Critique: The 53% Illusion

Hutter RNN Project

2026-01-31

Call and Response

CALL: "53% improvement from ES features"

We claimed augmented RNN (256 bytes + 5 ES one-hot) achieved 3.44 bpc vs 7.31 bpc baseline. A 53% improvement.

RESPONSE: Bayesian analysis shows ES features provide at most 0.19 bits/char improvement. That's 3.7%, not 53%.

The 53% measured *training dynamics* on 1M chars, not information content. The baseline had exploding weights.

CALL: "ES explains 59% of model compression"

From archive 20260131: "5 character classes explain 59% of model compression."

RESPONSE: ES captures 15.9% of byte-level Markov mutual information. Or 34.7% of joint bigram entropy if you're generous.

$$I(\text{ES}; \text{prev\_ES}) = 0.19 \text{ bits}$$
$$I(\text{byte}; \text{prev}) = 1.19 \text{ bits}$$
$$\text{Ratio} = 15.9\%$$

CALL: "Top pattern: 'b' $\to$ 'n' (strength 255)"

Our pattern extraction found 'b'$\to$'n' as the strongest learned pattern.

RESPONSE: 'b'$\to$'n' is not even in the top 100 English bigrams.

We computed influence $= \sum_h W_x \cdot W_y$, which measures weight geometry, not data statistics. After 1M chars of training, this is noise.

Actual top bigram: 'e'$\to$' ' (191,521 occurrences). We had 'b'$\to$'x' at #3.

CALL: "Quotient spaces = marginalization = bias terms"

We claimed ES input weights are bias terms implementing marginalization. The 53% gap = "cost of learning to marginalize."

RESPONSE: This may be theoretically sound, but the 53% is wrong. If ES weights implement marginalization, the benefit should be $\sim$0.19 bits/char, not 3.87 bits/char.

Something else caused the experimental gap: unstable baseline, training dynamics, or bugs.

CALL: "ES weights are 10-100$\times$ larger than byte means"

Probe showed ES weights amplify class-level signals far beyond marginalization.

RESPONSE: This observation stands. But it doesn't prove the marginalization theory—it may just mean the model learned to use ES features differently than we theorized.

What Went Wrong

1. **Compared apples and oranges.** Barely-trained models (1M chars = 0.1% of data).
2. **Baseline was broken.** Hidden weights exploded to $\pm$100.
3. **Confused dynamics with content.** ES features helped training stability, not compression.
4. **Didn't validate patterns.** Accepted nonsense patterns as "learned structure."
5. **Bayesian analysis came last.** Should have computed bounds first.

What Stands

1. ES transition matrix P is real and meaningful.
2. Data-based pattern extraction works: top patterns match actual bigrams.
3. The visualization infrastructure is solid.
4. ES *does* provide some structure ($\sim$0.19 bits/char).

Original	Corrected
53% improvement	3.7% theoretical max
ES explains 59%	ES explains 15.9% of Markov MI
'b'$\rightarrow$'n' is top pattern	'e'$\rightarrow$' ' is top pattern
Marginalization = 3.87 bpc	Marginalization $\leq$ 0.19 bits

Corrected Claims

Path Forward

1. Train to convergence before comparing
2. Use Bayesian bounds as sanity checks
3. Validate extracted patterns against data
4. Separate "makes training easier" from "captures information"