A Provable Expressiveness Hierarchy in Hybrid Linear-Full Attention

Finite-state communication and sequential retrieval

Paper: Xiaowei Ye, Xiaoyu He, Chao Liao,
Chen Wu, and Pinyan Lu

arXiv:2602.01763v1 | 2 February 2026
Research discussion prepared on 15 October 2026

On a constructed task family, many recurrent layers cannot replace enough rounds of query-dependent access to the original input.

Paper reading presentation; presenter is not a paper author. arxiv.org/abs/2602.01763v1

The question is whether recurrent depth buys retrieval rounds

A concrete dependency

Retrieve one value, turn it into the next address, and retrieve again.

- Full attention accesses prefix tokens using the *current* query.
- A recurrent layer passes a finite summary formed before future queries.
- A hybrid combines both: can many summaries compensate for fewer full-attention layers?

Reading: Count both state capacity and opportunities to access the input.

Paper, Secs. 1–2, PDF pp. 1–5. The paper is theoretical; it reports no training experiments.

The headline separation comes with a capacity condition

For a constructed L-step task, let $A = 2^{3L^2}$ and $c = Hdp$.

Architecture	Capacity result
L full-attention layers	Polynomial capacity necessary
$L + 1$ full-attention layers	$c = \text{polylog}(n)$ is sufficient
$L - 1$ full layers, each followed by A recurrent layers	Fails if $c \leq n^{2^{-4L-2}}$

Fixed L; causal models; deterministic, exact correctness on every input.

Paper, Theorem 1.1, PDF pp. 1–2; Table 1, p. 2. The first two rows are attributed to Chen et al.

Finite precision turns representations into bit budgets

Model quantities

H: attention heads

d: dimension per head

p: bits per number

n: prompt tokens

Two budgets

Token representation:

$$c = Hdp.$$

Standard linear-attention state:

$$C = Hd(d+1)p.$$

Local MLPs may be arbitrary functions. The lower bound restricts information flow between positions.

Paper, Secs. 2.1–2.2 and Lemma 2.2, PDF pp. 3–4. These quantities are not total parameter counts.

A matrix and a vector carry the linear-attention prefix

With column-vector notation and $q_i, k_i, v_i \in \mathbb{R}^d$:

$$S_i = S_{i-1} + v_i \varphi(k_i)^\top, \qquad Z_i = Z_{i-1} + \varphi(k_i),$$

$$S_i \in \mathbb{R}^{d \times d}, \quad Z_i \in \mathbb{R}^d, \qquad y_i = \frac{S_i \varphi(q_i)}{Z_i^\top \varphi(q_i)}.$$

- Each head carries $d^2 + d$ numbers across a token boundary.
- Arbitrary recurrent updates can fit the communication framework.
- For a particular architecture, count its *actual total state bits*.

Paper, Lemma 2.2, p. 4; Appendix D, p. 24. Formula orientation normalized for column vectors.

Each function lookup determines the next address

$$N_0 = m, \qquad N_\ell = m \prod_{j=1}^{\ell} n_j, \qquad z_\ell : [N_{\ell-1}] \to [N_{\ell-1}].$$

$$i_0 = z_0, \qquad i_1 = z_1(i_0), \qquad i_{\ell+1} = z_{\ell+1}(w_\ell, i_\ell).$$

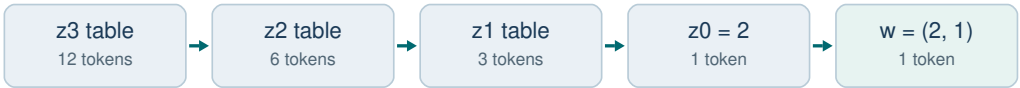
- Query: $w_\ell \in [n_\ell]$. Encode (w_ℓ, i_ℓ) in $[N_\ell]$.
- Return i_L at the final query position.
- Prompt order: $z_L \mid \cdots \mid z_1 \mid z_0 \mid w$.

The causal scan encounters the tables before it knows the needed addresses.

Paper, Definition 2.5, PDF pp. 4–5. Product indices follow the consistent Eq. (7), p. 17.

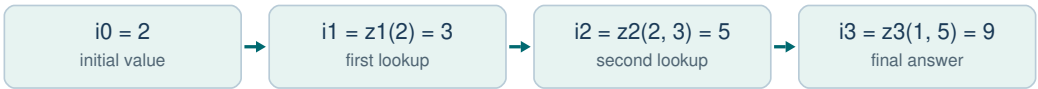
A three-step example makes the order mismatch visible

Input order: tables come before the query



m = 3, n1 = n2 = 2; total prompt length = 23 tokens

Computation order: every lookup supplies the next address



Teaching example based on Definition 2.5 (PDF pp. 4-5); not a hard-instance parameter setting.

Original teaching diagram based on Definition 2.5, PDF pp. 4–5. Other function entries may be arbitrary.

One initial read plus L lookups gives the upper bound

1. At the query token, retrieve the initial value z_0.
2. Use a full-attention layer to retrieve $z_1(z_0)$.
3. For each subsequent layer, construct (w_ℓ, i_ℓ) and retrieve from $z_{\ell+1}$.

Why retrieval is cheap in representation size

Binary addresses use $O(\log n)$ dimensions. A scaled score gap concentrates softmax on the matching address; local readout recovers the discrete value.

Paper, Table 1, p. 2 (Chen et al. upper bound); Appendix C, pp. 23–24 (retrieval head).

The quantitative guarantee is strongly asymptotic

L	Recurrent layers per block	Failure condition
2	$2^{12} = 4{,}096$	$Hdp \leq n^{1/1024}$
3	$2^{27} = 134{,}217{,}728$	$Hdp \leq n^{1/16384}$
4	2^{48}	$Hdp \leq n^{1/262144}$

Total recurrent depth is $(L-1)2^{3L^2}$.

A scale check

At $L = 2$ and $n = 2^{20}$, the threshold is $Hdp \leq 1.0136$, below the proof's assumption $Hdp \geq 2$.

Values substituted into Theorem 1.1; the scale check is our calculation, not an experiment.

The protocol exposes two different information channels

Players $L, L-1, \ldots, 0, -1$ hold $z_L, z_{L-1}, \ldots, z_0, w$.

Channel	Information flow	Message budget
Full attention	Query-dependent replies from prefix blocks	$2Hdp\, m_{(i)}$ bits per sender
Recurrent layer	State passed along $L \to \cdots \to -1$	$Hd(d+1)p$ bits per edge

One epoch is one full-attention round followed by a_ℓ recurrent rounds.

Prefix players forget received future queries after replying: causal states remain causal.

Paper, Sec. 3, PDF p. 5. $m_{(i)}$ is the number of tokens held by the receiving player.

Indistinguishability must survive later full-attention rounds

Construct a pair of input sets:

$$R_{\geq \ell} \subseteq A_L \times \cdots \times A_\ell, \qquad Z_{<\ell} = Z_{-1} \times \cdots \times Z_{\ell-1}.$$

A_j contains all possible functions held by player j.

Invariant

For each fixed $z_{<\ell}$, all inputs in $R_{\geq \ell}$ give the same relevant soft and recurrent transcripts through epoch ℓ.

Preserve both:

- many possible prefix functions in $R_{\geq \ell}$;
- many reachable intermediate values in $\mathcal{I}_{\ell-1}(Z_{<\ell})$.

Paper, Definition 4.1, p. 6; Definition B.4 and Lemmas B.5–B.7, pp. 18–19.

The induction fixes messages while retaining input diversity

1. **Initialize.** Fix a first-round message pattern by pigeonhole; retain many functions and reachable values.
2. **Extract a rectangle.** Choose $S_1 \times S_2 \subseteq R_{\geq \ell}$ and set $Z_\ell = S_2$.
3. **Fix new soft transcripts.** Earlier indistinguishability removes their dependence on the choice of z_ℓ.
4. **Fix recurrent transcripts.** Charge their total bits to the allowable set-size loss.

Reading: Lemma B.11 tracks recurrent messages so they cannot reintroduce hidden distinctions.

Paper, Secs. 4.2–4.3, pp. 7–8; Appendix B.2–B.3, pp. 20–23. Base-case qualification on slide 14.

Too many functions must agree on too many addresses

At the final epoch, let $N = N_{L-1}$ and

$$t = n_{L-1}|\mathcal{I}_{L-1}(Z_{<L})|.$$

- Identical transcripts force all surviving z_L to agree on t reachable addresses.
- At most N^{N-t} functions can share those values.
- The invariant retains at least N^N/Δ_L functions.

Contradiction

The parameters ensure $\Delta_L < N^t$, hence

$$N^N/\Delta_L > N^{N-t}.$$

Teaching expansion of Lemma B.5, PDF p. 19, using Eqs. (10)–(11), p. 17.

The first-epoch distinction matters in the v1 proof

- Sec. 4.1 uses $(L, 0, A, \ldots, A)$: no recurrent layer after the first full layer.
- Appendix B states the stronger allowance $a_1 \leq 1$.
- Its base case fixes a soft transcript and claims the query state is independent of z_1.
- With $a_1 = 1$, a recurrent message can still carry z_1; this dependence needs an additional argument.

How we use the argument

Explain the main route with $a_1 = 0$. Prepending an identity full layer embeds the headline $L - 1$-full-layer architecture in this layout.

Our reading concern: compare Sec. 4.1, p. 6; Theorem B.2, pp. 17–18; base case, p. 20. Not a counterexample to Theorem 1.1.

Lookup and permutation tasks yield simpler state lower bounds

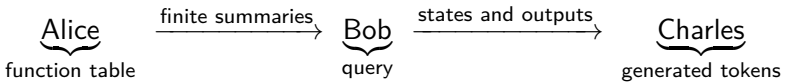
$\mathrm{PerCom}(\sigma, \tau) = (\sigma(\tau(1)), \ldots, \sigma(\tau(n)))$ for two permutations.

Task	Linear attention needs	Log-linear needs
$\mathrm{Eva}(f, x) = f(x)$	$LHd(d+1)p \geq n \log_2 n$	$LHd^2 p = \Omega(n)$
$\mathrm{PerCom}(\sigma, \tau)$	$LHd(d+1)p \geq \log_2(n!)$	$LHd^2 p = \Omega(n)$

- A prefix summary must distinguish n^n functions or $n!$ permutations.
- Log-linear communication carries $O(\log n)$ matrix states per layer.
- One full-attention layer suffices with $Hdp = \mathrm{polylog}(n)$.

Paper, Theorems A.3–A.8, PDF pp. 12–15; Appendix C, pp. 23–24.

Appended CoT cannot restore information absent from the state



- Charles receives only information already available to Bob.
- Bob could simulate all of Charles's subsequent computation.
- Thus these pure recurrent/log-linear retrieval bounds also hold with appended CoT under this access model.

The main hybrid theorem is a single-forward-pass result.

Paper, Appendix A.2–A.4, PDF pp. 13–15. A new channel to original tokens changes the model.

Sparse retrieval first compresses, then chooses blocks

For $q = n/B$ blocks, the query player receives:

1. an Hdp-bit summary from each block;
2. all raw tokens from k selected blocks.

The collision construction

Two blocks a, b have the same summary but different value sets. Put b in every inspected block. An uninspected block can switch from b to a and change membership of a value v.

Querying $-v \bmod M$ converts that hidden difference into a 2-Sum answer.

Paper, Definition 2.4, p. 4; Sec. 5, p. 8; Appendix A.6, p. 16. Shared compression and summary-based selection.

The sparse counting bound needs a range for B and k

For a value universe of size n, the collision argument requires

$$Hdp \geq \log_2 \left(\sum_{j=1}^{\min(B,n)} \binom{n}{j} \right).$$

- There must be an uninspected block: for one head, $k < n/B$.
- For $1 \leq B \leq n/2$, the count scales as $\Theta(B \log(n/B))$.
- It yields $\Omega(B \log n)$ when $B \leq n^{1-\varepsilon}$ for fixed $\varepsilon > 0$.

Theorem 1.2 states $\Omega(B \log n)$ without these qualifications.

Our range analysis of Appendix A.6, p. 16; compare Theorem 1.2, p. 2. With multiple heads, use the union of inspected blocks.

The 2-Sum appendix changes the value universe

Definition 2.6

$M = O(n)$.

At most $2^M = 2^{O(n)}$ distinct value sets.

Proof of Theorem A.9

Takes $M = n^2$.

$$\log_2 \binom{n^2}{n-1} = \Theta(n \log n).$$

- The larger universe supports the proof's $n \log n$ information count.
- That count does not establish the same bound under the stated $M = O(n)$ definition.
- The proof also needs a consistent modulus M.

Our consistency analysis: Definition 2.6, p. 5; Theorem A.9 and proof, pp. 15–16.

The useful distinction is access depth versus state capacity

1. **Core idea.** Preserve indistinguishable input families across both query-dependent replies and recurrent state transfers.
2. **Architectural implication.** Count full-attention opportunities, their placement, and actual recurrent state bits.
3. **Evidence scope.** Exact worst-case tasks and very large asymptotic instances; no measured accuracy or optimal mixing ratio.
4. **Reading judgment.** Keep the main proof route separate from the v1 strengthening and auxiliary statements needing clarification.

Synthesis of the paper and the qualifications on slides 14, 18, and 19.

References and discussion route

- Xiaowei Ye, Xiaoyu He, Chao Liao, Chen Wu, Pinyan Lu.
  A Provable Expressiveness Hierarchy in Hybrid Linear-Full Attention.
  arXiv:2602.01763v1, 2026.
- Lijie Chen, Binghui Peng, Hongxun Wu.
  Theoretical limitations of multi-layer Transformer.
  arXiv:2412.02975v1; cited as FOCS 2025 in the paper.
- Clayton Sanford, Daniel Hsu, Matus Telgarsky.
  Representational Strengths and Limitations of Transformers. NeurIPS 2023; source cited for the full-attention 2-Sum bound.

Main discussion: slides 1–20, approximately 25 minutes. Backup: parameters, invariants, and retrieval construction.

Backup: the task size absorbs recurrent transcript bits

Let $c = Hdp$ and $C = Hd(d+1)p$.

$$K = (cL)^8 8^{2L^2}, \qquad m = K^{1+\sum_{r=0}^{L-1} 8^r},$$

$$n_\ell = K^{4 \cdot 8^{L-\ell-1}}, \qquad x_\ell = K^{8^{L-\ell-1}}.$$

The communication condition is

$$C \sum_{r=1}^{\ell+1} a_r \le \sqrt{K}(x_0 \cdots x_{\ell-1})(n_1 \cdots n_{L-1}).$$

- Fixing s transcript bits costs a factor of at most 2^s.
- Lemma B.1 bounds the constructed prompt size by $c^{4 \cdot 16^L}$.

Paper, Eqs. (5)–(12), pp. 17–18. In Lemma B.3, use $C \le 2c^2$ to repair a dropped factor of d.

Backup: two size bounds are maintained together

$$|R_{\geq \ell}| \geq \frac{|A_L| \cdots |A_\ell|}{\Delta_\ell}, \qquad |\mathcal{I}_{\ell-1}| \geq \Theta_{\ell-1}.$$

$$\Delta_\ell = 2^{4\sqrt{K}(x_0 \cdots x_{\ell-2})(n_1 \cdots n_{L-1})},$$

$$\Theta_\ell = 8^{-L\ell}(x_0 \cdots x_\ell)(n_1 \cdots n_{\ell-1}).$$

- A large R preserves ambiguity about prefix functions.
- A large $\mathcal{I}$ prevents all ambiguity from hiding at unreachable addresses.
- Rectangle extraction lets new local inputs vary independently.

Paper, Eqs. (10)–(11), p. 17; Lemmas B.5–B.9, pp. 19–21.

Backup: a consistent scaling implements the retrieval head

For binary addresses, define $u(a) = (a, \mathbf{1} - a)$ and set

$$q = \sqrt{\lambda}\, u(a), \qquad k_i = \sqrt{\lambda}\, u(a_i), \qquad v_i = b_i.$$

Then

$$q^\top k_i = \lambda(D - \mathrm{Ham}(a, a_i)), \qquad \alpha_* \geq \frac{1}{1 + (n-1)e^{-\lambda}}.$$

- Choose $\lambda = \Theta((\log n)^2)$; the unique match dominates.
- Binary readout can recover the retrieved value at finite precision.
- The v1 text scales both Q, K by $\log^2 n$ but writes a $\log^2 n$ dot-product factor; the formula above normalizes that mismatch.

Teaching rewrite of Appendix C, PDF p. 24. The scaling repair does not change the polylogarithmic capacity claim.