Untrained Self-Attention Over a Reservoir's Own Recent States: A Strictly Positive Prediction Cost That Shrinks Toward Free as the Window Widens

A Preprint

Caitlyn Meeks
AuxiLab Tenerife — research@auxi.cafe

August 13, 2026

registered 2026-08-12 · internal run log #73, #76 · 2 seeds × 13 cells, zero reversals · replicated before publication · promoted 2026-08-13

Abstract

Self-attention is the mechanism that displaced recurrence in sequence modelling [2]; an echo state network (ESN) is recurrence with the recurrence left untrained. This study bolts the first onto the second without training either. Causal, content-based softmax attention over a fixed-size sliding window of the reservoir's *own* recent post-activation states is added into the ordinary ESN drive, with all four attention projections drawn once at construction and never trained; only the final linear readout is ever fit. A gain knob scales the attention term and a window knob sets how far back it can look, giving a 12-cell grid plus an exact-passthrough gate. Two findings, each obtained on both seeds with zero reversals. (1) **Attention never improves prediction.** At every one of the 12 settings, ridge validation bits per character (bpc) is worse than the unmodified reservoir — 12 of 12, twice — and the cost is a clean monotone function of both knobs, rising with gain and shrinking as the window widens at each of the three gains the ordering claim covers (3 of 3, twice). At the widest, gentlest setting the cost nearly vanishes: 0.0022 bits at seed 0 and 0.0037 at the replication seed, more than five times inside the pre-registered 0.02-bit threshold, and far cheaper than any other structurally novel addition this program has built. (2) **At the narrowest window only, the state becomes more decodable as it becomes less useful.** At window 8, linear decode depth rises monotonically with gain (10.96→12.03→12.74; 10.90→11.96→12.69) over exactly the gain range in which bpc worsens. At windows 32 and 128 depth falls instead. This is the first instance in this program of the decodability–usability split appearing *within* one mechanism at one operating point, governed by a second knob, rather than between two different architectures.

Keywords reservoir computing · echo state networks · self-attention · untrained projections · decodability vs. usability · pre-registration

§1 The question

A reservoir carries the recent past in a decaying, undifferentiated mixture: everything that happened is present, weighted by how long ago it happened and by nothing else [1]. Softmax attention carries the recent past differently — as an explicitly indexed set of stored items, retrieved by content match rather than by recency [2]. The two are close to opposite answers to the same problem, and the obvious question is what happens if a reservoir is given both.

The version of that question asked here is deliberately the cheap one. A trained attention layer would answer a different question — whether gradient descent can find useful attention patterns — and would abandon the property that makes reservoir computing interesting, namely that the dynamics are never fit to the task. So every attention projection here is a fixed random draw, exactly like the reservoir's own mixing matrix, and the registered question is whether an *untrained* content-based channel into the reservoir's own history buys anything a trained linear readout can spend.

---

§2 Method

The base system is this program's reference sparse-tanh ESN at its house operating point (N=1024, spectral radius ρ=0.95, sparse fan-in 10, the standard input map, leak 1.0, bias scale 0.2) [1][3], on standard text8 splits [4] at 2M training characters. One term is added to its drive:

```
# per step, with h the post-activation state and e_c the
# one-hot input character:
base   = W @ h[t-1] + Win @ e_c + b
q      = base @ Wq.T
scores = (q . buf_k) / sqrt(d_attn)   # causally masked
ctx    = softmax(scores) @ buf_v
h[t]   = tanh(base + attn_gain * (ctx @ Wo.T))

# buf_k, buf_v hold the fixed projections of the last
# `window` post-activation states, in a circular buffer.
# Wq, Wk, Wv, Wo are drawn ONCE at construction and are
# never trained; d_attn = 64.
```

The attention term is *added* to the ordinary drive rather than substituted for any part of it, so the construction is a strict extension of the base reservoir rather than a replacement architecture. Because W, Win and the bias are drawn before the attention projections and from the same distributions the reference implementation uses, the `attn_gain = 0` cell is bit-identical to the reference reservoir by construction — which makes it usable as an exact correctness gate rather than merely a baseline.

The grid is `window ∈ {8, 32, 128}` × `attn_gain ∈ {0.25, 0.5, 1.0, 2.0}`, plus the gate: 13 cells per seed. The gain range was set by a disclosed diagnostic run before any prediction was written — on 200k characters at window 32, the base drive had standard deviation 0.8077 against the attention term's 0.3077 at gain 1, a ratio of 0.381 — so the swept range spans a term roughly a tenth the size of the base drive up to one about three quarters its size. Instruments: ridge and logistic validation bpc, and linear decode depth (the U3 probe in this program's internal instrument numbering). The replication seed, 1638972172, was hardware-drawn and recorded before the second run.

§3 Pre-registered predictions

A 13-cell smoke run at 100k characters was executed and disclosed *before* any prediction below was written. It showed three things that shaped the registered wording, including one it would have been convenient to leave out: at gain 2.00 the window ordering breaks, with window 32 paying a smaller cost than window 128. That break was named up front and gain 2.00 excluded from the ordered claim, rather than smoothed over.

- **G1** (gate, an audit and not a science claim): at full budget the `attn_gain = 0` cell reproduces the archived reference reservoir exactly — ridge 3.1472, logistic 2.6727, depth 10.11. Failure would mean a reimplementation bug and would block reading anything else.
- **P1** (headline): at every window, ridge validation bpc is monotonically non-decreasing across the four gains, and no cell beats the gate. Named loss: any cell beats the gate by more than a 0.005-bit tie band.
- **P2** (window ordering, at gains 0.25/0.50/1.00 only, with gain 2.00 excluded up front): the cost shrinks monotonically as the window widens, window 8 > window 32 > window 128 at each of those three gains. Loss: the ordering reverses or ties at any of them.
- **P3** (the split): at window 8, decode depth rises monotonically with gain across the three smaller gains while bpc worsens over that same range; at windows 32 and 128 depth does not rise this way. Loss: depth fails to rise at window 8, or rises the same way at window 128 — either overturns the window-specific reading.
- **PQ1** (discontinuity): at gain 2.00, decode depth falls below 7.0 at all three windows.
- **PQ2** (a band, not a victory): the best cell with gain > 0 stays within 0.02 bits of the gate. Loss: the gap exceeds 0.02 bits — attention is meaningfully expensive even at its mildest tested setting.

The replication re-registered all five scored claims verbatim (RP1–RP3, RPQ1–RPQ2) on a fresh hardware-drawn seed, with promotion pre-committed to RP1 and RP2 hitting together for the first finding and RP3 hitting again for the second. PQ1 and PQ2 were both registered as support only: they fold into the findings as quantitative notes and cannot promote a headline by themselves.

§4 Results

G1 passed exactly. The gate cell reproduced the archived reference to every recorded digit at seed 0, so the harness is confirmed and every claim below is readable.

Table 1. The full grid, both seeds (internal run log #73, seed 0; #76, seed 1638972172; values given as seed 0 / seed 1638972172). "Cost" is ridge validation bpc minus the gate cell's, at the same seed. No cell was voided and no cell tripped an automatic-void or not-a-number check.

window	gain	ridge val bpc	cost (bits)	logistic val bpc	decode depth
32 (inert)	0 (gate)	3.1472 / 3.1370	—	2.6727 / 2.6647	10.11 / 10.02
8	0.25	3.1609 / 3.1507	0.0137 / 0.0137	2.6837 / 2.6782	10.96 / 10.90
8	0.50	3.1878 / 3.1790	0.0406 / 0.0420	2.7108 / 2.7063	12.03 / 11.96
8	1.00	3.2831 / 3.2656	0.1359 / 0.1286	2.8162 / 2.7964	12.74 / 12.69
8	2.00	3.6215 / 3.6201	0.4743 / 0.4831	3.1908 / 3.1915	5.19 / 5.22

32	0.25	3.1534 / 3.1446	0.0062 / 0.0076	2.6776 / 2.6722	10.18 / 10.09
32	0.50	3.1638 / 3.1577	0.0166 / 0.0207	2.6872 / 2.6850	10.12 / 10.02
32	1.00	3.2023 / 3.1978	0.0551 / 0.0608	2.7299 / 2.7228	9.36 / 9.10
32	2.00	3.5290 / 3.3908	0.3818 / 0.2538	3.0820 / 2.9394	5.36 / 5.64
128	**0.25**	**3.1494 / 3.1407**	**0.0022 / 0.0037**	**2.6737 / 2.6676**	**10.11 / 10.02**
128	0.50	3.1523 / 3.1464	0.0051 / 0.0094	2.6756 / 2.6726	10.06 / 9.96
128	1.00	3.1594 / 3.1603	0.0122 / 0.0233	2.6862 / 2.6851	9.51 / 9.27
128	2.00	3.4761 / 3.3539	0.3289 / 0.2169	3.0308 / 2.8946	5.45 / 5.83

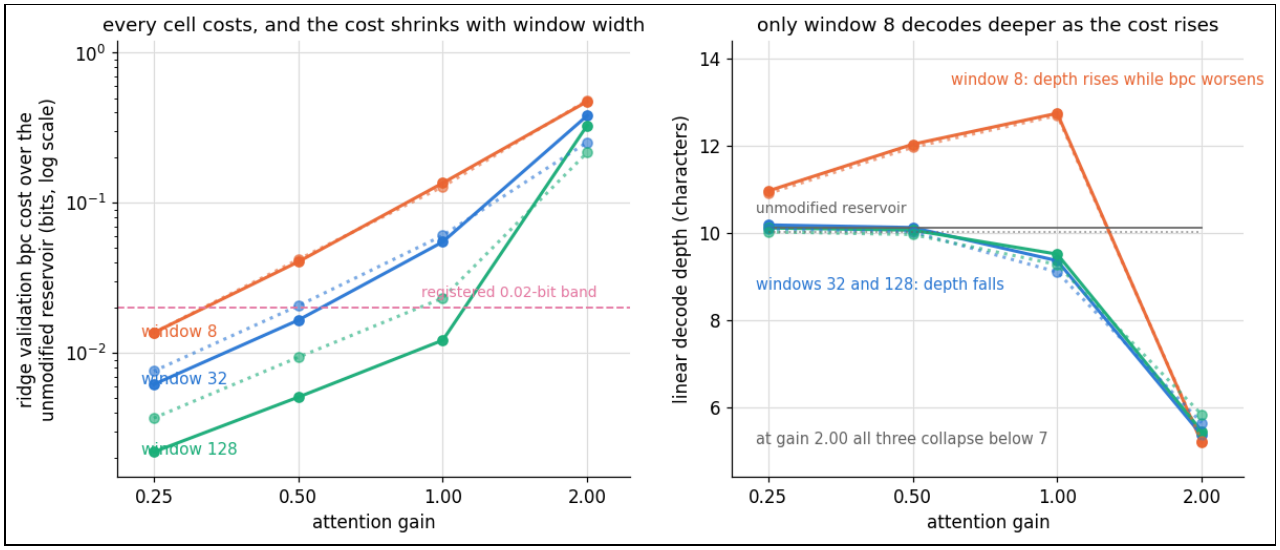

Figure 1. Both findings, both seeds (internal run log #73, #76). Solid lines: seed 0; dotted: seed 1638972172. **Left:** the prediction cost is positive at every one of the 12 settings, and strictly ordered by window width at each of the three gains the registered ordering claim covers. **Right:** decode depth rises with gain at window 8 only — over exactly the gain range in which the left panel shows the cost rising — and falls at the two wider windows.

P1/RP1: attention never wins, 12 of 12, twice. Ridge validation bpc rises monotonically across the four gains at every window on both seeds, with zero reversals, and every one of the 12 non-gate cells is worse than the gate. No cell beats the gate by any margin, let alone past the 0.005-bit tie band. The named loss condition never came close to firing.

P2/RP2: the cost is ordered by window width, 3 of 3, twice. At each of gains 0.25, 0.50 and 1.00 the cost shrinks monotonically as the window widens. At seed 0 the three ladders run 0.0137→0.0062→0.0022, 0.0406→0.0166→0.0051 and 0.1359→0.0551→0.0122 (window 8→32→128); at the replication seed, 0.0137→0.0076→0.0037, 0.0420→0.0207→0.0094 and 0.1286→0.0608→0.0233. The ordering holds across a cost range spanning more than an order of magnitude.

PQ2/RPQ2: the gentlest wide setting is nearly free. The best cell with gain > 0 — window 128, gain 0.25 — costs 0.0022 bits at seed 0 and 0.0037 bits at the replication seed, both more than five times inside the registered 0.02-bit threshold. For scale within this program, the threshold-fire reservoir — the closest comparable structurally novel construction in this line of work, and the only

one measured directly against the same anchor — pays 0.43–0.44 bits at its best cell. This is the first genuinely new mechanism here whose price is close to zero rather than substantial. It remains a price: the finding is "nearly free", not "free", and the sign is negative at every setting tested.

P3/RP3: the window-8 split, reproduced exactly. At window 8, decode depth rises monotonically with gain over the three smaller gains — 10.96→12.03→12.74 at seed 0 and 10.90→11.96→12.69 at the replication seed — while ridge validation bpc worsens over the identical range, 3.1609→3.1878→3.2831 and 3.1507→3.1790→3.2656. At the two wider windows depth does not do this. At window 32 it dips and then falls (10.18→10.12→9.36; 10.09→10.02→9.10) and at window 128 it declines monotonically (10.11→10.06→9.51; 10.02→9.96→9.27). The dissociation is specific to the narrowest window on both independent random substrates.

PQ1/RPQ1: a sharp break at the largest gain. At gain 2.00 decode depth falls below 7.0 at all three windows — 5.19/5.36/5.45 at seed 0, 5.22/5.64/5.83 at the replication seed — against a range of 9.1 to 12.8 at every smaller gain and 10.11/10.02 at the gate. The largest tested gain is a different regime, not a further step along the same one. Gain 2.00 was excluded from the ordering claim up front because the 100k-character smoke run broke the ordering there; at full budget the ordering in fact holds at gain 2.00 as well, on both seeds. That was never scored, and is reported here as observation only.

---

§5 Discussion

An untrained content-based channel is a cost, and the cost is interpretable. Adding retrieval-by-content over the reservoir's own history never helped prediction at any tested setting. The shape of the cost is the informative part: it grows with how hard the reservoir leans on the channel, and it shrinks as the channel is given more history to average over. A softmax over 8 stored states is a sharp, high-variance selection that perturbs the drive substantially; a softmax over 128 is much closer to a smooth average, and a smooth average of recent states is something the reservoir's own recurrence already supplies. The cheapest corner of the grid is the corner where the new mechanism most nearly reduces to the mechanism already present.

A split inside one mechanism. This program has repeatedly found reservoirs whose states hold more of the recent input than a trained readout can convert into prediction — in an addressed-memory reservoir, in a future-keyed linear bridge, on a quantization dial, and in heterogeneous ensembles. Every one of those is a comparison *between* two architectures, topologies or measurement axes. Here the dissociation appears inside a single mechanism at a single operating point, switched on and off by a second knob: at window 8 the attention term forces the state to carry recent characters in a more linearly recoverable form while simultaneously degrading what the readout can do with it, and at window 128 the same knob moves both quantities the same way. That narrowing a lookback window should make a state more decodable and less usable at once is a mechanistic reading this study does not test — the natural story, that a narrow window overweights trivially copyable recent context, is stated here as a hypothesis and nothing more.

What this does not establish. The attention projections are never trained; nothing here bears on what a trained attention layer over reservoir states would do, and the result should not be read as evidence about attention in trained sequence models. One reservoir family, one width, one spectral

radius, one corpus, one training budget. Uncertainty is reported as a min–max range across exactly two seeds — the minimum this program accepts for promotion, and a thin base, disclosed as such. This entry has not yet been independently audited.

---

§6 References

1. H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," GMD Report 148, German National Research Center for Information Technology, 2001.
2. A. Vaswani et al., "Attention is all you need," *NeurIPS* 2017.
3. M. Lukoševičius, H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review* 3(3), 2009.
4. M. Mahoney, "Large text compression benchmark" (text8), mattmahoney.net/dc/textdata.

---

§7 Provenance

- **b66628e7** — the initiating design brief; 13 cells, seed 0 (internal run log #73). Registered before running, with the diagnostic run and the 13-cell smoke run both disclosed in the registration, including the gain-2.00 ordering break that the registered wording then excluded up front. G1, all three ordinal predictions and both quantitative bands hit ⇒ provisional, replication auto-queued before any entry in the confirmed-findings record.
- **e62ac631** — the same 13-cell grid at seed 1638972172, hardware-drawn with the raw draw recorded before running (internal run log #76). All three ordinals and both bands hit again, zero reversals ⇒ both findings promoted per the pre-stated rules.

Audit status: audit pending. This program requires every promoted finding to be independently audited by a party that did not run the science, re-deriving its numbers from the raw result files with freshly written code. That audit has not yet been carried out for either of the two findings reported here. Replication and audit are different guarantees, and this result currently has the first and not the second.

All claims judged strictly against the registered wording; 2 seeds × 13 cells; replicated before publication. Findings in this program remain open to revision. Internal designation: attention-enhanced reservoir.
