Retrospective: January 31, 2026 Sessions

Hutter RNN Project

January 31, 2026

1 Overview

Five archives were created on January 31, 2026, exploring RNN interpretability through the lens of Universal Models. This retrospective summarizes what worked, what didn't, and key insights.

2 Session Timeline

Archive	Focus	Key Result
_1	Initial ES experiments	5 ESs explain 59% of compression
_2	Tock methodology	Critique: ES explains 15.9%, not 59%
_3	Memory traces, factor maps	Pattern injection theory developed
_4	Pattern injection empirical	1 bit/char head start; Q = λ unification
_5	Retrospective	(this document)

3 What Worked

3.1 Pattern Injection via SVD

Injecting bigram statistics into RNN weights via SVD factorization:

$$P^T = U \cdot S \cdot V^T \implies W_{ho} = U\sqrt{S}, \quad W_{ih} = \sqrt{S}V^T$$

Result: 1 bit/char head start (5.46 $\rightarrow$ 4.47 bpc) without any training.
This validates that UM patterns can be directly translated to RNN weights.

3.2 SVD Component Interpretation

The singular values have real meaning:

- Component 0 (97.5%): Frequency baseline
- Component 1: ASCII vs UTF-8 encoding
- Component 2: Letters vs digits/XML
- Component 3: Bracket structure (Wikipedia citations)
- Components 4–5: UTF-8 internals, phonotactics

Rank-64 truncation keeps all interpretable structure, loses only noise.

3.3 Theoretical Unification

The identity $Q = \lambda$ ("the quotient IS the luck") unifies:

- Bayesian inference (probability updates)
- Thermodynamics (microstate counting)
- Neural network layers (dimensionality reduction)
- Arithmetic coding (interval narrowing)

3.4 Time-Energy Connection

$$\text{bits} \propto \frac{h}{\text{time}}$$

The depth limit $d_{\max} = 24/H_{\text{avg}}$ is simultaneously:

- A bit budget (information)
- An energy budget (thermodynamics)
- A temporal horizon (time)
- A precision limit (computation)

4 What Didn't Work

4.1 W_{hh} Injection

Attempts to inject trigram patterns into W_{hh} failed. The recurrent weights encode temporal patterns that span multiple timesteps, hitting precision limits.

4.2 Memory Depth Prediction

Predicted memory depth ∼12 characters based on $d_{\max} = 24/H_{\text{avg}}$.

Observed: dependency roughly flat out to 30 characters.

Possible explanations:

1. The precision limit applies to gradients (training), not inference
2. The trained W_{hh} has learned efficient encoding
3. The measurement method (conditional variance) has issues
4. tanh normalization prevents precision loss accumulation

4.3 Initial 59% Claim

Session _1 claimed ESs explain 59% of compression. Session _2's critique showed this was wrong—ESs capture only 15.9% of 1-step mutual information.

Lesson: Validate claims against proper baselines.

5 Key Insights

5.1 Carrying Entropy Through Time

"Carrying entropy through layers = carrying entropy through time."

The RNN hidden state h is analogous to the arithmetic coding interval Q. Both accumulate context from the past, both are limited by precision.

5.2 Q Unbounded, f32 Isn't

The fundamental limit: ideal Q carries unbounded information, but float32 has only 24 mantissa bits.

This explains why temporal patterns (W_{hh}) are harder to inject than local patterns (W_{ih}, W_{ho}).

5.3 Dimension as Bandwidth

Embedding dimension d = number of frequencies = bandwidth.

SVD shows English bigrams have effective dimension ∼64. Higher dimensions capture noise, not signal.

6 Current State

Model	Elman RNN 256→128→256
Performance	5.7 bpc (far from 1.1 bpc SOTA)
Understanding	Pattern injection, SVD interpretation, unification

7 Open Questions

1. Why doesn't memory depth show predicted decay?
2. How do LSTMs "select" what entropy to carry?
3. Can we inject longer patterns within precision budget?
4. What is ω (learning function) in gradient terms?

8 Lessons Learned

1. **Validate against baselines**: The 59% claim was wrong because it wasn't compared to proper Markov baselines.
2. **Perturbation ≠ dependency**: Flipping context at distance k measures perturbation propagation, not memory. Need different methods.
3. **Theory predicts, experiments test**: The $d_{\max}$ formula made a testable prediction. It didn't match, which is valuable information.
4. **Unification is powerful**: Seeing Bayes/Thermo/Quotient/AC as the same thing opens new reasoning paths.
5. **Write it up**: Papers force clarity. Every session should produce documented artifacts.