

# Loss Scaling and Long-Context Retrieval Across Attention Variants



Billy Dickson<sup>1</sup> and Zoran Tiganj<sup>1,2</sup>

<sup>1</sup> Department of Computer Science   <sup>2</sup> Department of Psychological and Brain Sciences

RESEARCH QUESTION

Do attention variants with similar loss curves also retrieve information equally well from long contexts?

01 Study design

REPEATING FOUR-LAYER PATTERN

|                |   |   |   |   |
|----------------|---|---|---|---|
| Full softmax   | F | F | F | F |
| Sliding-window | S | S | S | F |
| HGDN           | G | G | G | F |
| HC4            | G | G | G | F |

HC4 adds Identity Hyper-Connections (four streams).

F full attention · S 4096-token window · G Gated DeltaNet

Four architectures share a training recipe while changing attention and residual connections.

- All hybrids use full attention every fourth layer, including the final layer.
- HGDN is hybrid Gated DeltaNet.
- HC4 uses learned weights to combine the four residual streams into each sublayer’s input and to add its output back to the streams.

60M · 100M · 190M · 370M

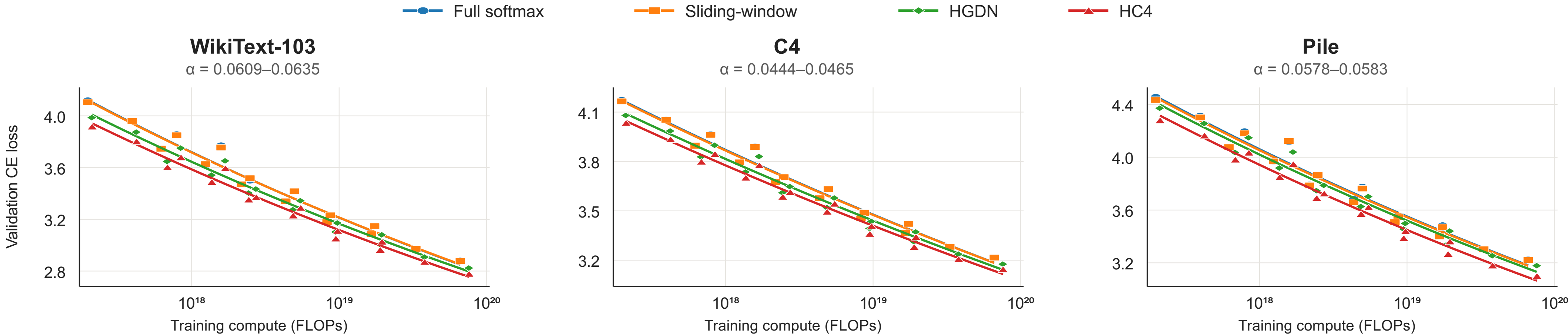
10, 20, 40, 80 training tokens per parameter

16 runs → 64 evaluated checkpoints

02 Loss improves at similar rates

HGDN and HC4 reach lower fitted losses, while full softmax and sliding-window curves nearly overlap.

Built on OLMo-core · FineWeb · Dolma2 tokenizer · 8192-token sequences · Muon + AdamW · WSD-S learning-rate schedule



TRAINING COMPUTE

$C = 6ND$

N excludes token embeddings.  
D counts cumulative training tokens.

LOSS-COMPUTE FIT

$L(C) = A C^{-\alpha}$

A sets the loss scale;  $\alpha$  sets its rate of improvement with compute.

EXPONENT COMPARISON

$\Delta\alpha_v = \alpha_v - \alpha_{full}$

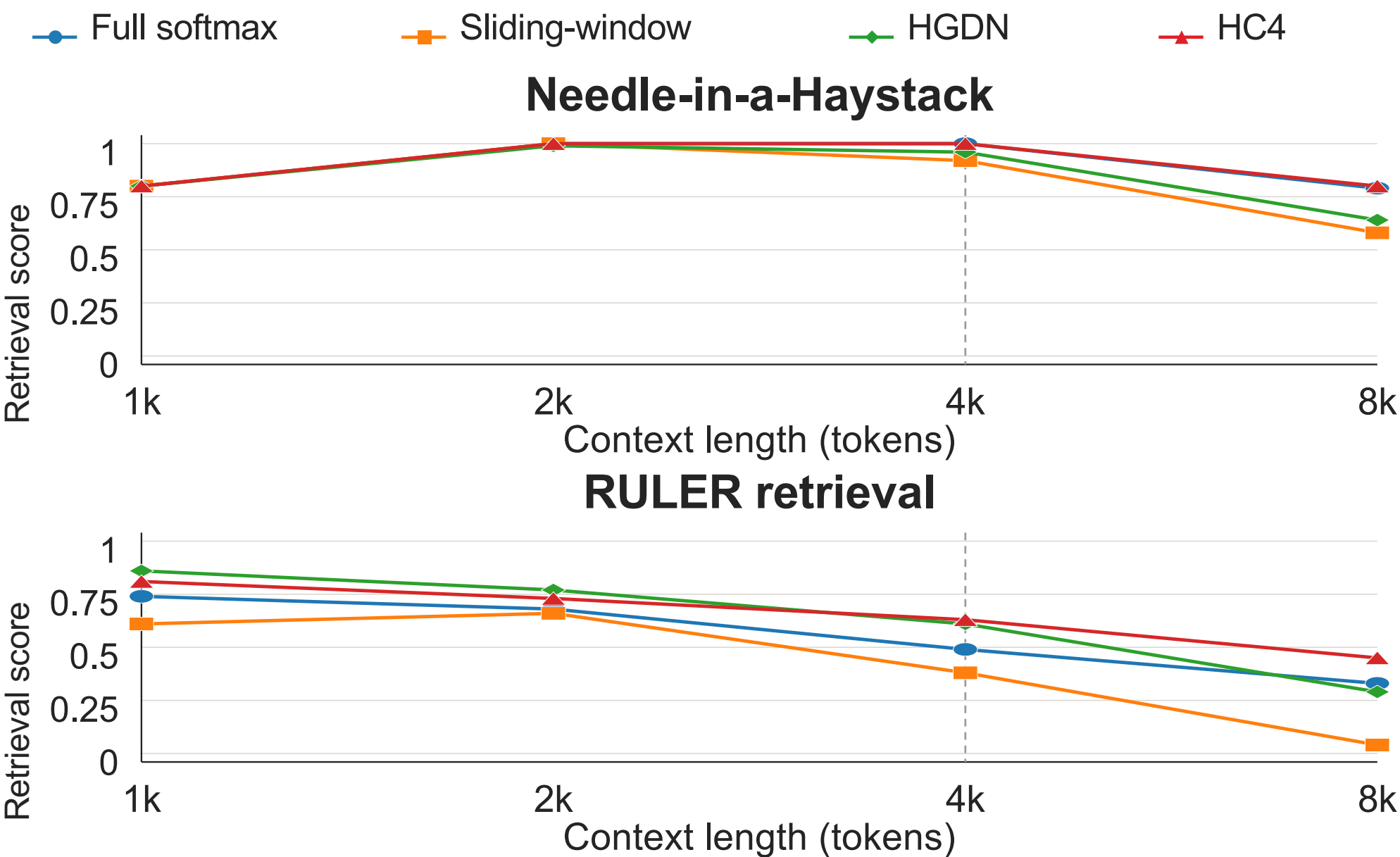
Compare each variant’s exponent with the full-softmax baseline.

Each fit uses all 16 checkpoints. Loss is validation cross-entropy in nats per token.

Across nine datasets, bootstrap comparisons do not establish different rates of loss improvement.

03 Retrieval separates the architectures

At 370M, sliding-window retrieval drops sharply beyond its 4096-token window.



Scores range from 0 to 1. Per length, NIAH averages 100 examples across five insertion positions; RULER averages 40 examples across multi-key and multi-value retrieval. All inputs are within the 8192-token training length. Dashed lines mark the 4096-token window.

RULER at 8k   Full 0.33   Sliding 0.04   HGDN 0.29   HC4 0.45

04 Compute at matched loss

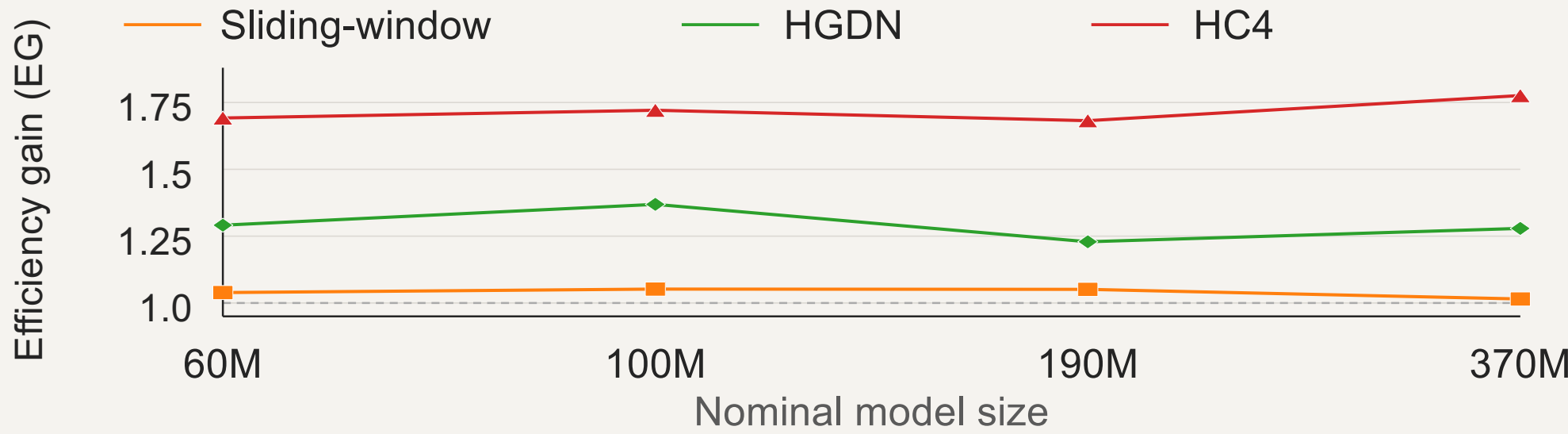
Following The Microsoft AI Team (2026), §2.2.2.

Mean validation loss across nine datasets.  
80 training tokens per parameter ( $D/N = 80$ ).

$$f_{full}(C) = A_{full} C^{-\alpha_{full}} + E_{full}$$
$$EG = \frac{f_{full}^{-1}(L')}{C'} = \frac{1}{C'} \left( \frac{A_{full}}{L' - E_{full}} \right)^{1/\alpha_{full}}$$

$L'$  is observed mean loss;  $C'$  is checkpoint training compute.

$E_{full}$  is the fitted loss floor;  $L' > E_{full}$ .  $EG > 1$  favors the variant.



1.78× estimated full-softmax compute to match 370M HC4 loss

Fit-based training-FLOP estimate; the 370M comparison extends beyond the measured full-softmax baseline.

05 Takeaways

Similar prediction loss can accompany substantial differences in long-context retrieval.

Loss improves at similar fitted rates.

HGDN and HC4 reach lower losses than full softmax under the shared training recipe.

Retrieval needs its own evaluation.

At 8k tokens, sliding-window RULER scores 0.04 versus 0.33 for full softmax, despite similar loss.

Residual connections also affect loss and retrieval.

HC4 adds four residual streams to HGDN and reaches lower fitted loss. At 370M and 8k tokens, its RULER retrieval score is 0.45 compared with HGDN’s 0.29.

Selected references

Merrill et al. (2026), Olmo Hybrid · Yang et al. (2025), Gated Delta Networks · Zhu et al. (2025), Hyper-Connections  
Hsieh et al. (2024), RULER · Wen et al. (2024), Warmup-Stable-Decay learning rates  
The Microsoft AI Team (2026). MAI-Thinking-1: Building a Hill-Climbing Machine. Microsoft AI technical report.

dicksonb@iu.edu · ztiganj@iu.edu

Supported by a Modal for Academics compute grant.