Predictions: Testable Claims for Future Validation

Hutter RNN Project

January 31, 2026

1 Purpose

This document records specific, testable predictions based on the theoretical framework developed in sessions _1 through _4. Future work should validate or refute these claims.

2 Predictions

2.1 P1: Random RNN Shows Predicted Memory Decay

Claim: An *untrained* RNN with random weights will show exponential memory decay at the predicted rate:

$$\text{dependency}(k) \propto e^{-k/d_{\max}}, \quad d_{\max} = \frac{24}{H_{\text{avg}}}$$

Rationale: The trained RNN may have learned to preserve information efficiently. A random W_{hh} should show raw precision limits.

Test: Run the conditional variance experiment on a randomly initialized (untrained) model.

Expected result: Decay to 1/e around $k = 12$ for English text.

2.2 P2: W_{hh} Spectral Radius Near 1

Claim: The trained W_{hh} has spectral radius $|\lambda_{\max}| \approx 1$.

Rationale: Information preservation requires eigenvalues near the unit circle. Smaller eigenvalues cause exponential decay; larger cause explosion.

Test: Compute eigenvalues of trained W_{hh}.

Expected result: $0.9 < |\lambda_{\max}| < 1.1$.

2.3 P3: LSTM Forget Gate Correlates with Entropy

Claim: In an LSTM, the forget gate f_t will be lower (more forgetting) after high-entropy inputs and higher (more remembering) after low-entropy inputs.

Rationale: LSTMs must be selective about what to carry. High-entropy inputs consume more of the bit budget.

Test: Train LSTM, record f_t values, correlate with local entropy.

Expected result: Negative correlation between $H(x_t)$ and f_t.

2.4 P4: Pattern Injection Improves with Rank

Claim: Pattern injection performance improves monotonically with SVD rank up to ∼64, then plateaus.

Rationale: First 64 components capture signal; rest is noise.

Test: Inject at ranks 1, 2, 4, 8, 16, 32, 64, 128, 256. Measure initial bpc.

Expected result:

Rank	Expected bpc
1	∼6.5
8	∼5.0
64	∼4.5
256	∼4.4

2.5 P5: Training Destroys Injection Advantage

Claim: After sufficient training, random-init and injected-init models converge to similar performance.

Rationale: Session _4 showed the gap shrinking (0.99 → 0.28 bpc after 10 epochs). Training overwrites the injected patterns.

Test: Train both models for 100 epochs.

Expected result: Gap < 0.1 bpc after convergence.

2.6 P6: Hidden Size Determines Effective Rank

Claim: The RNN cannot exploit more than H singular components, where H is the hidden size.

Rationale: The hidden state is H-dimensional; it can only represent H independent directions.

Test: Compare H=32, H=64, H=128, H=256 on pattern injection.

Expected result: Performance saturates when rank ≈ H.

2.7 P7: Bigram Injection Fails for Non-English

Claim: Pattern injection from English bigrams will *hurt* performance on non-English text (e.g., Chinese, code).

Rationale: The injected patterns encode English-specific statistics.

Test: Evaluate English-injected model on Chinese Wikipedia or source code.

Expected result: Worse than random init on non-English data.

2.8 P8: Depth Limit Scales with Precision

Claim: Using float64 instead of float32 will double the effective memory depth.

$$d_{\max}^{64} = \frac{53}{H_{\text{avg}}} \approx 26 \text{ characters}$$

Rationale: float64 has 53 mantissa bits vs 24 for float32.

Test: Implement RNN in float64, repeat memory depth experiment.

Expected result: Memory decay threshold shifts from ∼12 to ∼26.

3 Validation Protocol

For each prediction:

1. Record the prediction *before* running the experiment
2. Run the experiment exactly as specified
3. Compare results to prediction
4. Document whether prediction was **confirmed**, **refuted**, or **inconclusive**
5. If refuted, update theory accordingly

4 Scoring

After validation:

8/8 confirmed	Theory is solid
6–7/8 confirmed	Theory is mostly correct, minor refinements needed
4–5/8 confirmed	Theory captures something but has gaps
<4/8 confirmed	Theory needs major revision

5 Summary

ID	Prediction	Status
P1	Random RNN shows memory decay	Untested
P2	W_{hh} spectral radius ≈ 1	Untested
P3	LSTM forget gate correlates with entropy	Untested
P4	Injection improves with rank, plateaus at 64	Untested
P5	Training destroys injection advantage	Partially tested
P6	Hidden size determines effective rank	Untested
P7	English injection hurts non-English	Untested
P8	float64 doubles memory depth	Untested