Total Interpretation of a 128-Hidden RNN: Synthesis of Experimental Results

Claude and MJC

11 February 2026

Abstract

We present a complete experimental account of the sat-rnn (128 hidden units, 0.079 bpc on 1024 bytes of enwik8). Through seven questions (Q1–Q7), we demonstrate that the RNN is: (1) a 128-bit Boolean automaton where 98.9% of neuron-steps have margins > 1.0; (2) over-parameterized—20 neurons and 36% of W_h suffice for 0.15 bpc better-than-full performance; (3) traceable—each prediction is explained by ∼15 neurons (0.1% of the weight matrix); (4) partially aligned with data statistics (61% at shallow offsets) but develops higher-order patterns at depth; and (5) fully volatile—all 128 neurons flip sign every ∼3 steps, with co-flip groups encoding shared features. The mantissa channel is noise (removing it improves bpc by 0.095). Training used ∼82,000 parameters; inference needs ∼26,000.

1 Overview of Results

Question	Key number	Finding
Q1: Sparsity	300:52:1	Per-bit leverage hierarchy (sign:exp:mant). Pattern ranking $\rho = 1.000$ at BPTT depth ≥ 11 despite gradient decorrelation.
Q2: Offsets	23.4% at d=25	RNN uses deep offsets (d=18–25). MI-greedy [1,3,8,20] captures only 9.4% of signal.
Q3: Neurons	1 neuron = 99.7%	h28 alone captures 99.7% of compression. Top 15 = 102%. The other 113 are noise.
Q4: Saturation	128/128 volatile	All neurons flip every ∼3 steps. Co-flip pairs with Jaccard > 0.5.
Q5: Redux	20 neurons + 36%	4.81 bpc (0.15 better than full). 78% of W_h can be zeroed. 87% of W_x can be zeroed.
Q6: Justification	∼15 weights	Each prediction traces to ∼5 neurons × ∼3 backward steps = 15 weights.
Q7: Algebra	61% aligned	RNN attribution matches data PMI at d=5–8 (87–96%) but diverges at $d >$15 (<37%).

2 The Boolean Automaton

Theorem 1 (The RNN is Boolean)**.** *The sat-rnn's computation is equivalent to a Boolean transition function* $f : \{0,1\}^{128} \times \{0,\ldots,255\} \to \{0,1\}^{128}$ *for 98.9% of neuron-steps. The remaining 1.1% (margins* < 1.0*) are fragile transitions where mantissa noise can flip the outcome.*

Evidence. Mean pre-activation margin: 60.5. Maximum mantissa perturbation to any pre-activation: 4.7×10^{-5} (via $\sum |W_h| \cdot 2^{-23}$). The margin exceeds the perturbation by a factor of 10^6 on average. Only at the 0.11% of neuron-steps with margin < 0.1 can the mantissa change the Boolean outcome.

The mantissa mechanism. At fragile transitions, mantissa noise flips a sign bit with sensitivity $\sim$4.6 (each flip cascades to 4.6 others). Over 520 positions, this injects $\sim$0.13 random flips per step, degrading bpc by 0.095. Removing the mantissa (keeping only sign + exponent) yields 4.870 bpc vs full f32's 4.965.

3 The Minimal Model

Configuration	bpc	Δ	Parameters
Full f32 (baseline)	4.965	0	82,304
Top 20 neurons	4.882	-0.083	37,689
Top 20 + W_h prune (> 3.0)	**4.811**	-0.154	25,857
Top 15 + W_h prune (> 3.0)	4.879	-0.086	22,017
W_h prune (> 3.0) alone	4.903	-0.063	49,753
W_h prune (> 4.0) alone	4.963	-0.002	40,002

Observation 1 (The full model is suboptimal)**.** *Every pruned variant in the table outperforms the full model. The best redux (20 neurons,* W_h *threshold 3.0) achieves 0.15 bpc better with 31% of the parameters. The extra parameters were needed for training (gradient flow through* W_h*) but are noise for inference.*

What the redux removes.

- 108 of 128 W_y columns (the neurons that add noise to readout).
- 64% of W_h entries ($|W_h| < 3.0$, below the 64th percentile).
- These entries don't affect the Boolean dynamics (margins absorb them) and don't help prediction (they contribute noise through W_y).

4 The Routing Backbone

The backward attribution chains (Q6) pass through a small set of neurons:

Neuron	Dominates	Role
h54	7/12 predictions	Decision point (smallest margin, most volatile)
h121	h54's source	Relay for h54
h78	h121's source	Deep source
h0, h30	Secondary routes	Alternative paths
h7, h75	Tertiary routes	Through h56, h30

Observation 2 (h54 is the bottleneck)**.** *h54 has the smallest mean margin (26.7), is the most volatile (234 flips), and dominates 7/12 sampled predictions. It is the point where the RNN's computation is most "undecided"—where context and input battle to determine the sign. The routing backbone h54 ← h121 ← h78 carries the plurality of prediction-relevant information.*

5 The Information Flow

Combining all seven questions:

Step 1: Input enters. The input byte x_t activates the corresponding column of W_x (only ∼13% of W_x entries matter; the rest can be zeroed). This perturbs the pre-activation of each neuron by $W_x[j][x_t]$.

Step 2: Boolean transition. The pre-activation $z_j = b_h^{(j)} + W_x^{(j)} e_{x_t} + \sum_k W_h^{(j,k)} h_t^{(k)}$ is computed. For 98.9% of neurons, the margin $|z_j|$ is so large that $\text{sgn}(z_j)$ determines the next sign bit unambiguously. The top ∼36% of W_h entries ($|W_h| > 3.0$) carry all the dynamically relevant signal; the rest are absorbed by margins.

Step 3: Readout. The output distribution is computed via $W_y \cdot h_t + b_y$. Only ∼20 neurons contribute meaningfully; the other 108 add noise. The sign of h_t (not the magnitude) determines the logit contribution: $W_y[o][j] \cdot (\pm 1)$.

Step 4: Attribution. For each prediction, ∼5 neurons dominate through $|\Delta\text{bpc}|$. Each traces backward through 2–3 W_h hops to input bytes at depths 1–25. The total explanation involves ∼15 weight entries (0.1% of W_h).

6 What Training Did

From ∼10 million random bits (the initial ϵ-field of weight initialization), training by BPTT-50 + Adam selected a map $\phi : H^{128} \times \{0, \ldots, 255\} \to H^{128}$ where:

1. The sign channel carries 99.7% of the compression.
2. The mantissa channel carries 0.3% and actively interferes.
3. 20 of 128 readout neurons suffice for better-than-full prediction.
4. 36% of W_h entries suffice for unchanged dynamics.
5. All 128 neurons are volatile (no frozen features).

6. The dynamics is ergodic (no attractors, no cycles).
7. A 5-neuron routing backbone carries the plurality of signal.
8. The effective offsets are $d = 18$–25, deeper than skip-k-grams.
9. At shallow offsets, the learned patterns match data PMI (87–96%).
10. At deep offsets, the RNN develops higher-order patterns (24–37% PMI alignment).

Training gave us too much. The full model has 82,304 parameters. Inference needs ∼26,000 (the redux). The remaining 56,000 parameters were scaffolding for gradient flow—needed to navigate the optimization landscape but pure overhead once the good map is found.

The mantissa was the ladder. tanh and its mantissa enable gradient-based optimization: BPTT computes $\partial\mathcal{L}/\partial W$ through the Jacobian $\text{diag}(1 - h^2) \cdot W_h$, which requires continuous h values. But the resulting map is Boolean. The mantissa is the ladder; inference is the landing.

7 Open Questions

1. Can we *train* the 26,000-parameter redux directly, or does gradient flow require the full 82,304?
2. What is the cycle length of the Boolean dynamics for a fixed input? (We found no cycles up to period 100 in 50 random starts.)
3. Can the routing backbone (h54, h121, h78) be explained in terms of the data's character-level statistics?
4. Does the 61% PMI alignment improve with higher-order (3-gram, 4-gram) data statistics?
5. Is there a principled way to identify the 20 "good" neurons without running the full model first?