_\ell}^2 + \sigma_{\text{Attn}}^2 + \rho_1 \cdot \sigma_{x_\ell} \cdot \sigma_{\text{Attn}}, \end{aligned} \quad (14)$$

474 where ρ_1 is the correlation factor. Similarly, let $\text{Var}(\text{FFN}(\text{LN}(x'_\ell))) = \sigma_{\text{FFN}}^2$. Then we have:

$$\sigma_{x_{\ell+1}}^2 = \sigma_{x'_\ell}^2 + \sigma_{\text{FFN}}^2 + \rho_2 \cdot \sigma_{x'_\ell} \cdot \sigma_{\text{FFN}}, \quad (15)$$

475 where ρ_2 is the correlation factor. Thus, the relationship between $\text{Var}(x_{\ell+1})$ and $\text{Var}(x_\ell)$ becomes:

$$\sigma_{x_{\ell+1}}^2 = \sigma_{x_\ell}^2 + \sigma_{\text{Attn}}^2 + \sigma_{\text{FFN}}^2 + \rho_1 \cdot \sigma_{x_\ell} \cdot \sigma_{\text{Attn}} + \rho_2 \cdot \sigma_{x'_\ell} \cdot \sigma_{\text{FFN}}. \quad (16)$$

476 A.1.1 Variance of the Attention

477 The scaled dot-product attention mechanism is defined as:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \quad (17)$$

478 The softmax function outputs a probability distribution over the keys. Let the softmax output be
479 $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$, where A is a matrix with each row summing to 1. The final attention output
480 is obtained by multiplying the softmax output A with the value matrix V :

$$\text{Attn}(Q, K, V) = AV. \quad (18)$$

481 **Lemma 2** ([24]). *Let $\{X_i\}_{i=1}^N$ be independent and identically distributed random variables with mean*
482 *m and variance $\sigma^2 < \infty$. Define the softmax weights $p_i = \frac{e^{X_i}}{\sum_{j=1}^N e^{X_j}}$, and let $p = (p_1, \dots, p_N)$.*
483 *Then, as $N \rightarrow \infty$, with high probability, the softmax vector p concentrates around the uniform*
484 *distribution on N elements. In particular,*

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\sum_{i=1}^N \left(p_i - \frac{1}{n} \right)^2 \right] = 0, \quad (19)$$

485 *which implies that the softmax output becomes asymptotically indistinguishable, in expectation, from*
486 *the uniform distribution.*

487 According to the above lemma, to simplify the analysis, we make the following additional assumptions:
488 The softmax output A is approximately uniform, meaning each element of A is roughly $1/n$, where
489 n is the number of keys/values. Given this assumption, the variance of the attention is:

$$\text{Var}(\text{Attn}(Q, K, V)) \sim \text{Var}(AV) = \frac{1}{n} \sum_{i=1}^n d_{\text{head}} \text{Var}(V_i) = \frac{1}{n} \cdot n \sigma_V^2 \cdot d_{\text{head}} = d_{\text{head}} \sigma_V^2 = \sigma_W^2 d. \quad (20)$$

490 where W is the universal weight matrix defined as before.

491 A.1.2 Variance of the Feed-Forward Network

492 The feed-forward network (FFN) in transformers typically consists of two linear transformations with
 493 a ReLU activation in between. The FFN can be written as:

$$\text{FFN}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2. \quad (21)$$

494 where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors.

495 Using the result obtained by Wang et al. [46], we get:

$$\sigma_{\text{FFN}}^2 \sim \sigma_{W_1}^2 \cdot \sigma_{W_2}^2 = \sigma_W^4. \quad (22)$$

496 In conclusion:

$$\begin{aligned} \sigma_{x'_\ell}^2 &= \sigma_{x_\ell}^2 + \sigma_W^2 + \rho_2 \cdot \sigma_{x_\ell} \cdot \sigma_W \\ &= \sigma_{x_\ell}^2 \left(1 + \frac{\sigma_W}{\sigma_{x_\ell}} + \frac{\sigma_W^2}{\sigma_{x_\ell}^2}\right) \\ &= \sigma_{x_\ell}^2 \Theta\left(1 + \frac{1}{\sigma_{x_\ell}}\right). \end{aligned} \quad (23)$$

497 For simplicity, we set the numerator part to 1. Substitute $\sigma_{x'_\ell} = \sigma_{x_\ell} \sqrt{1 + \frac{\sigma_W^2}{\sigma_{x_\ell}^2} + \rho_2 \cdot \frac{\sigma_W}{\sigma_{x_\ell}}}$. into
 498 Equation (16) we get:

$$\begin{aligned} \sigma_{x_{\ell+1}}^2 &= \sigma_{x_\ell}^2 + \sigma_W^2 + \sigma_W^4 d^2 + \rho_1 \cdot \sigma_{x_\ell} \cdot \sigma_W + \rho_2 \cdot \sigma_{x'_\ell} \cdot \sigma_W^2 d \\ &= \sigma_{x_\ell}^2 + \sigma_W^2 + \sigma_W^4 d^2 + \rho_1 \cdot \sigma_{x_\ell} \cdot \sigma_W + \rho_2 \cdot \sigma_{x_\ell} \cdot \sigma_W^2 d + \frac{\rho_2 \sigma_W^4 d^2}{2\sigma_{x_\ell}} + \frac{\rho_2^2 \sigma_W^3 d \sigma_{x_\ell}}{2} \\ &= \sigma_{x_\ell}^2 \Theta\left(1 + \frac{1}{\sigma_{x_\ell}}\right). \end{aligned} \quad (24)$$

499 From the result we can generally infer that the variance accumulates layer by layer. The variance
 500 with regard to σ_{x_1} :

$$\sigma_{x_\ell}^2 = \sigma_{x_1}^2 \Theta\left(\prod_{k=1}^{\ell-1} \left(1 + \frac{1}{\sigma_{x_k}}\right)\right). \quad (25)$$

501 We can also obtain a similar result for $\sigma_{x'_\ell}^2$.

502 We observe that for any $\sigma_{x_k}^2 \geq 1$, the sequence is increasing, meaning each term in the product is
 503 bounded. Consequently, the entire product is bounded above by:

$$\sigma_{x_\ell}^2 \leq \sigma_{x_1}^2 \prod_{k=1}^{\ell-1} \left(1 + \sqrt{\frac{1}{\sigma_{x_1}^2}}\right) = \sigma_{x_1}^2 \left(1 + \sqrt{\frac{1}{\sigma_{x_1}^2}}\right)^{\ell-1} = \exp \Theta(L). \quad (26)$$

504 Taking the natural logarithm of both sides:

$$\begin{aligned} \log(\sigma_{x_\ell}^2) &= \log\left(\sigma_{x_1}^2 \prod_{k=1}^{\ell-1} \left(1 + \sqrt{\frac{1}{\sigma_{x_k}^2}}\right)\right) = \sum_{k=1}^{\ell-1} \log\left(1 + \sqrt{\frac{1}{\sigma_{x_k}^2}}\right) + \log(\sigma_{x_1}^2) \\ &\geq \sum_{k=1}^{\ell-1} \left(\sqrt{\frac{1}{\sigma_{x_k}^2}} - \frac{1}{2} \left(\sqrt{\frac{1}{\sigma_{x_k}^2}}\right)^2\right) + \log(\sigma_{x_1}^2). \end{aligned} \quad (27)$$

Exponentiating both sides to find the lower bound for $\sigma_{x_\ell}^2$, we obtain:

$$\sigma_{x_\ell}^2 \geq \sigma_{x_1}^2 \exp\left(\sum_{k=1}^{\ell-1} \left(\sqrt{\frac{1}{\sigma_{x_k}^2}} - \frac{1}{2\sigma_{x_k}^2}\right)\right).$$

505 This provides a tighter lower bound for $\sigma_{x_\ell}^2$ compared to the upper bound of Equation (26). Since we
 506 know the upper bound of variance grows exponentially, the lower bound must be sub-exponential.
 507 Therefore, for $\sigma_{x_\ell}^2 = \ell$, we must have:

$$\sigma_{x_\ell}^2 \geq \sigma_{x_1}^2 \exp \left(\sum_{k=1}^{\ell-1} \left(\frac{1}{k} - \frac{1}{2k} \right) \right) = \Theta(\exp(\sqrt{L})) \geq \Theta(L).$$

508 □

509 Therefore, the increasing lower bound for $\sigma_{x_\ell}^2$ must grow faster than a linear function. So, the increase
 510 of variance is sub-exponential. A large increase in such bound will lead to gradient spikes, which can
 511 connect to previous studies in Huang et al. [19, 20].

512 A.2 Proof of Theorem 1

513 In this proof, we will divide the argument into two parts: first, the calculation of the Lemma 3, and
 514 second, the analysis of $\frac{\partial y_\ell}{\partial x_1}$.

515 **Lemma 3.** For an L -layered Pre-LN Transformer, $\frac{\partial y_L}{\partial x_1}$ using Equations (2) and (3) is given by:

$$\frac{\partial y_L}{\partial x_1} = \prod_{n=1}^{L-1} \left(\frac{\partial y_\ell}{\partial x'_\ell} \cdot \frac{\partial x'_\ell}{\partial x_\ell} \right). \quad (28)$$

516 The upper bound for the norm of $\frac{\partial y_L}{\partial x_1}$ is:

$$\begin{aligned} \left\| \frac{\partial y_L}{\partial x_1} \right\|_2 &\leq \prod_{l=1}^{L-1} \left(\left(1 + \frac{\sigma^2}{\sigma_{x'_\ell} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2} \right) \right. \\ &\quad \times \left. \left(1 + 2dh \left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \frac{\sigma^2}{\sigma_{x_\ell}} \left(\sigma^2 d \sqrt{d_{\text{head}}} + \left(1 + \sqrt{d_{\text{head}}/d} \right) \right) \right) \right). \end{aligned} \quad (29)$$

517

518 Here, h denotes the number of heads, s is the sequence length, and d , d_{FFN} , and d_{head} are the
 519 dimension of the embedding, FFN layer and multi-head attention layer, respectively. The standard
 520 deviation of W_Q , W_K , W_V , and W_{FFN} at layer ℓ is σ based on Assumption 1.

521 A.2.1 Proof of Lemma 3

522 *Proof.* Our derivation follows results in [40], specifically Equation (7), which provides an upper
 523 bound on the norm of $\frac{\partial y_\ell}{\partial x_1}$ as:

$$\left\| \frac{\partial y_\ell}{\partial x_1} \right\|_2 = \left\| \prod_{l=1}^{L-1} \frac{\partial y_\ell}{\partial x'_\ell} \frac{\partial x'_\ell}{\partial x_\ell} \right\|_2. \quad (30)$$

524 Thus, we can estimate the upper bound of the gradient norm of $\frac{\partial y_\ell}{\partial x_1}$ by analyzing the spectral norms
 525 of the Jacobian matrices for the FFN layer and the self-attention layer, namely,

$$\text{FFN: } \left\| \frac{\partial y_\ell}{\partial x'_\ell} \right\|_2 \quad \text{Attention: } \left\| \frac{\partial x'_\ell}{\partial x_\ell} \right\|_2. \quad (31)$$

526 We now derive an upper bound of $\left\| \frac{\partial y_\ell}{\partial x'_\ell} \right\|_2$ as follows:

$$\left\| \frac{\partial y_\ell}{\partial x'_\ell} \right\|_2 \leq 1 + \left\| \frac{\partial \text{FFN}(\text{LN}(x'_\ell))}{\partial \text{LN}(x'_\ell)} \right\|_2 \left\| \frac{\partial \text{LN}(x'_\ell)}{\partial x'_\ell} \right\|_2. \quad (32)$$

Let σ_{w1_ℓ} and σ_{w2_ℓ} be the standard deviations of W_ℓ^1 and W_ℓ^2 , respectively. From Assumption 1, the
 spectral norms of W_ℓ^1 and W_ℓ^2 are given by their standard deviations and dimensions [45], so we
 have:

$$\|W_1\|_2 \sim \sigma_1 \sqrt{d + \sqrt{d_{\text{FFN}}}}.$$

527 .

528 For simplicity, we assume that d , and d_{FFN} are equal, thus,

$$\left\| \frac{\partial \text{FFN}(\text{LN}(x'_\ell))}{\partial \text{LN}(x'_\ell)} \right\|_2 = \|W_\ell^1 W_\ell^2\|_2 \leq \sigma_1 \sigma_2 (\sqrt{d} + \sqrt{d_{\text{ffn}}})^2. \quad (33)$$

529 Finally, we have the following bound:

$$\left\| \frac{\partial y_\ell}{\partial x'_\ell} \right\|_2 \leq 1 + \frac{\sigma_{w1_\ell} \sigma_{w2_\ell}}{\sigma_{x'_\ell} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2} = 1 + \frac{\sigma_\ell^2}{\sigma_{x'_\ell} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2}. \quad (34)$$

530 Following a similar procedure for the FFN, we rewrite $\|\frac{\partial x'}{\partial x}\|_2$ in Equation (31) as:

$$\left\| \frac{\partial x'}{\partial x} \right\|_2 \leq 1 + \left\| \frac{\partial \text{Attn}(\text{LN}(x))}{\partial \text{LN}(x)} \right\|_2 \left\| \frac{\partial \text{LN}(x)}{\partial x} \right\|_2. \quad (35)$$

531 Let $Z(\cdot) = \text{concat}(\text{head}_1(\cdot), \dots, \text{head}_h(\cdot))$ and J^Z denote the Jacobian of the $Z(\cdot)$. We can now
532 express the spectral norm of the Jacobian matrix of attn as:

$$\left\| \frac{\partial \text{Attn}(\text{LN}(x_\ell))}{\partial \text{LN}(x_\ell)} \right\|_2 = \left\| W_\ell^O Z(\text{LN}(x_\ell)) \frac{\partial Z(\text{LN}(x_\ell))}{\partial \text{LN}(x_\ell)} \right\|_2 = \|W_\ell^O J_\ell^Z\|_2. \quad (36)$$

533 From [45], we know that:

$$\|J_\ell^Z\|_2 \leq h \left(\left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \sigma^3 \sqrt{d^3 d_{\text{head}}} + \sigma_x^\ell \left(\sqrt{d} + \sqrt{d_{\text{head}}} \right) \right). \quad (37)$$

534 Here h is the number of heads, s is the sequence length, and the standard deviation of W_Q , W_K , and
535 W_V is σ .

536 By combining the inequalities (34), (37) and (35), and assuming that all σ values are the same for
537 simplicity, we obtain:

$$\begin{aligned} \left\| \frac{\partial y_L}{\partial x_1} \right\|_2 &\leq \prod_{l=1}^{L-1} \left(\left(1 + \frac{\sigma^2}{\sigma_{x'_\ell} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2} \right) \right. \\ &\quad \times \left. \left(1 + 2dh \left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \frac{\sigma^2}{\sigma_{x_\ell}} \left(\sigma^2 d \sqrt{d_{\text{head}}} + \left(1 + \sqrt{d_{\text{head}}/d} \right) \right) \right) \right). \end{aligned} \quad (38)$$

538 \square

539 A.2.2 Analysis of the Upper Bound

540 As discussed in [40], σ should be sufficiently small, and the standard deviation, $\sigma_{x'_\ell}$ or σ_{x_ℓ} should
541 satisfy the condition $\sigma^2 \ll \sigma_{x'_\ell}$ to maintain the lazy training scheme. Thus, we obtain the following
542 bound for the product over ℓ from 1 to L :

543 To find the bound for $\left\| \frac{\partial y_\ell}{\partial x_1} \right\|_2$ with respect to ℓ , we simplify the given inequality by approximating
544 σ_{x_ℓ} and $\sigma_{x'_\ell}$. Based on Equation (23), σ_{x_ℓ} is only one layer ahead of $\sigma_{x'_\ell}$, and this layer does not
545 significantly affect the overall performance of deep Transformer networks. Furthermore, based on
546 Lemma 1, we assume that $\sigma_{x'_\ell} = \sigma_{x_\ell}$.

547 Equation (3) can be expressed in a traditional product form [48] for σ_{x_ℓ} :

$$\left\| \frac{\partial y_L}{\partial x_1} \right\|_2 \leq \prod_{l=1}^{L-1} \left(1 + \frac{1}{\sigma_{x_\ell}} A + \frac{1}{\sigma_{x_\ell}^2} B \right), \quad (39)$$

548 where

$$A = \frac{\sigma^2}{(\sqrt{d} + \sqrt{d_{\text{FFN}}})^2} + 2dh \left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \sigma^2 \left(d\sqrt{d_{\text{head}}} + 1 + \sqrt{d_{\text{head}}/d} \right), \quad (40)$$

549 and

$$B = 2dh \left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \sigma^4 d \sqrt{d_{\text{head}}}, \quad (41)$$

550 where A and B are independent of σ_{x_ℓ} , and under our assumption, are treated as constants.

551 From classical infinite series analysis, it is known that as σ_{x_ℓ} grows at a faster rate, the upper bound of
 552 the product decreases. The proof is omitted here for brevity. For the upper bound on the convergence
 553 rate of $\sigma_{x_\ell}^2$, we assume $\sigma_{x_\ell}^2 = \exp(\ell)$ without loss of generality. Under this condition, we can derive
 554 the following result:

555 Taking the natural logarithm of the product:

$$\log \left(\prod_{k=1}^{L-1} \left(1 + \frac{A}{e^k} + \frac{B}{e^{2k}} \right) \right) = \sum_{k=1}^{L-1} \log \left(1 + \frac{A}{e^k} + \frac{B}{e^{2k}} \right).$$

Using the Taylor series expansion for $\log(1+x)$, and applying this to our sum, we get:

$$\sum_{k=1}^{\infty} \log \left(1 + \frac{A}{e^k} + \frac{B}{e^{2k}} \right) = \sum_{k=1}^{\infty} \left(\frac{A}{e^k} + \frac{B}{e^{2k}} - \frac{1}{2} \left(\frac{A}{e^k} + \frac{B}{e^{2k}} \right)^2 + \frac{1}{3} \left(\frac{A}{e^k} + \frac{B}{e^{2k}} \right)^3 - \dots \right).$$

556 By evaluating the sums for each order of terms, we find that the result is a constant. Carrying this out
 557 for each term, we obtain:

$$\log \left(\prod_{k=1}^{L-1} \left(1 + \frac{A}{e^k} + \frac{B}{e^{2k}} \right) \right) \sim \frac{A}{e-1} + \frac{B}{e^2-1} - \frac{1}{2} \left(\frac{A^2}{e^2-1} + 2 \frac{A \cdot B}{e^3-1} + \frac{B^2}{e^4-1} \right).$$

558 Thus, the product is approximately:

$$\left\| \frac{\partial y_L}{\partial x_1} \right\|_2 \leq \exp \left(\frac{A}{e-1} + \frac{B}{e^2-1} - \frac{1}{2} \left(\frac{A^2}{e^2-1} + 2 \frac{A \cdot B}{e^3-1} + \frac{B^2}{e^4-1} \right) \right) = M, \quad (42)$$

559 where M is a constant.

560 For the lower bound on the convergence rate of $\sigma_{x_\ell}^2$, we assume $\sigma_{x_\ell}^2 = \ell$ without loss of generality.
 561 Under this condition, we derive the following result. Taking the logarithm of the product, applying
 562 the Taylor series expansion for $\log(1+x)$, and applying this to our sum:

$$\sum_{k=1}^{\infty} \log \left(1 + \frac{A}{k} + \frac{B}{e^{k^2}} \right) = \sum_{k=1}^{\infty} \left(\frac{A}{k} + \frac{B}{e^{k^2}} - \frac{1}{2} \left(\frac{A}{k} + \frac{B}{e^{k^2}} \right)^2 + \frac{1}{3} \left(\frac{A}{k} + \frac{B}{e^{k^2}} \right)^3 - \dots \right).$$

563 For the first-order terms:

$$\sum_{k=1}^{\infty} \left(\frac{A}{k} + \frac{B}{e^{k^2}} \right) = A \sum_{k=1}^{\infty} \frac{1}{k} + B \sum_{k=1}^{\infty} \frac{1}{e^{k^2}}.$$

564 The series $\sum_{k=1}^{\infty} \frac{1}{k}$ is the harmonic series, which diverges. However, we approximate it using the
 565 Euler-Mascheroni constant γ and the fact recognize that the harmonic series grows logarithmically:

$$\sum_{k=1}^{\infty} \frac{1}{k} \sim \log n + \gamma \quad (\text{for large } n).$$

566 The other series such as $\sum_{k=1}^{\infty} \frac{1}{e^{k^2}}$ converge because e^{k^2} grows very rapidly.

567 For higher-order terms, they converge to constant, involving the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$ converges to $\frac{\pi^2}{6}$, so
 568 they contribute a constant. Exponentiating both sides, we get:

$$\prod_{k=1}^{\infty} \left(1 + \frac{A}{k} + \frac{B}{e^{k^2}}\right) \sim \exp(A(\log n + \gamma) + \text{const}).$$

569 Thus, the growth rate of the upper bound for $\left\|\frac{\partial y_L}{\partial x_1}\right\|_2$ is:

$$\left\|\frac{\partial y_L}{\partial x_1}\right\|_2 \leq \Theta(L). \quad (43)$$

570 B Theoretical Analysis of LayerNorm Scaling

571 **Lemma 4.** *After applying our scaling method, the variances of x'_ℓ and x_ℓ , denoted as $\sigma_{x'_\ell}^2$ and $\sigma_{x_\ell}^2$,
 572 respectively, exhibit the same growth trend, which is:*

$$\sigma_{x_\ell}^2 = \sigma_{x_1}^2 \Theta\left(\prod_{k=1}^{\ell-1} \left(1 + \frac{1}{\sqrt{k}\sigma_{x_k}}\right)\right), \quad (44)$$

573 with the following growth rate bounds:

$$\Theta(L) \leq \sigma_{x_L}^2 \leq \Theta(L^{(2-\epsilon)}). \quad (45)$$

574 where ϵ is a small number with $1/2 \leq \epsilon < 1$.

575 From Lemma 4, we can conclude that our scaling method effectively slows the growth of the variance
 576 upper bound, reducing it from exponential to polynomial growth. Specifically, it limits the upper
 577 bound to a quadratic rate instead of an exponential one. Based on Theorem 1, after scaling, we obtain
 578 the following:

579 **Theorem 2.** *For the scaled Pre-LN Transformers, the Euclidean norm of $\frac{\partial y_L}{\partial x_1}$ is given by:*

$$\left\|\frac{\partial y_L}{\partial x_1}\right\|_2 \leq \prod_{\ell=1}^{L-1} \left(1 + \frac{1}{\ell\sigma_{x_\ell}}A + \frac{1}{\ell^2\sigma_{x_\ell}^2}B\right), \quad (46)$$

580 where A and B are dependent on the scaled neural network parameters. Then the upper bound for
 581 the norm is given as follows: when $\sigma_{x_\ell}^2$ grows at $\ell^{(2-\epsilon)}$, (i.e., at its upper bound), we obtain:

$$\sigma_{x_\ell}^2 \sim \ell^{(2-\epsilon)}, \quad \left\|\frac{\partial y_L}{\partial x_1}\right\|_2 \leq \omega(1), \quad (47)$$

582 where ω denotes that if $f(x) = \omega(g(x))$, then $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \infty$. Meanwhile, when $\sigma_{x_\ell}^2$ grows
 583 linearly (i.e., at its lower bound), we obtain:

$$\sigma_{x_\ell}^2 \sim \ell, \quad \left\|\frac{\partial y_L}{\partial x_1}\right\|_2 \leq \Theta(L). \quad (48)$$

584

585 The detailed descriptions of A and B , and ϵ , along with the full proof, are provided in Appendices
 586 B.1 and B.2.

587 By comparing Theorem 1 (before scaling) with Theorem 2 (after scaling), we observe a substan-
 588 tial reduction in the upper bound of variance. Specifically, it decreases from exponential growth
 589 $\Theta(\exp(L))$ to at most quadratic growth $\Theta(L^2)$. In fact, this growth is even slower than quadratic
 590 expansion, as it follows $\Theta(L^{(2-\epsilon)})$ for some small $\epsilon > 0$.

When we select a reasonable upper bound for this expansion, we find that $\left\| \frac{\partial y_L}{\partial x_1} \right\|_2$ no longer possesses a strict upper bound. That is, as the depth increases, $\left\| \frac{\partial y_L}{\partial x_1} \right\|_2$ continues to grow gradually. Consequently, fewer layers act as identity mappings compared to the original Pre-LN where nearly all deep layers collapsed into identity transformations. Instead, the after-scaled network effectively utilizes more layers, even as the depth approaches infinity, leading to improved expressivity and trainability.

B.1 Proof of Lemma 4

Proof. After scaling, the equation becomes:

$$\begin{aligned} y &= x_{\ell+1} = x'_\ell + \text{FFN}\left(\frac{1}{\sqrt{\ell}} \text{LN}(x'_\ell)\right), \\ x'_\ell &= x_\ell + \text{Attn}\left(\frac{1}{\sqrt{\ell}} \text{LN}(x_\ell)\right). \end{aligned} \quad (49)$$

Following the same analysis as before, we scale the Attention and FFN sub-layers, yielding:

$$\sigma_{\text{Attn}}^2 = \frac{1}{n\ell} \cdot n \cdot \sigma_V^2 = \frac{1}{\ell} \sigma_V^2 = \frac{\sigma_W^2}{\ell}, \quad \sigma_{\text{FFN}}^2 \sim \frac{\sigma_{W_1}^2}{\ell} \cdot \frac{\sigma_{W_2}^2}{\ell} = \frac{\sigma_W^4}{\ell^2}. \quad (50)$$

In conclusion:

$$\sigma_{x'_\ell}^2 = \sigma_{x_\ell}^2 + \sigma_W^2 + \rho_2 \cdot \sigma_{x_\ell} \cdot \frac{\sigma_W}{\sqrt{\ell}} = \sigma_{x_\ell}^2 \Theta\left(1 + \frac{1}{\sqrt{\ell} \sigma_{x_\ell}}\right). \quad (51)$$

Similarly, we obtain:

$$\sigma_{x_{\ell+1}}^2 = \sigma_{x'_\ell}^2 \Theta\left(1 + \frac{1}{\sqrt{\ell} \sigma_{x'_\ell}}\right). \quad (52)$$

From the result we can generally infer that the variance accumulates layer by layer. The variance with regard to σ_{x_1} :

$$\sigma_{x_\ell}^2 = \sigma_{x_1}^2 \Theta\left(\prod_{k=1}^{\ell-1} \left(1 + \frac{1}{\sqrt{k} \sigma_{x_k}}\right)\right), \quad (53)$$

We can also obtain a similar result for $\sigma_{x'_\ell}^2$.

Taking the natural logarithm of both sides:

$$\begin{aligned} \log(\sigma_{x_\ell}^2) &= \log\left(\sigma_{x_1}^2 \prod_{k=1}^{\ell-1} \left(1 + \sqrt{\frac{1}{k \sigma_{x_k}^2}}\right)\right) = \sum_{k=1}^{\ell-1} \log\left(1 + \sqrt{\frac{1}{k \sigma_{x_k}^2}}\right) + \log(\sigma_{x_1}^2) \\ &\geq \sum_{k=1}^{\ell-1} \left(\sqrt{\frac{1}{k \sigma_{x_k}^2}} - \frac{1}{2} \left(\sqrt{\frac{1}{k \sigma_{x_k}^2}}\right)^2\right) + \log(\sigma_{x_1}^2). \end{aligned} \quad (54)$$

To establish a lower bound for $\sigma_{x_\ell}^2$, we exponentiate both sides. Setting $\sigma_{x_\ell}^2 = \ell$, we must have:

$$\sigma_{x_\ell}^2 \geq \sigma_{x_1}^2 \exp\left(\sum_{k=1}^{\ell-1} \left(\frac{1}{k} - \frac{1}{2k}\right)\right) = \Theta(\exp(\log L)) \geq \Theta(L). \quad (55)$$

Therefore, the increasing lower bound $\sigma_{x_\ell}^2$ is greater than a linear function.

Similarly, assuming $\sigma_{x_\ell}^2 = \ell^{(2-\epsilon)}$, we have:

$$\begin{aligned} \sigma_{x_\ell}^2 &= \sigma_{x_1}^2 \prod_{k=1}^{\ell-1} \left(1 + \frac{1}{k^{2-\epsilon/2}}\right) \sim \exp\left(\sum_{k=1}^{\ell-1} \frac{1}{k^{2-\epsilon/2}}\right) \sim \exp\left(\frac{\ell^{\epsilon/2-1} - 1}{\epsilon/2 - 1}\right) \\ &\leq \Theta(\ell^{(2-\epsilon)}) \leq \Theta(\ell^2). \end{aligned} \quad (56)$$

Here ϵ is a small constant with $1/2 \leq \epsilon < 1$. Therefore, the increasing upper bound of $\sigma_{x_\ell}^2$ is slower than the ℓ^3 function, leading to:

$$\sigma_{x_\ell}^2 \leq \Theta(L^2)$$

609 .

610

□

611 B.2 Proof of Theorem 2

612 *Proof.* Similarly, after applying the scaling transformation, we derive an upper bound for $\|\frac{\partial y_\ell}{\partial x_\ell}\|_2$ as
613 follows:

$$\begin{aligned} \left\| \frac{\partial y_\ell}{\partial x_\ell} \right\|_2 &\leq 1 + \left\| \frac{\partial \text{FFN}(\text{LN}(x'_\ell))}{\partial \text{LN}(x'_\ell)} \right\|_2 \left\| \frac{1}{\sqrt{\ell}} \right\|_2 \left\| \frac{\partial \text{LN}(x'_\ell)}{\partial x'_\ell} \right\|_2 \\ &= 1 + \frac{\sigma_\ell^2}{\ell \sigma_{x'_\ell} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2}. \end{aligned} \quad (57)$$

614 Similarly, rewriting Equation (31) after scaling, we have

$$\left\| \frac{\partial x'}{\partial x} \right\|_2 \leq 1 + \left\| \frac{\partial \text{Attn}(\text{LN}(x))}{\partial \text{LN}(x)} \right\|_2 \left\| \frac{1}{\sqrt{\ell}} \right\|_2 \left\| \frac{\partial \text{LN}(x)}{\partial x} \right\|_2. \quad (58)$$

615 By combining the bound (57), and inequality (58), and assuming all σ are equal for simplicity, we
616 obtain:

$$\begin{aligned} \left\| \frac{\partial y_L}{\partial x_1} \right\|_2 &\leq \prod_{l=1}^{L-1} \left(\left(1 + \frac{\sigma^2}{\ell \sigma_{x'_l} (\sqrt{d} + \sqrt{d_{\text{FFN}}})^2} \right) \right. \\ &\quad \times \left. \left(1 + 2dh \left(\sqrt{s} + 2 + \frac{1}{\sqrt{s}} \right) \frac{\sigma^2}{\ell \sigma_{x_l}} \left(\sigma^2 d \sqrt{d_{\text{head}}} + \left(1 + \sqrt{d_{\text{head}}/d} \right) \right) \right) \right). \end{aligned} \quad (59)$$

617 Equation (59) is a traditional product form [48] for σ_{x_ℓ} . After scaling, it becomes:

$$\left\| \frac{\partial y_L}{\partial x_1} \right\|_2 \leq \prod_{l=1}^{L-1} \left(1 + \frac{1}{\ell \sigma_{x_\ell}} A + \frac{1}{\ell^2 \sigma_{x_\ell}^2} B \right), \quad (60)$$

618 where A and B retain their forms from Equation (40) and Equation (41) and are treated as constants.

619 Regarding the upper bound on the convergence rate of $\sigma_{x_\ell}^2$, we assume $\sigma_{x_\ell}^2 = \ell^{(2-\epsilon)}$ without loss of
620 generality. For large L , the product can be approximated using the properties of infinite products:

$$\prod_{\ell=1}^{L-1} \left(1 + \frac{A}{\ell^{2-\epsilon/2}} + \frac{B}{\ell^{4-\epsilon}} \right) \sim \exp \left(\sum_{\ell=1}^{L-1} \left(\frac{A}{\ell^{2-\epsilon/2}} + \frac{B}{\ell^{4-\epsilon}} \right) \right). \quad (61)$$

621 Then, by evaluating the sum in the exponent, we obtain:

$$\prod_{\ell=1}^{L-1} \left(1 + \frac{A}{\ell^{2-\epsilon/2}} + \frac{B}{\ell^{4-\epsilon}} \right) \sim \exp \left(A \cdot \frac{\ell^{\epsilon/2-1} - 1}{\epsilon/2 - 1} + B \cdot \frac{\ell^{\epsilon-3} - 1}{\epsilon - 3} \right). \quad (62)$$

622 Therefore, we establish the upper bound:

$$\left\| \frac{\partial y_L}{\partial x_1} \right\|_2 \leq \Theta \left(\exp \left(A \cdot \frac{\ell^{\epsilon/2-1} - 1}{\epsilon/2 - 1} + B \cdot \frac{\ell^{\epsilon-3} - 1}{\epsilon - 3} \right) \right) = \omega(1), \quad (63)$$

623 where $\omega(1)$ denotes a growth strictly greater than a constant as defined before. □

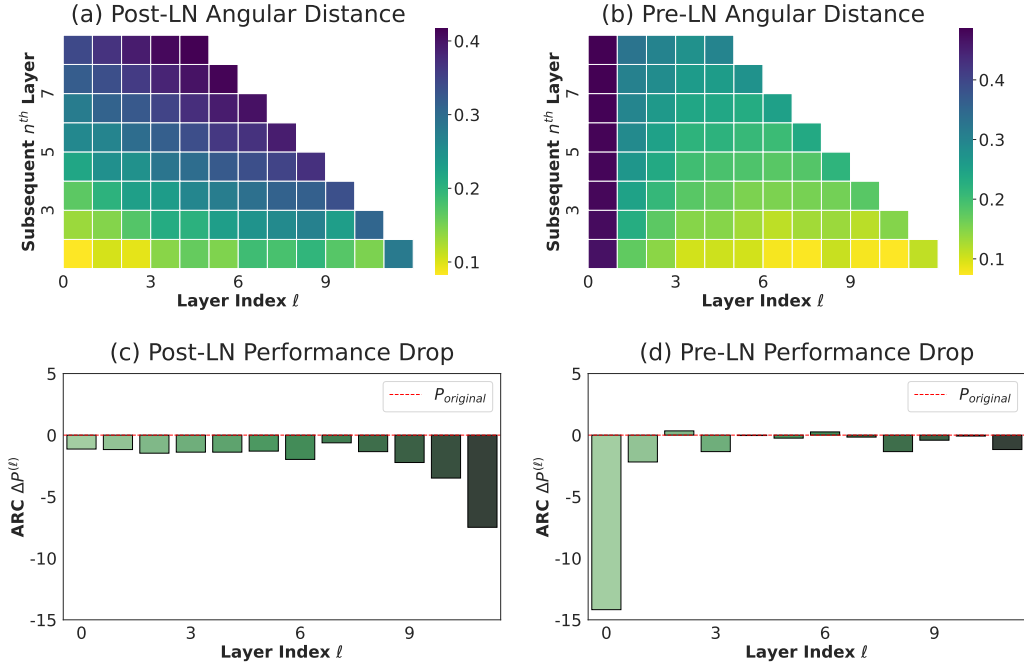


Figure 5: Results of in-house small-scale LLaMa-130M. **Angular Distance (a, b):** Each column represents the angular distance from the initial layer ℓ (x-axis) and its subsequent n^{th} layer (y-axis). The distance is scaled to the range $[0, 1]$, where yellow indicates smaller distances and purple indicates larger distances. **Performance Drop (c, d):** ARC-e performance drop of removing each single layer from LLaMa-130M.

624 C Results of In-house Small-scale LLaMa-130M

625 Figure 5 compares LLaMa-130M models differing only in Layer Normalization, clearly distinguishes
626 Post-LN from Pre-LN. In Post-LN models, early layers exhibit high similarity (low angular distance,
627 Fig. 5-a) and their removal causes minimal performance loss (Fig. 5-d), while deeper layers become
628 more distinct and critical. Post-LN also shows larger gradients in deeper layers but severe vanishing
629 in early layers at the start of training (Fig. 5-c). Conversely, Pre-LN LLaMa-130M demonstrates
630 a gradual decrease in angular distance with depth, resulting in highly similar deep layers (Fig. 5-
631 b). Removing most layers after the first in Pre-LN causes negligible performance loss (Fig. 5-d),
632 indicating their limited contribution. These consistent findings, observed in both open-weight and
633 in-house LLMs, lead to the conclusion that the widespread use of Pre-LN is the root cause of the
634 ineffectiveness of deep layers in LLMs.

635 D Training Loss Curve

636 We report the training loss curves of Pre-LN and LayerNorm Scaling in Figure 6. While LayerNorm
637 Scaling exhibits slightly slower convergence at the early stages of training, it consistently outperforms
638 Pre-LN as training progresses. We attribute this to the uncontrolled accumulation of output variance
639 in Pre-LN, which amplifies with depth and training steps, ultimately impairing the effective learning
640 of deeper layers. In contrast, LayerNorm Scaling mitigates this issue by scaling down the output
641 variance in proportion to depth, thereby enabling more stable and effective training across all layers
642 during training.

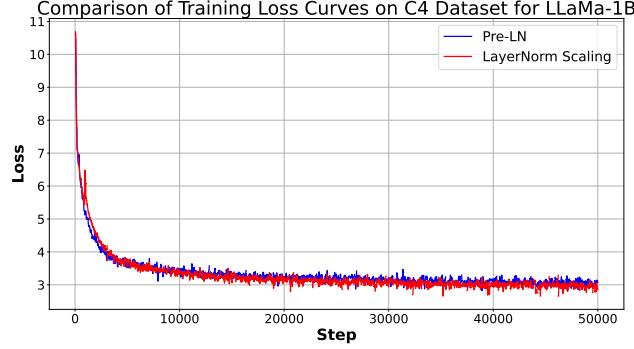


Figure 6: Training loss of LLaMA-1B with Pre-LN and LayerNorm Scaling.

E Variance Growth in Pre-LN Training

To analyze the impact of Pre-LN on variance propagation, we track the variance of layer outputs across different depths during training.

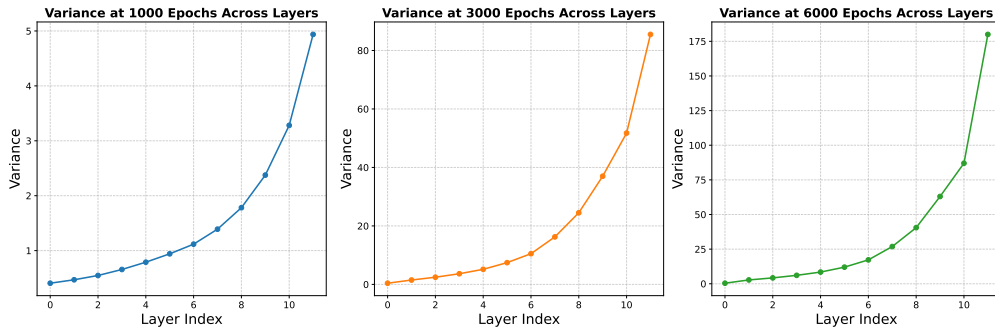


Figure 7: **Variance growth across layers in LLaMA-130M with Pre-LN.** Each subplot shows the variance at different training stages (1000, 3000, and 6000 epochs). In all cases, the variance follows an exponential growth pattern as depth increases, indicating that deeper layers experience uncontrolled variance amplification regardless of training progress.

Figure 7 illustrates the layer-wise variance in LLaMA-130M with Pre-LN at 1000, 3000, and 6000 epochs. Across all stages, variance remains low in shallow layers but grows exponentially in deeper layers, confirming that this issue persists throughout training rather than being a temporary effect. This highlights the necessity of stabilization techniques like LayerNorm Scaling to control variance and ensure effective deep-layer learning.

F Applicability to Vision–Language Models (Qwen 2.5-VL)

To examine whether the *Curse of Depth* also manifests in vision–language models (VLMs), we perform layer–pruning experiments on **Qwen 2.5-VL-7B** [4]. For both its vision encoder and language decoder, we prune one transformer layer at a time and directly evaluate the pruned model on the MMMU benchmark [53]. Figure 8 presents the resulting performance drops.

We observe that the **language branch** clearly suffers from the Curse of Depth, whereas the **vision branch** remains uniformly important. This suggests that the phenomenon is more pronounced in autoregressive language components of VLMs and may not directly transfer to vision encoders. A detailed modality–specific theoretical account is left to future work and community discussion.

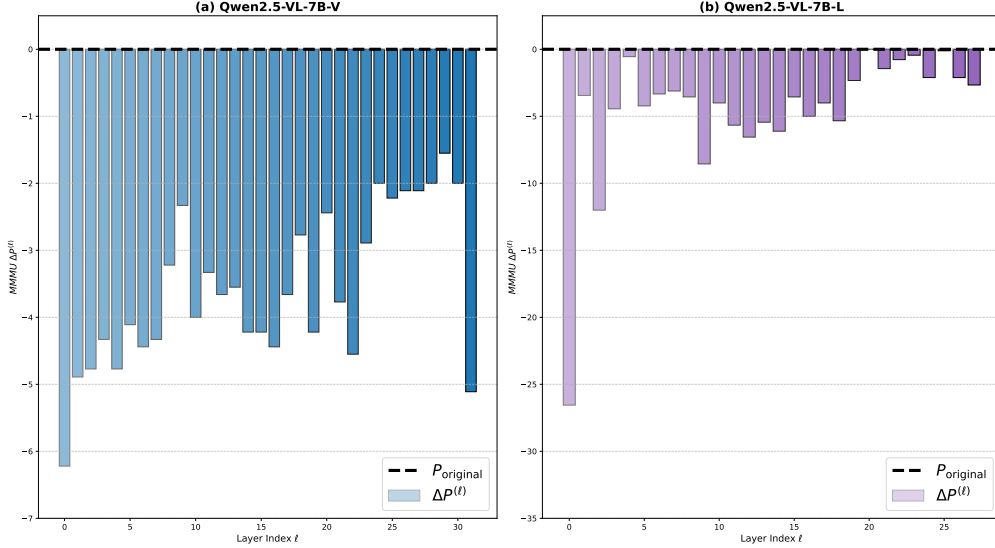


Figure 8: Performance drop of layer pruning on Qwen 2.5-VL-7B. (a) Vision branch shows relatively uniform sensitivity across layers. (b) Language branch exhibits a clear *Curse of Depth*: deeper layers contribute much less than early ones.

660 G LayerNorm Scaling in Vision Transformer

661 To evaluate whether LayerNorm Scaling (LNS) also benefits architectures beyond language models,
 662 we conduct experiments on ViT-S for image classification. Since ViT-S includes *LayerScale* by
 663 default—which may interfere with the effect of LNS—we remove *LayerScale* from all evaluated
 664 variants to ensure a fair comparison. We then test different insertion positions of LNS. The top-1
 665 accuracy results are summarized in Table 7. Whereas LNS in language models is typically most
 666 effective directly after normalization, in Vision Transformers, the best position is after the attention
 667 and MLP blocks. We next examine whether this performance gain correlates with better control of
 668 layer-wise variance.

Table 7: Top-1 accuracy (%) of ViT-S model with and without LNS.

Model Variant	LNS Position	Top-1 Accuracy
ViT (with LayerScale)	—	70.30
ViT (w/o LayerScale)	—	67.91
ViT (w/o LayerScale)	after LayerNorm	66.43
ViT (w/o LayerScale)	after Attn/MLP	68.75

669 Figure 9 plots the average output variance of each transformer block during training. Without
 670 LayerScale, variance in deeper layers grows rapidly—exceeding $\sim 3,000$ by 30K update steps.
 671 Applying LNS after Attn/MLP controls this growth to below ~ 150 , confirming that LNS stabilizes
 672 the forward signal even in vision transformers.

673 These preliminary findings indicate that the variance-control mechanism underlying LNS generalizes
 674 to vision transformers when the scaling is applied after Attn/MLP. We leave a more detailed theoretical
 675 understanding of this behavior to future work and community discussion.

676 H Post-LN Transformers

677 For Post-LN Transformers, we continue to adopt Assumption 1. In this setting, each layer is followed
 678 by a layer normalization (LN) step, ensuring that the variances $\sigma_{x_\ell}^2$ and $\sigma_{x'_\ell}^2$ remain fixed at 1 across

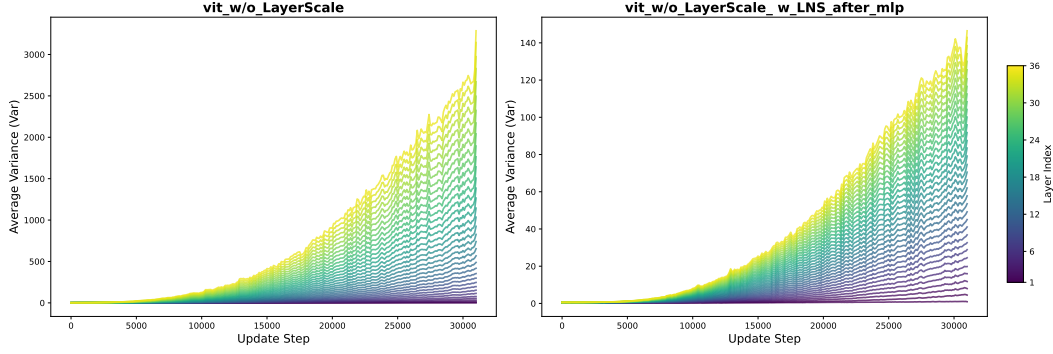


Figure 9: Layer-wise output variance of ViT-S **without** LayerScale (left) and with **LNS after Attn/MLP** (right). LNS significantly reduces the variance growth compared to the baseline.

all layers. Consequently, the norm $\left\| \frac{\partial y_\ell}{\partial x_\ell} \right\|_2$ exhibits minimal variation from one layer to the next, indicating stable gradient propagation.

Since the variance is effectively controlled by LN in Post-LN Transformers, an explicit variance-based analysis becomes less critical. Nonetheless, there remain other important aspects to investigate in deeper Post-LN architectures, such as the evolution of feature mappings and the behavior of covariance kernels over deep layers. These directions will be pursued in future work.

I Scaling Up Training

To confirm that our proposed **LayerNorm Scaling (LNS)** remains effective when both model capacity and training budget grow, we ran a controlled scaling study using the OLMO training stack on **The Pile** [15] corpus².

Setup. Four Transformer language models—60 M, 150 M, 300 M, and 1 B parameters—were pre-trained for a fixed 20 B-token budget sampled from The Pile. Except for the normalisation variant (Pre-LN vs. LNS), all hyper-parameters (optimizer, learning-rate schedule, batch size, sequence length) were identical across runs.

Results. Figure 10 shows training loss versus parameter count on a logarithmic scale. LNS (orange) outperforms the Pre-LN baseline (black) at every model size, reaching an **8 % relative improvement** at 1 B parameters ($3.13 \rightarrow 2.89$). This mirrors the larger-scale trends reported in Section 5.3 of the main paper, where LNS likewise reduced perplexity for 1 B–7 B parameter models and for an alternative architecture (Qwen2.5-0.5B).

J Limitations

While this work offers a comprehensive analysis of the Curse of Depth in LLMs and proposes LayerNorm Scaling as an effective remedy, several limitations remain:

Scope of Architectures. Our study primarily focuses on Transformer-based LLMs using Pre-LN. Although Pre-LN dominates modern architectures, our theoretical study does not cover models employing alternative normalization strategies (e.g., Post-LN only [13], normalization-free architectures [57]) or emerging paradigms such as mixture-of-experts or structured sparsity-based models.

Task Coverage. Most empirical evaluations, including pruning and angular distance analyses, were conducted using general-purpose benchmarks like MMLU. While these tasks reflect broad model capabilities, domain-specific or long-context reasoning tasks may reveal different dynamics in deep layer contributions, which we leave for future work.

²<https://pile.eleuther.ai/>

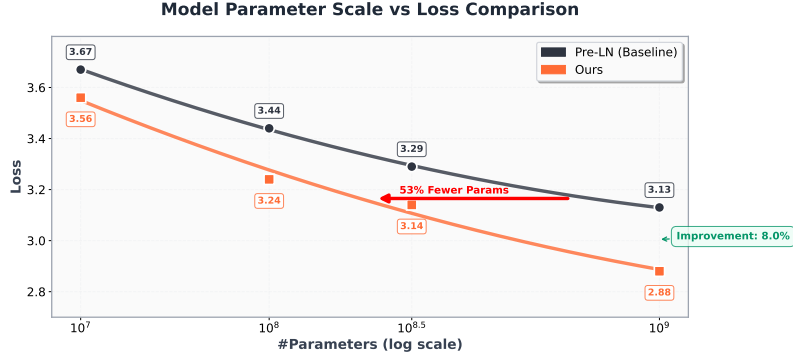


Figure 10: Training loss versus model size on The Pile. Dots: Pre-LN; squares: LNS. Numbers above markers indicate exact loss values.

Fine-grained Representation Quality. While LNS improves angular distance and performance sensitivity across layers, a deeper analysis of what types of information are represented or lost in deeper layers remains unexamined. For example, whether LNS helps preserve syntactic, semantic, or factual knowledge across depth is unclear.

K Broader Impact

The Curse of Depth phenomenon, identified and addressed in this work, has significant implications for the design, training, and deployment of LLMs. By revealing that deeper layers in modern Pre-LN Transformers often fail to contribute meaningfully to learning, our study prompts a reevaluation of how model capacity is allocated and utilized. This has both practical and ethical consequences.

From a computational efficiency perspective, the insights offered by this work can lead to more principled model pruning, layer reuse, or architecture design strategies that improve training and inference efficiency without compromising performance. In particular, LayerNorm Scaling enables deeper layers to be trained more effectively, maximizing the utility of each parameter and reducing unnecessary resource expenditure. This can help democratize access to powerful models by reducing the cost of pretraining and fine-tuning, especially for institutions or communities with limited computational resources.

From a sustainability standpoint, addressing CoD has the potential to lower the environmental impact of large-scale model training by mitigating wasteful computation. With LLMs increasingly deployed in industrial-scale applications, these gains can scale into substantial reductions in energy consumption and carbon footprint.

In terms of scientific understanding, this work contributes to the growing body of research that seeks to interpret and improve the internal dynamics of deep neural networks. By identifying the gradient-preserving failure modes induced by Pre-LN at depth, we provide both a diagnosis and a remedy that could influence future research in deep optimization, normalization strategies, and interpretability.

We also caution that increasing the efficiency of LLM training and deployment through techniques like LNS may further accelerate the proliferation of powerful LLMs, raising concerns around misuse, disinformation, or labor displacement. As such, our findings should be accompanied by responsible deployment practices and continued ethical oversight in the broader AI community.