From RNN to Universal Model via Pattern Chains

Claude and MJC

8 February 2026

1 Definitions

bpc Bits per character. $\text{bpc} = -\frac{1}{N}\sum_t \log_2 p(x_t|x_{<t})$. Lower is better. A uniform predictor scores 8.0 bpc on bytes.

UM Universal model. A prediction framework based on explicit patterns (source → destination with strength). Not a single algorithm but a specification: any model that maintains per-event accumulators updated by patterns qualifies.

SN Spectral notation. A concrete syntax for UM patterns: event declarations, pattern strengths in $[0, 255]$, and an event space structure. The SN is a *designed* macrostate—we chose its structure.

ES Event space. The set of events over which the model reasons. Can be factored into product spaces (e.g., input bytes × output bytes).

Binary ES An event space with exactly two events $\{e^+, e^-\}$. Used to represent each hidden neuron under the doubled-E isomorphism. Softmax over the pair gives $P(e^+) = 2^{A^+}/(2^{A^+} + 2^{A^-})$.

Pattern A triple (source, destination, strength) asserting: "when source fires, destination receives `strength` bits of log-support." Patterns are additive: multiple patterns accumulate into per-event accumulators.

Data-term A pattern grounded in dataset frequencies. A probabilistic syllogism: "A was observed; when A is observed, B follows with support s; therefore B receives s bits of evidence." The strength s is the log-luck of the co-occurrence.

Luck (λ) The reciprocal of probability: $\lambda(e) = 1/p(e)$. Log-luck $\Lambda = \log_2 \lambda = -\log_2 p$ is the surprisal.

Quotient (Q) The compression ratio from collapsing an event space. $Q = \lambda$: the quotient equals the luck (unification identity). Positions in the dataset are microstates; an event partitions them.

Doubled-E The isomorphism from RNN to UM: each hidden neuron h_j maps to a binary ES $\{h_j^+, h_j^-\}$ via $\tanh(x) = 2\sigma(2x) - 1$. With float weights, this is exact (0.000% bpc difference).

BPTT Backpropagation through time. The RNN's training algorithm. Truncated at 50 steps for the sat-rnn, imposing a hard cutoff on learnable pattern length.

Atomic pattern A depth-1 pattern in the backward trie: a single (input byte, output byte) pair. The elementary building block.

Skip-pattern A pattern that chains atomic patterns at different time offsets: "byte a at offset k and byte b at offset $j \Rightarrow$ predict output y." Formed by multiplying atomic patterns by time.

Skip-k-gram A skip-pattern using k input bytes at specified offsets to predict the output. Greedy offset selection by complementary MI achieves high compression of the pattern inventory.

Neuron permutation symmetry RNNs with identical architectures trained on overlapping data converge to equivalent representations under neuron relabeling. Individual W_h weights are meaningless across training runs; patterns are the correct level of abstraction.

DSS Dataset size. The number of bytes on which a model is trained. At DSS = 1024, generalization is minimal — the model memorizes. This makes artifact detection straightforward: double the DSS and check which patterns survive.

Readout loss The gap between a constructed model's bpc and the information-theoretic floor, caused by the linear softmax readout's inability to fully exploit hash features.

2 Model Taxonomy

Name	Description	bpc	Notes
sat-rnn	Saturated RNN, 128 hidden, BPTT-50	0.079	Trained model
doubled-E	Exact float UM isomorphic to sat-rnn	0.079	Mathematically exact
sn-quant	$[0, 255]$ SN-quantized UM	0.09–2.1	Chaotic (export gap)
ngram-um	Direct n-gram counting UM	0.081	Order 11
pattern-chain	Explicit $i \to \cdots \to o$ from data	0.067	Order 12, backoff
skip-4-gram	Greedy skip offsets $[1, 8, 20, 3]$	0.069	712 patterns
constructed-8	Hash + greedy-8 offsets, W_y only	0.190	No BPTT training

Table 1: Models in this track of work.

3 Results So Far

1. **Exact isomorphism** (doubled-E): The sat-rnn IS a UM with float-valued pattern strengths. No information loss. (20260131)

2. **SN export gap** (export-gap.pdf): Quantizing to $[0, 255]$ integer strengths causes chaotic bpc due to recurrent error amplification through W_h. The chaos is a BPTT-50 artifact: errors accumulate over 1024 uncontrolled timesteps beyond the 50-step training horizon. 8-bit resolution is adequate per-weight; the loss is a signal about skip connections through time that we haven't captured as explicit patterns.

3. **Pattern-chain UM** (pattern-chain.pdf): Building patterns directly from data bypasses the export gap entirely. Order-by-order: 4.74 (marginal) → 2.05 (bigram) → 0.56 (trigram) → 0.067 (order 12 with backoff). Surpasses sat-rnn at order 10. Uses 6,180 data-term patterns.

4. **Pattern priors** (pattern-prior.pdf): The backward trie decomposes prediction into atomic patterns at each offset. MI decays from 2.69 bits (offset 1) to 1.6 bits (offset 10). This structure is what attention implements at runtime.

5. **Skip-k-grams** (pattern-prior.pdf §4): Greedy offset selection by complementary MI. Four non-contiguous bytes at offsets $[1, 8, 20, 3]$ reach 0.069 bpc with 712 patterns—nearly matching 12 contiguous bytes (0.067 bpc, 6,180 patterns). A 9× compression of the pattern inventory. Offset 8 is chosen before offset 2 because XML tag structure makes distant bytes more informative than adjacent ones.

6. **Pattern survival under DSS doubling** (pattern-prior.pdf §3.2): Approximately 80% of skip-2-gram patterns at DSS = 1024 are artifacts that vanish when the dataset doubles to 2048 bytes. The surviving 20% carry 60–71% of occurrences, with count correlation $r > 0.8$. This cleanly separates structural patterns from overfitting artifacts.

7. **Neuron universality** (pattern-prior.pdf §5): The same RNN neurons (h52, h8, h68) carry skip-pattern information regardless of the skip distance. The W_h highway h8→h52 (+1.28) is the dominant information pathway. Spectral norm $\sigma_1 = 5.5$ explains the chaotic amplification in the export gap.

8. **Neuron permutation symmetry** (pattern-prior.pdf §8): W_h correlation between 1024-trained and 2048-trained models (same architecture, same seed) is $r = 0.06$—effectively zero. Individual weights cannot be compared across training runs. Patterns, not weights, are the correct abstraction for understanding what the RNN has learned.

9. **Write-back construction** (pattern-prior.pdf §7): RNN weights can be constructed from data patterns without gradient-based training. Random binary hashes encode input bytes; a shift register or direct construction provides multi-offset context; only the readout W_y is optimized. Greedy-8 offsets $[1, 8, 20, 3, 27, 2, 12, 7]$ achieve 0.190 bpc (vs contiguous-8 at 0.228, sat-rnn at 0.079). The readout loss (gap to UM floor 0.043) is 0.147 bpc. All constructed models generalize better than the trained sat-rnn on unseen data (5–6 vs 8.2 bpc). Construction decomposes the RNN into three independent problems: offset selection, encoding, and readout.

4 Skeletal Argument

The overall arc, with aspirational points marked [TODO]:

1. An RNN trained on data learns patterns that predict output from input.

2. The doubled-E isomorphism shows these patterns live in a UM.

3. The UM patterns factor through hidden neurons—this is compression, not a fundamental requirement.

4. Building patterns directly from data (pattern-chain) recovers and surpasses the RNN's predictive power.

TODO Reverse isomorphism: map pattern-chain patterns back onto RNN weights. Partial progress: write-back construction achieves 0.190 bpc without BPTT. Three gaps remain: (a) readout loss from linear softmax (0.147 bpc); (b) the constructed W_h (shift register) is a crude approximation of what BPTT learns; (c) the trained model uses all 128 neurons as a single adaptive encoding rather than k independent hash groups. Neuron permutation symmetry ($r = 0.06$) means there is no canonical assignment to reverse-map onto.

5. The backward trie gives the ground-truth attention map. The RNN compresses this into a fixed W_h highway (spectral norm 5.5), using the same neurons regardless of skip distance. This is lossy: the RNN evaluates at 0.079 bpc while the full backward trie (via pattern-chains) reaches 0.067 bpc. The gap quantifies what the 128-neuron bottleneck misses.

TODO The exponential distribution over pattern lengths (RNN prior) can be derived from the architecture. Empirically, skip-4 with 712 patterns nearly matches contiguous order-12 with 6,180 patterns, suggesting that the prior selects for patterns that combine *complementary* offsets, not adjacent ones.

6. Attention mechanisms generalize the backward trie to runtime. The backward trie's MI-by-offset profile (2.69 bits at offset 1, decaying to 1.6 at offset 10) is the static ground truth that learned attention heads approximate. Greedy skip-k-gram offset selection—choosing offsets by complementary MI—is an explicit implementation of what attention does implicitly.

7. The $Q = \lambda$ framework gives a Bayesian account: each pattern performs a quotient on dataset positions (microstates), and the total bpc is the average log-luck. Pattern survival under DSS doubling directly tests this: artifacts have Q values that don't survive because they partition the wrong microstates. See export-gap.pdf §6.6.

TODO Scale: extend from 1024 bytes to enwik9 (10^9 bytes). The pattern-chain approach scales polynomially in data but exponentially in order; skip-k-grams mitigate this by achieving the same bpc with far fewer patterns at non-contiguous offsets. The RNN scales polynomially in both but with a compression bottleneck. Finding the crossover is the engineering question.

5 Open Questions

- **Neuron permutation.** Can we solve the neuron permutation problem by matching neurons by functional role (input/output sensitivity profiles) rather than index? This unblocks reverse isomorphism.

- **Pattern length distribution.** Is the distribution exponential, and with what rate? Skip-k-grams suggest the effective length (in terms of information) is much shorter than the contiguous order.

- **Offset selection mechanism.** Why does greedy MI choose offset 8 before offset 2? Is this a property of XML structure specifically, or of hierarchical markup in general?

- **Scaling.** How does the MI-by-offset curve change with dataset size? At DSS = 1024 there is minimal generalization; at what DSS does the transition to genuine pattern learning occur?

- **Readout gap.** The linear softmax loses 0.147 bpc relative to the UM floor. Can a nonlinear readout (e.g., a small MLP over the hash features) close this gap without reintroducing BPTT-style training?

- **Adaptive encoding.** The trained sat-rnn uses all 128 neurons as a single adaptive representation. The construction partitions them into independent hash groups. What encoding between these extremes achieves the best bpc/generalization trade-off?

Papers in This Archive

1. **export-gap.pdf** — The SN export gap (8 pages)
2. **pattern-chain.pdf** — The pattern-chain UM (5 pages)
3. **pattern-prior.pdf** — Pattern priors, skip-patterns, and write-back (7 pages)
4. **sn-visibility-sat.pdf** — Full SN visibility (7 pages, earlier)