DepthBench

Measuring How Residual Connections Enable More Computational Depth

A research paper reading presentation

arXiv:2609.32534v1 | 19 October 2026

Paper authors: Keyu Wang, Yangyi Huang, Jiale Kang,
David González-Martínez, Weiyang Liu, and Shiwei Liu.

Central finding

Residual design changes whether trading width for depth improves model quality.

Paper version: arxiv.org/abs/2609.32534v1. Main discussion: slides 2–21; supporting details: slides 22–26.

Trainable depth can still be weakly utilized

$$h_\ell = h_{\ell-1} + F_\ell\big(\mathrm{LN}(h_{\ell-1})\big)$$

- A direct residual path supports optimization of deep networks.
- Later updates can become weak, similar, or interchangeable.
- Adding layers at fixed width also adds parameters and compute.

DepthBench's question

At a similar parameter budget, which architectures benefit from a deeper, narrower shape?

Paper, Secs. 1–2.1, PDF pp. 2–3; Eq. (1).

The controlled variable is width–depth allocation

$$r = \frac{d}{L}, \qquad N \approx 12Ld^2 + 2Vd, \qquad V = 50280$$

Fixed total parameters

$$(L, d) = (16, 1216) \rightarrow (70, 640)$$

$$r : 76.0 \rightarrow 9.1$$

Backbone: $284\mathrm{M} \rightarrow 345\mathrm{M}$

Fixed backbone control

Hold the Transformer backbone near 300M.

Embedding/head parameters shrink with width.

Separates shape gains from backbone growth.

Approximate parameter matching; the formula describes the sub-1B MHA backbone.

Paper, Sec. 2.2, Tables 3 and 7, PDF pp. 4–5, 21.

Ten variants change how a layer accesses its past

Family	Architectures	Main intervention
Baseline	Pre-LN	One stream; fixed residual addition
Normalization	Sandwich-LN, LNS, DeepNorm, KEEL	Normalization, scaling, initialization
Multi-stream	HC, mHC	Learned read, write, and stream mixing
Cross-layer	Full/Block AttnRes, MoDA	Read earlier outputs or depth KV

Normalization adjusts scale; richer residual paths change history access.

Paper, Table 1 and Appendix A, PDF pp. 4, 19–20. MoDA uses Pre-LN in this benchmark.

Full AttnRes retrieves individual earlier outputs

$$a_{\ell i} = \mathrm{softmax}_i\left(w_\ell^\top \mathrm{RMSNorm}_{\mathrm{key},\ell}(v_i)\right), \qquad z_\ell = \sum_{i<\ell} a_{\ell i} v_i$$

$$f_\ell = F_\ell(\mathrm{RMSNorm}_{\mathrm{pre},\ell}(z_\ell))$$

- Sources: the embedding and individual sublayer branch outputs.
- Learned query $w_\ell \in \mathbb{R}^d$; token-dependent mixture weights.
- Zero-initialized queries start from uniform averaging; no $1/\sqrt{d}$ score scaling.
- The output head also reads a learned mixture of sources.

Implementation restated from Appendix A, PDF p. 19. z_ℓ denotes the mixed input; normalization is fused in the implementation.

HC propagates history through interacting streams

$$H_\ell = H_{\ell-1} A_\ell + F_\ell\big(\mathrm{LN}(H_{\ell-1}\alpha_\ell)\big)\beta_\ell^\top$$

- $H \in \mathbb{R}^{d\times m}$: $m = 4$ streams for each token.
- $\alpha_\ell \in \mathbb{R}^m$: read; $A_\ell \in \mathbb{R}^{m\times m}$: mix; $\beta_\ell \in \mathbb{R}^m$: write.

Explicit path from branch i to a later branch ℓ

$$w_{i\to\ell} = \beta_i^\top A_{i+1} \cdots A_{\ell-1}\alpha_\ell$$

History is recursively transported and recombined across streams.

Paper, Table 1, Sec. 4.2 and Appendix A, PDF pp. 4, 10, 20. Path coefficients describe realized residual transport, not the full network Jacobian.

Four suites test shape, scale, and budget effects

Suite	Budget	Shape coverage
Main comparison	∼400M total	10 variants at $L = 16$–32; deeper extensions
Backbone control	∼300M backbone	Configuration pool: $L = 16$–84
Multiple scales	200M–500M	Three aspect ratios near 28, 43, 76
Larger scale	∼1.6B total	$(L, d) = (28, 2048), (40, 1728), (54, 1504)$

- FineWeb-Edu; about 20 tokens/parameter: 8B tokens at 400M, 32B at 1.6B.
- Training context 2048; global batch 512 sequences; AdamW.
- Architecture-specific LR selection at 400M; 5×10^{-4} at 1.6B.

Paper, Sec. 2.3, Tables 2–4 and 7–9, PDF pp. 4–5, 21–22. The 1.6B suite uses GQA rather than MHA.

At 400M, the same shape change has opposite effects

$(16, 1216) \longrightarrow (32, 896)$ at approximately fixed total size

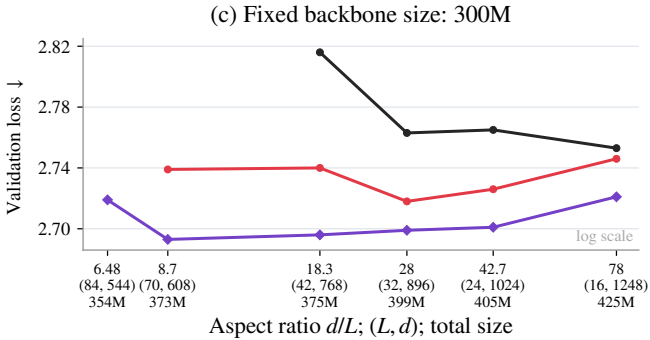
Architecture	Loss, $L = 16$	Loss, $L = 32$	Change
Pre-LN	2.759	2.782	$+0.023$
Full AttnRes	2.751	2.718	-0.033
HC	2.729	2.699	-0.030

Shape preference depends on residual design

Full AttnRes and HC improve between these endpoints; Pre-LN worsens. Validation loss is lower-is-better.

Paper, Sec. 3, PDF pp. 5–6; changes calculated from the reported losses.

Fixed-backbone scaling has a preferred depth

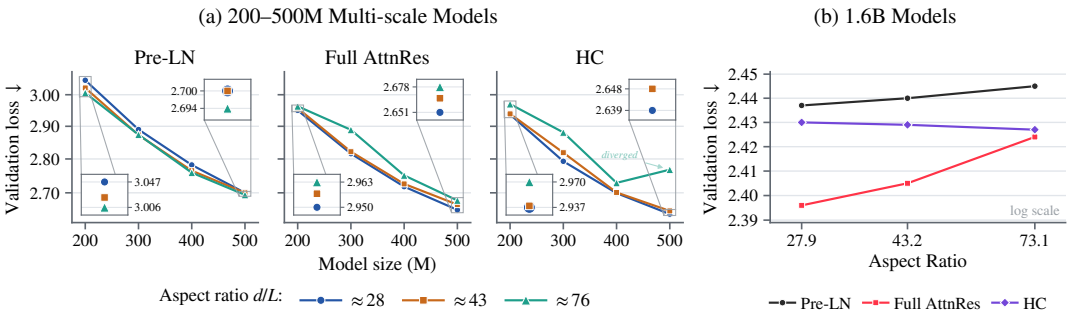


Black: Pre-LN Red: Full AttnRes Purple: HC

- Leftward means deeper; total model size decreases at similar backbone size.
- Full AttnRes favors $L \approx 32$; HC worsens from 70 to 84 layers.

Paper, Fig. 2(c), PDF p. 6. Original panel cropped; axes and configuration labels retained.

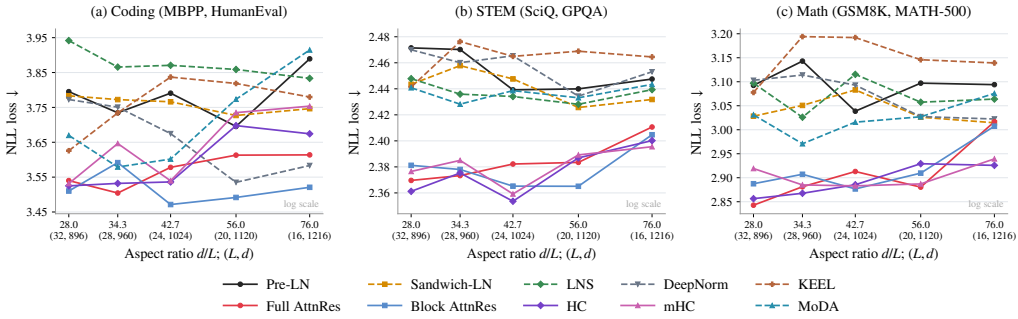
Full AttnRes has the more consistent scale evidence



- Full AttnRes benefits from deeper shapes across 200M–500M and at 1.6B.
- HC: the widest 500M run diverges; the 1.6B curve has no clear depth gain.
- Pre-LN also gains slightly from depth at 1.6B; the 400M trend is not universal.

Paper, Fig. 3 and Sec. 3, PDF pp. 6–7. Optimization sensitivity is the authors' explanation for HC's instability.

The gains extend to domain-specific likelihood

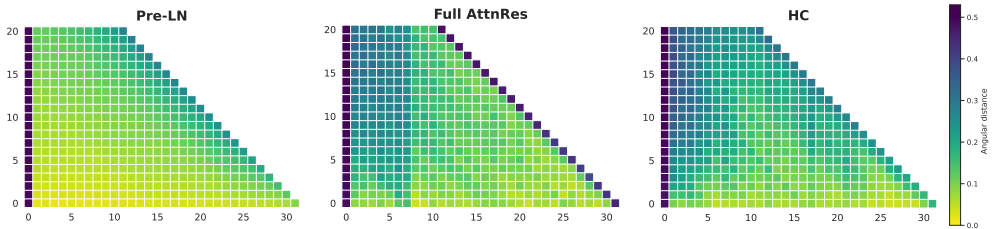


- Each panel averages teacher-forced NLL over two tasks; lower is better.
- Deep HC/Full AttnRes generally help, with non-monotonic intermediate points.
- These are reference-sequence likelihoods, not generation pass rates.

Paper, Fig. 4, PDF p. 7. Coding: MBPP/HumanEval; STEM: SciQ/GPQA; math: GSM8K/MATH-500.

Representations retain more layer-specific structure

$$d(h_\ell, h_{\ell+n}) = \frac{1}{\pi} \arccos \frac{h_\ell \cdot h_{\ell+n}}{\|h_\ell\|_2 \|h_{\ell+n}\|_2}$$



- x: starting layer; y: layer offset. Yellow means similar directions.
- Full AttnRes and HC show less smooth structure than Pre-LN.
- This probe is learning-rate sensitive; distance alone is not utility.

Paper, Eq. (2), Fig. 5, PDF p. 8. Selected original panels rearranged; shared colorbar retained. LR 2×10^{-3} for this comparison.

Layer interventions probe dependence and ordering

Skip a layer: how much does a later update change?

$$C(s,\ell) = \frac{\|(h_{\ell+1} - h_\ell) - (\bar{h}_{\ell+1} - \bar{h}_\ell)\|_2}{\|h_{\ell+1} - h_\ell\|_2}$$

Swap two layers: how much does the loss change?

$$P(\ell_1,\ell_2) = \frac{|\mathcal{L}(M) - \mathcal{L}(M_{\text{swap}})|}{\mathcal{L}(M)}$$

- $\bar{h}$ denotes hidden states after skipping layer s.
- The absolute-value permutation score measures change, not its sign.

Paper, Eqs. (3)–(4), PDF p. 9. Causal score measures relative update perturbation, not task accuracy.

Late layers are more dependent and less interchangeable

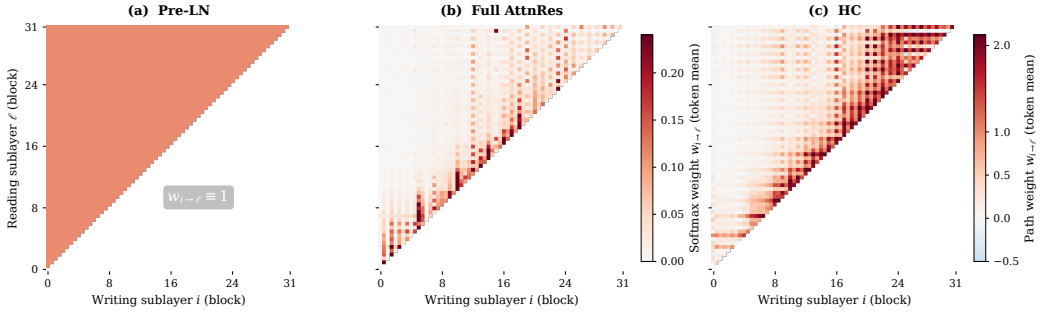
Restrict layer pairs to the last 3/4 of a 32-layer model.

Architecture	Fraction with $C > 0.45$	Fraction with $P > 0.15$
Pre-LN	$\sim 1\%$	$\sim 8\%$
Full AttnRes	$\sim 16\%$	$\sim 46\%$
HC	$\sim 7\%$	$\sim 35\%$

- Removing early layers from the summary exposes differences in later computation.
- These are thresholded *pair fractions*, not percentages of effective layers.
- Strong perturbation responses do not by themselves imply favorable shape scaling.

Paper, Fig. 6 and Eqs. (5)–(8), PDF pp. 9–10. Approximate percentages reported in the text.

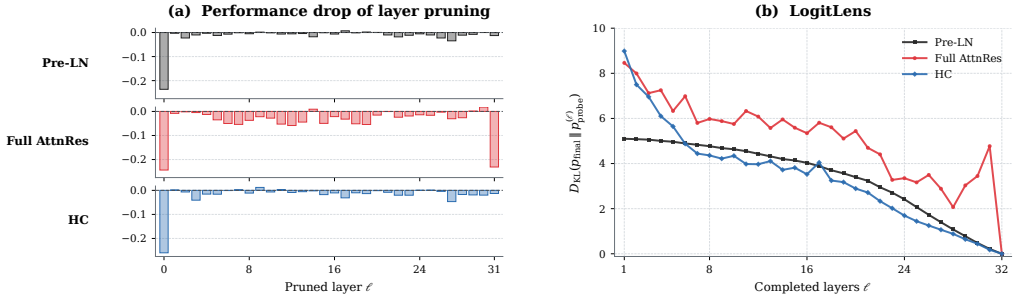
Full AttnRes retrieves; HC transports and recombines



- Pre-LN: explicit residual coefficients $w_{i\to\ell} = 1$.
- Full AttnRes: non-negative, normalized weights over stored sources.
- HC: products of write, transport, and read maps; coefficients may be signed.

Paper, Fig. 7 and Sec. 4.2, PDF pp. 10–11. Panels use different color scales; coefficient magnitude is not a common importance score.

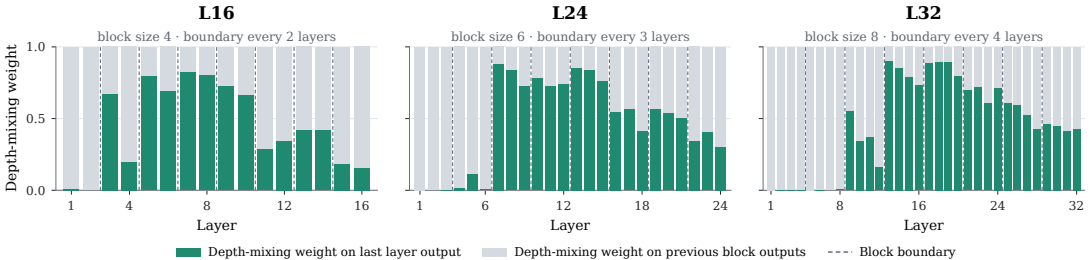
Stronger cooperation does not make every layer essential



- ARC-Easy pruning: HC and Pre-LN depend strongly on the first layer.
- Full AttnRes is sensitive over more layers, including the final layer.
- LogitLens: Full AttnRes follows a less monotonic prediction trajectory.

Paper, Fig. 8, PDF p. 11. KL is $D_{\mathrm{KL}}(p_{\mathrm{final}} \| p_{\mathrm{probe}})$, so it measures agreement with the final model prediction.

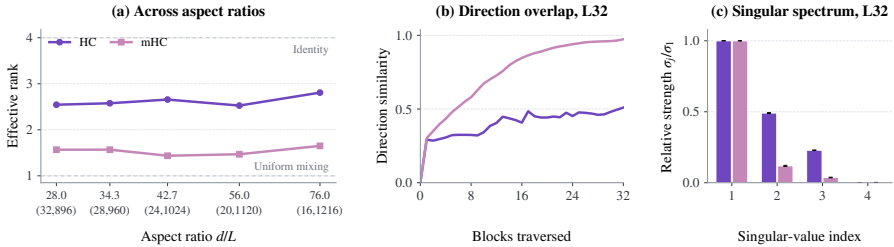
Block aggregation can weaken early sequential reuse



- Fixed eight blocks: sublayer block size $b = 2L/8$ grows with depth.
- Green: weight on the current running sum. Early near-zero spans grow to about 2, 5, and 8 layers.
- Plausible mechanism: a new update is computed but receives little immediate use.

Paper, Fig. 9, PDF p. 12; Appendix A, p. 19. Evidence is for this schedule; no isolated block-count intervention is reported.

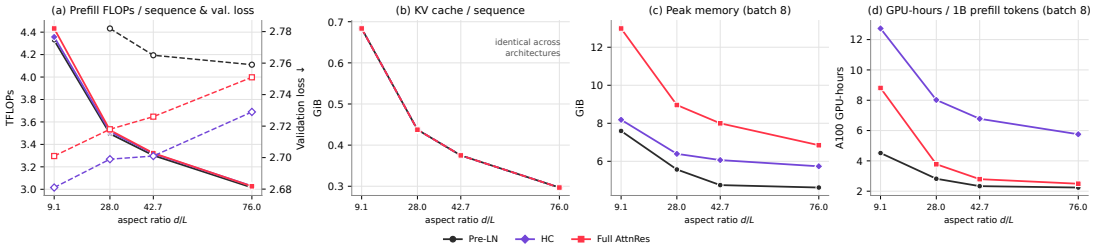
mHC's residual transport concentrates stream directions



- Effective rank: HC 2.53–2.81; mHC 1.44–1.65.
- mHC directions approach collinearity; smaller singular values decay.
- For doubly stochastic A, $AU = UA = U$, with $U = \mathbf{1}^\top/m$.

Paper, Fig. 10 and Eq. (10), PDF pp. 12–13. Convergence toward U requires contraction of nonuniform directions. Intervals use document bootstrap.

Depth increases prefill costs at similar parameter size



- $r : 76.0 \to 9.1$: about $3.0 \to 4.4$ TFLOPs/sequence; KV cache $> 2\times$.
- Full AttnRes has the largest peak memory; HC the highest measured GPU-hours.
- Near-fixed Ld^2 does not fix attention T^2Ld or cache TLd.

Paper, Fig. 11, PDF p. 13. 400M, $T = 4096$, BF16, A100-SXM4-40GB; batch 8 for memory/time. Time is per 1B prefill tokens, not training time.

The scope of control matters for interpretation

- **Parameter overhead varies.** At $L = 24$: Pre-LN 405.41M; Full AttnRes 405.51M; HC 405.70M; mHC 410.13M; MoDA 453.64M.
- **Optimization differs.** LNS needs a larger LR; HC becomes less stable at larger scales; some LR choices are transferred.
- **Shape and compute change together.** Matching parameters does not match prefill FLOPs, activation memory, or latency.
- **Diagnostics answer different questions.** Similarity, internal dependence, pruning sensitivity, and predictive quality need not rank models identically.

Paper, Table 6, Appendix C, Secs. 3–4, PDF pp. 6–13, 20, 22–23. MoDA's overhead is about 11.9% relative to Pre-LN.

Accessible history is part of the depth-scaling design

1. **Residual design changes shape preference.**
   Full AttnRes and HC make deeper shapes useful in the 400M benchmark.
2. **History access has different realizations.**
   Explicit retrieval and recursive multi-stream transport preserve differentiated routes.
3. **The practical optimum also depends on cost.**
   Choose depth, routing granularity, optimization, and hardware budget together.

Next experiments suggested by this reading

Vary block count; isolate mHC constraints; compare at matched FLOPs or GPU-hours.

Synthesis of the paper's results. Proposed experiments are reading discussion, not reported findings.

Reference and author-provided resources

Keyu Wang, Yangyi Huang, Jiale Kang, David González-Martínez, Weiyang Liu, Shiwei Liu.

DepthBench: Measuring How Residual Connections Enable More Computational Depth.

arXiv:2609.32534v1, submitted 26 September 2026.

- Versioned paper on arXiv
- DepthBench code repository
- Author checkpoint collection

Figures are from the paper's v1 TeX source. Figure 2(c) is cropped; Figure 5 uses selected panels rearranged for readability.

Paper identity and resource links: PDF p. 1. All page references count from the first PDF page.

Backup: normalization methods change different scales

$$\text{Sandwich-LN:} \quad h_\ell = h_{\ell-1} + \mathrm{LN}_{\mathrm{out}}(F_\ell(\mathrm{LN}_{\mathrm{in}}(h_{\ell-1})))$$

$$\text{LNS:} \quad h_\ell = h_{\ell-1} + F_\ell(\ell^{-1/2}\mathrm{LN}(h_{\ell-1}))$$

$$\text{DeepNorm:} \quad h_\ell = \mathrm{LN}(\alpha h_{\ell-1} + F_\ell(h_{\ell-1}))$$

$$\text{KEEL:} \quad h_\ell = \mathrm{LN}_{\mathrm{post}}(\alpha h_{\ell-1} + F_\ell(\mathrm{LN}_{\mathrm{pre}}(h_{\ell-1})))$$

- LNS shares the block-index scale between attention and FFN.
- DeepNorm: $\alpha = (2L)^{1/4}$, with depth-aware initialization.
- KEEL: gain based on sublayer count, with a special first block.

Paper, Table 1 and Appendix A, PDF pp. 4, 19. Generic sublayer equations; method-specific indexing/initialization follows Appendix A.

Backup: learning rate and probe definitions matter

- 400M optimal LR: most methods 2×10^{-3}; DeepNorm 10^{-3}; LNS 10^{-2}.
- Figure 5 instead fixes LR at 2×10^{-3}; LNS is suboptimal and DeepNorm is excluded after divergence.
- Block AttnRes borrows Full AttnRes's LR. mHC LR sweeps use a near-uniform initialization; main results use GAP-8.

LogitLens uses architecture-dependent states

Pre-LN: current residual stream. HC: sum of four streams. Full AttnRes: the next layer's depth-query mixture.

$$p_{\text{probe}}^{(\ell)} = \text{softmax}\big(W_{\text{LM}}\,\text{Norm}(h_\ell)\big)$$

Paper, Table 4; Sec. 4.2; Fig. 5 and Fig. 12 captions; Appendix C, PDF pp. 5, 8, 10, 22–23.

Backup: doubly stochastic does not imply collapse alone

$$U = \frac{1}{m}\mathbf{1}\mathbf{1}^\top, \qquad U^2 = U$$

An explanatory example, not a measured model

For $A = (1-\varepsilon)I + \varepsilon U$, $0 < \varepsilon < 1$,

$$A^k = U + (1-\varepsilon)^k (I - U).$$

The uniform direction survives; differences between streams contract.

- Identity and permutation matrices are also doubly stochastic, without this contraction.
- Measured transport rank is not the rank of the full nonlinear network.

Teaching derivation motivated by Eq. (10), PDF p. 13. The paper's explanation explicitly conditions on contraction of other directions.

Backup: the FLOP accounting explains the trade-off

$$F_{\mathrm{Pre-LN}} = L(8Td^2 + 6Td\,d_{\mathrm{ff}} + 2T^2d) + F_{\mathrm{head}}$$

$$\Delta F_{\mathrm{HC}} = 192TdL, \qquad \Delta F_{\mathrm{mHC}} = 480TdL \quad (m = 4),$$

$$\Delta F_{\mathrm{Full}} = 4TdL(2L + 3) \approx 8TdL^2.$$

- One multiply–add counts as two FLOPs; causal attention evaluates about half of query–key pairs.
- $F_{\mathrm{head}} = 2TdV$ for all-token logits, or $2dV$ for last-token logits.
- Counts omit normalization, softmax, and other elementwise operations.
- For fixed backbone Ld^2: $Ld \propto \sqrt{L}$, increasing KV and sequence attention costs.

Paper, Appendix D, PDF pp. 23–24. The final proportionality is an explanatory fixed-backbone approximation.