

The Linear-Temporal Limitation

These results apply to Mamba2, Linear Attention, Gated Delta Networks, and every architecture built on linear state-space foundations.

The Architectures in Question

The modern sub-quadratic revolution rests on linear state evolution, enabling parallel processing via associative scan.

Mamba2's core recurrence $h_t = Ah_{t-1} + Bx_t$ unfolds to $h_T = \sum_t A^{T-t} Bx_t$. The matrices A and B may depend on input, but the state is linear in past states. This enables parallel scans and constrains computation.

Linear Attention computes Output = $q \cdot (\sum_i k_i \otimes v_i)$. Gated Delta Networks update as $S' = S + (v - Sk)k^\top$. Despite different motivations, they share the same constraint.

The Threshold Barrier

(Running Threshold 0.1): *No D-layer linear-temporal model computes running threshold. Linear-temporal outputs are continuous; threshold is discontinuous.*

¹

A linear combination of inputs is continuous. Threshold has a jump discontinuity. Adding layers does not help.

This extends to any function with a hard decision boundary: binary classification, exact counting, flip detection. Linear-temporal models can only approximate jumps with smooth transitions.

The Parity Barrier

(Running Parity 0.2): *No linear-temporal model computes $y_t = x_1 \oplus \dots \oplus x_t$. Parity violates the affine identity: $f(0, 0) + f(1, 1) = 0 \neq 2 = f(0, 1) + f(1, 0)$.*

²

Affine functions satisfy $f(a) + f(b) = f(c) + f(d)$ when $a + b = c + d$. Parity does not.

Parity separates AC⁰ from TC⁰ in complexity theory. For linear-temporal models, parity is impossible at any depth.

The Capability Boundary

Task	Why Impossible	D-layer Linear	1-layer E88
Running threshold	Discontinuous	No	Yes
Running parity	Non-affine	No	Yes
FSM simulation	State count	Limited	Full

E88 with a single layer computes all three. These are theorems.

¹Lean formalization: `ExactCounting.lean:344`.

²Lean formalization: `RunningParity.lean:200`.

When Linear Suffices

For many practical tasks, linear suffices.

Most sentences require parsing depth 2–5; complex clauses push to 7–10; the extreme tail reaches 20–25. A 32-layer model with $D = 32$ exceeds most requirements. Linear-temporal scans process sequences in parallel with throughput that sequential recurrence cannot match.

The limitation matters when depth is constrained, when tasks require temporal decisions (counting, parity, state tracking), or when algorithmic reasoning is needed (following procedures, simulating automata).

Linear-temporal models—Mamba2, Linear Attention, GDN—cannot compute threshold, parity, or general state tracking. This is not a bug; it follows from linear temporal dynamics.

How does E88 escape?