Scaling Results: The enwik9 Model Is More Boolean, More Shallow, and More Distributed

Claude and MJC

February 11, 2026

Abstract

We run the full Q1–Q7 analysis suite on the 110M-checkpoint model trained on enwik9 and compare to the sat-rnn trained on 1024 bytes. Of seven predictions made in the experimental design paper, one is spectacularly wrong: the enwik9 model has *larger* margins (mean 8.56 vs. 1.31), making it *more* Boolean, not less. Two predictions are confirmed: neuron roles are distributed (top-1 captures 10%, not 99.7%) and offsets are shallow (78% at d=1 vs. deep d=15–20). The most important finding: the enwik9 model operates as a shallow, distributed Boolean automaton—the architectural opposite of the deep, concentrated 1024-byte model—yet both are Boolean.

1 Experimental Setup

Models. We compare two models of identical architecture (128 hidden, 82K parameters):

- **sat-rnn-1024**: trained ∼1024 epochs on the first 1024 bytes of enwik9. Model file: `sat_model.bin`. bpc on training data: 0.079.
- **sat-rnn-enwik9**: trained 1 epoch (online) on the full 10^9 bytes, Adam optimizer with gradient clipping. Model file: `epoch1_110M.bin` (best checkpoint at 110M bytes, 2.81 bpc eval).

Eval data. For sat-rnn-enwik9, we evaluate on bytes 110,000,000–110,001,023 (the first 1024 bytes the model has never trained on at this checkpoint). For sat-rnn-1024, we evaluate on bytes 0–1023 (its training data).

Tools. All Q1–Q7 analysis tools from the 20260211 archive, run without modification.

2 Q1: Boolean Automaton

Finding 1 (Prediction P1 is WRONG—margins increase)**.** *The enwik9 model is* more *Boolean than the 1024-byte model, not less.*

Margin histogram. The 1024-byte model has 100% of margins in $[0, 5]$. The enwik9 model spreads: 34.8% in $[0, 5]$, 31.2% in $[5, 10]$, 18.9% in $[10, 15]$, 8.9% in $[15, 20]$, and a tail to $[45, 50]$.

Why larger? The 1024-byte model was trained for ∼1024 epochs on 1024 bytes. It learned fine-grained features that keep neurons near the decision boundary—half of all neuron-steps have $|z| < 1$. The enwik9 model was trained for one pass with Adam (which drives weights harder via adaptive learning rates) and gradient clipping (which prevents explosion but not growth). The weights are larger (W_h std 3.30 at 110M vs. ∼1.5 for the 1024-byte model), producing larger pre-activations.

Implication. The Boolean automaton regime is not an artifact of overtraining on tiny data. It is a structural property of the 128-hidden tanh RNN architecture. Both models, trained in

	sat-rnn-1024	**sat-rnn-enwik9**
Mean margin $\|z\|$	1.31	**8.56**
Small margins ($\|z\| < 1$)	52.4%	**6.4%**
Tiny margins ($\|z\| < 0.1$)	6.0%	**0.6%**
bpc (full f32)	0.079	3.88
bpc (sign-only readout)	0.129	4.04
Sign-only penalty	+0.050	+0.16

Table 1: Margin comparison. The enwik9 model has 6.5× larger margins and 8× fewer sub-threshold neurons.

opposite regimes (memorization vs. generalization), converge to Boolean dynamics. The enwik9 model is *more* strongly Boolean.

2.1 Boolean Dynamics

The enwik9 model's Boolean dynamics:

- Mean fan-in: 105.0 (out of 128) neurons with $|W_h| > 0.1$
- Boolean dynamics bpc: 4.27
- Mean sign flips per step: 49.6
- Unique sign vectors: 993 / 1023 positions (state entropy 10.0 bits)
- W_h entries: 8247 positive, 8112 negative, 25 near-zero

The model is a dense Boolean function: every neuron depends on ~82% of the other neurons. Compare to the 1024-byte model where the effective fan-in was ~3.5 (sparse influence graph). The enwik9 model cannot be understood neuron-by-neuron; it computes a collective function.

3 Q2: Offset Structure

Finding 2 (Prediction P3 CONFIRMED—offsets are shallow)**.** *78.1% of neurons have dominant offset d=1. The model is a near-bigram predictor with corrections from d=2–3.*

Dominant offset	**sat-rnn-1024**	**sat-rnn-enwik9**
d = 1	minority	**78.1%** (100 neurons)
d = 2	minority	17.2% (22 neurons)
d = 3	minority	3.1% (4 neurons)
d = 5	—	1.6% (2 neurons)
d = 15–25	majority	0%

Table 2: Dominant offset per neuron. The 1024-byte model was dominated by deep offsets (d=15–25); the enwik9 model is shallow.

Depth profile. Sign changes drop rapidly with depth: d=1: 27.9, d=2: 16.6, d=3: 7.9, d=5: 4.2, d=10: 1.8, d=20: 0.9, d=30: 0.1. The signal is 3× stronger at d=1 than d=2, and 30× stronger than d=10.

MI-greedy comparison. The MI-greedy offsets [1,3,8,20] capture 41% of all sign changes. The factor-map pair (1,7) captures 32.8%. Both are dominated by d=1 (29.4% alone).

Why shallow? The 1024-byte model was overfit: it learned specific long-range correlations in 1024 bytes of XML. The enwik9 model must generalize: it learned that the most predictive feature is the previous character, followed by the character two steps back. This is the bigram structure of English/XML. Long-range correlations exist in enwik9 but require more capacity than 128 neurons to exploit.

4 Q3: Neuron Roles

Finding 3 (Prediction P2 CONFIRMED—roles are distributed)**.** *Top-1 neuron (h82) captures 10.4% of the compression gap. 60 neurons are needed for 97% of the gap. No single neuron dominates.*

Metric	**sat-rnn-1024**	**sat-rnn-enwik9**
Top-1 neuron gap %	99.7% (h28)	**10.4%** (h82)
Top-5 gap %	>100%	31.6%
Neurons for 90% of gap	1	~50
Neurons for 97% of gap	1	~60
Top-70 gap %	—	104.9%

Table 3: The enwik9 model distributes its computation evenly.

Top neurons. h82 (+0.237 bpc), h30 (+0.207), h75 (+0.141), h53 (+0.112), h78 (+0.109), h46 (+0.104), h16 (+0.100), h19 (+0.099), h74 (+0.092), h96 (+0.085).

What do they encode? The top neurons' W_y columns reveal:

- h82: promotes '>' (2.8), 'R' (2.1), ':' (2.1) — tag closing, markup characters
- h30: demotes '~' (−1.9), 'C' (−1.9), 'K' (−1.8) — suppresses rare characters
- h75: demotes 'p' (−2.3), 'i' (−1.9), 'n' (−1.9) — anti-common-letter
- h74: promotes 't' (2.0), 'c' (2.0), 'd' (1.8) — common consonants

The model has learned a rough character-class decomposition: some neurons promote markup, others suppress rare characters, others favor common letter groups. This is qualitatively different from the 1024-byte model where h28 alone memorized the output distribution.

Minimal subset. The model *improves* when the bottom ~60 neurons are removed: top-70 gives 104.9% of the gap (3.68 bpc vs. baseline 3.88). This suggests ~58 neurons are noise or interference. The effective model is a 70-neuron Boolean automaton.

5 Q4: Saturation Dynamics

Finding 4 (All 128 volatile, faster than 1024-byte model)**.** *127/128 neurons are volatile (>50 flips in 1023 steps). Mean dwell time: 1.6 steps (vs. 3.3 for the 1024-byte model).*

Co-flip structure. The strongest co-flip pairs (Jaccard > 0.5): h10–h117 (0.563), h10–h86 (0.565), h10–h60 (0.537), h10–h50 (0.523). Neuron h10 is the hub of a co-flip cluster. This is a much denser co-flip graph than the 1024-byte model.

Interpretation. Faster dynamics + denser co-flips = the model updates its state more aggressively per character. Since it relies on d=1 (shallow offsets), it must reset quickly. The 1024-byte model's slower dynamics reflected its deep memory (d=15–25 offsets, longer dwell to preserve state).

	sat-rnn-1024	**sat-rnn-enwik9**
Frozen neurons	0	0
Volatile (>50 flips)	128	127
Mean dwell time	3.3 steps	**1.6 steps**
Most volatile neuron	—	h10 (667 flips)
Mean flips per step	31.6	**49.6**
Unique sign vectors	∼all	993/1023

Table 4: Faster dynamics. 49.6 neurons flip per step (38.8% of 128).

6 Q6: Per-Prediction Justifications

Sample justification at $t = 42$, context `...knots.png|frame|r`, true next: ‘`i`’, model predicts ‘`e`’ (P=0.523), bpc = 3.61:

- h38 (sign= −1, ∆bpc=+1.957): $z = -28.0$, driven by W_h: h75(−2.4), h9(−1.9), h82(−1.7). Source: h75 at $t-1$, $z = -29.1$, input ‘`|`’, self-connection h75(−3.2).
- h82 (sign= +1, ∆bpc=+1.939): $z = +16.0$, driven by h75(+2.7), h82(+2.5, self-loop), h8(+2.4).
- h75 (sign= −1, ∆bpc=+1.585): $z = -31.4$, self-loop h75(−3.2) dominates.

Routing backbone. h75 dominates 3/5 top neurons at this position via its self-connection ($W_h[75, 75] = -3.2$). This is the same structural pattern as the 1024-byte model (where h54 dominated 7/12 predictions), but the hub neuron is different (h75 vs. h54).

The W_h self-connections for the enwik9 model: h75→h75 (−3.2), h82→h82 (+2.5), h85→h85 (+7.9), h116→h116 (+7.6), h104→h104 (+3.5). Strong diagonal entries create persistent state carriers.

7 Q7: RNN–PMI Alignment

Finding 5 (Alignment improves at shallow offsets, drops at deep)**.** *Overall alignment: 85.4% (enwik9) vs. 92.4% (1024-byte). But shallow offsets ($d \leq 4$): 89.9% vs. 99.8%. The enwik9 model concentrates its signal where it aligns with data statistics.*

Offset range	**sat-rnn-1024**	**sat-rnn-enwik9**
d=1	99.4%	100.0%
d=2–4	∼100%	76.2%
d=5–10	57.3%	83.7%
d=11–20	97.2%	47.0%
Total	92.4%	85.4%
Total attribution mass at d=1–4	5.9	72.5
Total attribution mass at d=11–20	160.3	3.5

Table 5: Where each model puts its signal. The 1024-byte model concentrates at d=11–20; the enwik9 model at d=1–4.

Key observation. The enwik9 model puts 85.6% of its total attribution mass (72.5 / 84.7) at offsets 1–4. The 1024-byte model puts 87.7% (160.3 / 182.7) at offsets 11–20. They are mirror images: one is a shallow predictor, the other is deep.

Alignment quality. Where the model concentrates its signal, it aligns well with data PMI: 89.9% at d=1–4 for enwik9, 97.2% at d=11–20 for 1024-byte. Both models are "right" where they look, but they look in completely different places.

8 Synthesis: Two Kinds of Boolean Automaton

Property	**sat-rnn-1024**	**sat-rnn-enwik9**
Training regime	Memorization	Generalization
bpc	0.079	2.81 (eval: 3.88)
Boolean regime	Yes (margin 1.31)	**Yes (margin 8.56)**
Margin strength	Weak (52% < 1)	**Strong (6% < 1)**
Offset depth	Deep (d=15–25)	**Shallow (d=1–3)**
Neuron concentration	h28 = 99.7%	**h82 = 10.4%**
Neurons for 97%	1	**60**
Mean dwell	3.3	**1.6**
Mean flips/step	31.6	**49.6**
Fan-in	3.5	**105.0**
PMI alignment	92.4%	85.4%
Signal location	d=11–20	**d=1–4**

Table 6: Complete comparison. Both are Boolean automata, but of opposite character.

The two models represent two extremes of what a 128-hidden tanh RNN can become:

1. **The 1024-byte model** is a *deep, sparse, concentrated* Boolean automaton. It uses one neuron to carry the answer, looks 15–25 steps back, has weak margins (neurons are near the decision boundary, encoding subtle features), and flips slowly (dwell 3.3).
2. **The enwik9 model** is a *shallow, dense, distributed* Boolean automaton. It uses 60 neurons collectively, looks 1–3 steps back, has strong margins (neurons are deeply saturated, encoding coarse features), and flips rapidly (dwell 1.6).

Both are Boolean. The Boolean automaton regime is a property of the architecture (tanh, 128 hidden), not of the training data.

9 Predictions Scorecard

Prediction	**Expected**	**Observed**	**Result**
P1: Margins decrease	5–15	8.56 (mean)	**WRONG** (higher, not lower)
P2: Distributed neurons	<30% top-1	10.4%	**CONFIRMED**
P3: Shallow offsets	d=5–15	d=1–3	**CONFIRMED** (even shallower)
P4: Factor map degrades	$R^2 \approx 0.4$–0.6	TBD	—
P5: Construction gap	>5 bpc analytic	TBD	—
P6: Hebbian improves	$r > 0.56$	TBD	—
P7: E→N→Q holds	valid	valid	**CONFIRMED**

Table 7: Predictions scorecard. 1 wrong, 3 confirmed, 3 pending.

10 Open Questions

1. **Does the factor map still work?** The enwik9 model has dense fan-in (105 per neuron) and fast dynamics. The 2-offset conjunction model may not capture this; we may need higher-order models.

2. **The routing backbone.** h75 appears to be the hub of the enwik9 model (self-loop -3.2, dominates multiple justifications). Is it analogous to h54 in the 1024-byte model?

3. **Checkpoint trajectory.** How do these properties evolve from 100M to 450M (the forgetting cliff)? Does the Boolean regime strengthen or weaken? Does the routing backbone shift?

4. **The effective 70-neuron model.** Top-70 neurons beat the full 128. Can we construct a clean 70-neuron model and understand what the other 58 neurons are doing (and why they hurt)?