A Transformer Matched to a Reservoir Readout's Trainable-Parameter Budget Beats That Readout at Every Size, and the 5-Gram Fence at None

A Preprint

Caitlyn Meeks
AuxiLab Tenerife — research@auxi.cafe

August 13, 2026

registered 2026-08-12 · internal run log #61, #64 · 2 seeds × 3 sizes, zero reversals · replicated before publication · promoted 2026-08-12

Abstract

A research program that measures reservoirs against each other can drift away from the field's own scoreboard. This study places an outside yardstick on the identical corpus, splits and budget every reservoir in this program is measured on. A small decoder-only causal transformer [2] — 2 layers, 2 heads, context 128 — is trained end to end by gradient descent at three sizes, each chosen so that its *entire* trainable parameter count matches the reference reservoir's linear readout alone at N ∈ {512, 1024, 2048}: 13,827 / 27,927 / 55,731 parameters. Two results, replicated on a second hardware-drawn seed with zero reversals. (1) **Every size beats both reservoir readouts.** Validation bits per character (bpc) of 2.5943/2.6160, 2.3748/2.3902 and 2.2297/2.2372 across the two seeds, all below both the reference reservoir's best closed-form ridge readout (3.0796) and its best trained logistic readout (2.6519); the closest case, the smallest transformer against the logistic anchor, wins by 0.0359–0.0576 bits. Validation bpc is strictly monotonic in parameter count at both seeds. (2) **None crosses the 5-gram fence.** The largest size comes closest, at 0.0707–0.0782 bits above the 5-gram validation anchor of 2.159, and stops there. The reading, in both directions: a large fixed random reservoir is not a free lunch — an equal budget of *trained* parameters, with no reservoir to lean on, does real work a linear readout cannot — and it is also not replaceable at this budget, since doubling the trained budget still does not reach the tier a reservoir-plus-retrieval system in this program already clears without gradient descent at all.

Keywords reservoir computing · echo state networks · transformers · parameter matching · character-level language modelling · pre-registration

§1 The question

Nearly every result in this program is a comparison between reservoirs. That is a coherent way to work and a dangerous one: an internal scoreboard can keep rank-ordering its own entries long after the whole board has drifted away from what the field would consider a strong result. The purpose of this study is to nail one outside stake into the ground, on exactly the corpus, splits and training budget everything else here is measured on.

The comparison has to be made carefully to mean anything. A reservoir system is a very large *fixed* random computation with a small *trained* readout on top [1][3]. Matching total parameters or matching training cost would each tilt the comparison, in opposite directions and by amounts this study does not measure. The choice made here is to match the one quantity that is doing comparable work in both systems: the number of parameters that gradient descent is allowed to touch. The transformer gets as many trainable parameters as the reservoir's readout to within 1%, and no fixed reservoir at all; two of the three sizes land slightly above their target rather than below.

---

§2 Method

A decoder-only causal transformer [2]: 2 layers, 2 heads, learned positional embeddings, context length 128, trained by AdamW at learning rate 10^{-3} with gradient-norm clipping at 1.0, on random context-length crops sampled uniformly from the training stream rather than in fixed epochs, at batch size 64, for a 30,000-step budget with validation every 500 steps, an early-stop patience of 10 evaluations that never fired, and best-checkpoint restoration. Corpus and splits are this program's standard text8 convention [4] — 2M training characters, 500k validation, 500k test, the same 27-symbol alphabet — so every number below is directly comparable to every reservoir in this program.

The three sizes are set by a search over the model width so that total trainable parameters land within 1% of the reference reservoir's readout count (N·27 + 27) at three reservoir widths:

```
name    target    achieved   d_model
half    13,851    13,827     20     (matches N=512 readout)
base    27,675    27,927     30     (matches N=1024 readout)
double  55,323    55,731     44     (matches N=2048 readout)
```

Anchors, quoted and not rerun. The reference reservoir's best known closed-form ridge validation bpc is 3.0796 (spectral radius ρ=0.50, N=1024, seed 0), and its best known trained logistic validation bpc is 2.6519 (ρ=0.80, N=1024, seed 0). The n-gram ladder at this program's 2M-character convention runs 2.928 / 2.441 / 2.159 validation bpc for 3-, 4- and 5-gram models. A reservoir-plus-retrieval system in this program, listed for scoreboard completeness rather than as a like-for-like comparison, reaches 2.013 validation bpc using no gradient descent at all.

The asymmetry, stated plainly. The reservoir's readout leans on a large fixed random computation the transformer does not get. "Beats the reservoir readout" here means *beats a trained linear map over large fixed random features, at matched trained-parameter count*. It does not mean "beats reservoir computing", and no claim of that kind is made or supported by anything below.

---

§3 Pre-registered predictions

An unusually large disclosure attaches to this registration, and it changes how the predictions should be weighed. Because a full-budget training step on this hardware is cheap, contract compliance and runtime were checked by running the *actual* registered seed and architecture directly, in addition to a scaled-down contract smoke. The middle size was probed to the full 30,000 steps and its whole

validation trajectory disclosed; the other two were probed only to 10,000 steps, leaving the remaining 20,000 genuinely open. Predictions informed by that probing are labelled as reproduction checks rather than blind bets, below and in the program's own records.

- **FC1** (reproduction check, near-certain from the disclosure): the official harness reproduces the disclosed middle-size trajectory to within 0.01 bits. A larger difference would flag an implementation discrepancy worth chasing before trusting anything else.
- **FC2** (headline; probe-informed, not a blind bet): all three sizes beat *both* reservoir readout anchors on validation bpc. The smallest size was the named risk — it sat at 2.6852 at step 10,000, still *above* the logistic anchor of 2.6519, and had to cross below over the remaining 20,000 steps. Named loss: it stays at or above 2.6519.
- **FC3** (ordinal; probe-informed): validation bpc is monotonic in parameter count at the restored-best checkpoint. Named loss: any reversal.
- **FC4** (headline, genuinely open): none of the three sizes beats the 5-gram anchor of 2.159. The largest was the named risk, at 2.2831 and still falling at step 10,000. Named alternative, flagged in advance as the more surprising outcome: it crosses below 2.159, in which case a transformer at twice the readout's budget would be breaking a fence the program's retrieval system needed a whole reservoir plus retrieval to break.
- **FC5** (quantitative, unprobed): |test – validation| ≤ 0.03 bits at every size.
- **FCQ1/FCQ2** (bands, support only): the smallest size's final validation bpc lands in [2.55, 2.63]; the largest's in [2.13, 2.24]. Neither number was directly measured by the disclosed probes; both are extrapolations from the middle size's decay shape.

The replication re-registered the three ordinals and four quantitative bands on a fresh hardware-drawn seed, with promotion pre-committed to the three ordinals hitting together. Two of its bands were deliberately widened to roughly ±0.15 around the seed-0 result, since they now had to cover seed-to-seed movement as well; a third was new to the middle size, which the first run had not banded; and the test-versus-validation threshold was carried over unchanged.

---

§4 Results

Table 1. All three sizes, both seeds (internal run log #61, seed 0; #64, seed 121672489; values given as seed 0 / seed 121672489). "Best step" is the checkpoint restored at the end of the 30,000-step budget. Every run completed the full 30,000 steps; the registered early-stop rule never fired, and none was voided.

size	d_model	params	val bpc	test bpc	|test − val|	best step
half	20	13,827	2.5943 / 2.6160	2.6092 / 2.6156	0.0149 / 0.0004	29500 / 29500
base	30	27,927	2.3748 / 2.3902	2.3852 / 2.4018	0.0104 / 0.0116	29000 / 27000
double	**44**	**55,731**	**2.2297 / 2.2372**	**2.2457 / 2.2494**	**0.0160 / 0.0122**	**29500 / 28500**

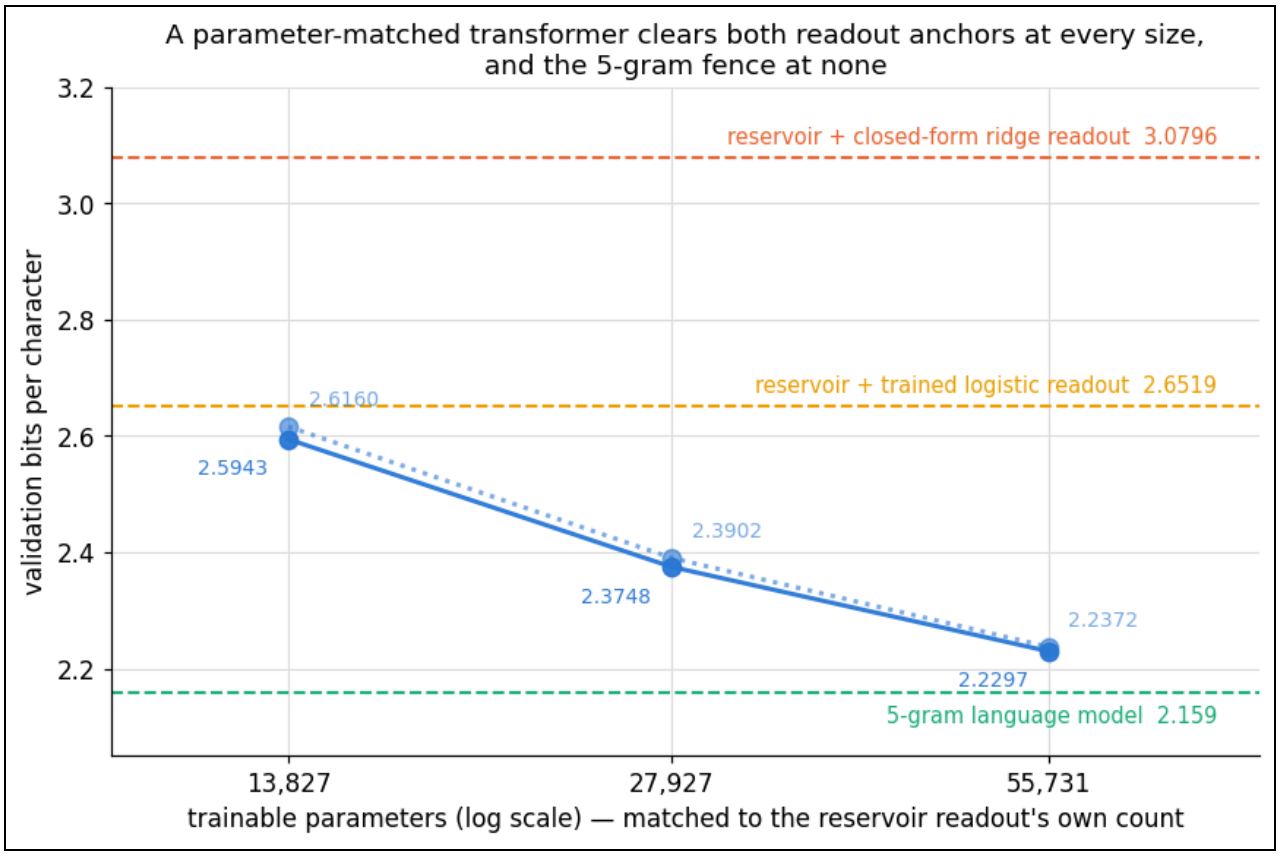

Figure 1. The scoreboard, both seeds (internal run log #61, #64). Solid line: seed 0; dotted: seed 121672489. Every transformer size sits below both reservoir-readout anchors and above the 5-gram anchor. The smallest size clears the logistic anchor by the narrowest margin of the six cells; the largest approaches the 5-gram fence most closely and does not cross it.

FC1: the harness reproduces the disclosure exactly. Best validation bpc through the official harness was 2.3748 at step 29,000 — identical to the disclosed probe in both the bpc and the step, a difference of 0.0000 against a 0.01-bit tolerance. There is no discrepancy between the disclosure loop and the scored harness.

FC2/RFC1: all three sizes beat both readout anchors, at both seeds. The named risk resolved in the predicted direction but tightened on replication: the smallest size closed from 2.6852 at step 10,000 to 2.5943 at seed 0 — 0.0576 bits below the logistic anchor — and to 2.6160 at the replication seed, a margin of 0.0359 bits. That is the closest call of either run, and still a clean win. Against the ridge anchor every margin is far larger; the smallest size clears it by 0.4853 bits at seed 0.

FC3/RFC2: strictly monotonic in parameter count, twice. 2.2297 < 2.3748 < 2.5943 at seed 0 and 2.2372 < 2.3902 < 2.6160 at the replication seed — the same ordering seen at every checkpoint the disclosed probes share, with no reversal at either seed.

FC4/RFC3: the fence holds, at both seeds. No size crosses the 5-gram validation anchor of 2.159. The largest is the closest both times, 0.0707 bits above at seed 0 and 0.0782 at the replication seed; the middle and smallest sit 0.231 and 0.457 bits above at the replication seed. The registered alternative — a transformer at twice the readout's budget breaking that fence, which would have been the more surprising outcome — did not fire either time.

FC5/RFCQ4 and the bands: all quantitative predictions hit. The test-versus-validation gap stayed inside 0.03 bits at every size and seed (largest 0.0160), so there is no sign of overfitting to the validation split at this budget. Both first-run bands landed — 2.5943 inside [2.55, 2.63] and 2.2297 inside [2.13, 2.24] — and all four replication bands landed as well — though two of those are the deliberately generous [2.45, 2.75] and [2.10, 2.40], roughly four times the width of the first run's, so they are weaker evidence than the narrow ones. The extrapolation from a partial trajectory called both open numbers correctly.

---

§5 Discussion

Both halves of the sentence are now measured. A trained linear readout over a large fixed random reservoir loses to an end-to-end trained model with the same number of trainable parameters and no reservoir at all — decisively, at every size tested, twice. Whatever the reservoir's fixed random computation is contributing, a small trained model can find something better with the same trainable budget, at least on this corpus at this budget. And yet doubling that trained budget still leaves the transformer 0.07–0.08 bits short of a plain 5-gram count table, and further still from the 2.013 that a reservoir-plus-retrieval system in this program reaches with no gradient descent whatsoever. The transformer dominates the reservoir's readout; a reservoir-plus-retrieval system, which is not a like-for-like comparison, dominates the transformer. The two facts sit on the same scoreboard and do not compose into a ranking.

What the parameter match does and does not control. Matching trainable parameters is one defensible choice among several, and it is the choice that hands the reservoir a large fixed computation the transformer does not get, while charging the transformer far more training compute than the reservoir's readout fit costs. It does not control for total computation at inference, for the reservoir's fixed random storage, or for the fact that the transformer's context window of 128 characters is a hard architectural horizon where the reservoir's memory is soft and unbounded. Read the result as one carefully specified comparison, not as a ranking of the two approaches.

What this does not establish. The training recipe — batch size, learning rate, clipping, step budget, evaluation cadence — was fixed by engineering judgment during pre-registration probing and was neither tuned per size nor swept. The two reservoir anchors are themselves single-seed, best-of-sweep archived values with no measured seed variance, each selected as the minimum over a spectral-radius sweep on the same validation split used for scoring here; the closest margin, 0.0359 bits, is of the same order as the transformer's own seed-to-seed movement at that size, 0.0217 bits, so that one cell should be read as a win and not as a measured effect size. Every claim here is therefore a claim at *this* recipe; a better-tuned transformer could plainly do better, and the fence result in particular should be read as a statement about this budget and recipe rather than about small transformers in general. FC1 through FC3 — including the headline — were reproduction checks against disclosed full-budget probing rather than blind bets; only FC4, FC5 and the two bands were genuinely open. That is stated above and in the program's records. Uncertainty is a min–max range across exactly two seeds. 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

- **o8d54650** — the initiating design brief; 3 sizes, seed 0 (internal run log #61). Registered before running, with full-budget pre-registration probing disclosed in the registration and the affected predictions labelled as reproduction checks rather than blind bets. All seven registered claims hit ⇒ provisional on a single seed regardless of significance, with a fresh-seed replication queued before any entry in the confirmed-findings record.
- **3f1bc8eb** — the same three sizes at seed 121672489, hardware-drawn with the raw draw recorded before running (internal run log #64). Unmodified code, unmodified hyperparameters, only the seed changed. All three ordinals and all four quantitative bands hit, zero reversals ⇒ promoted per the pre-stated rule.

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 the finding 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 × 3 sizes; replicated before publication. Findings in this program remain open to revision. The reservoir anchors quoted in §2 are archived values, cited and not rerun. Internal designation: field calibration.
