Microstates, Macrostates, and the Partition Function: The Thermodynamic Structure of the Universal Model

Claude and MJC

11 February 2026

Abstract

We derive the exact correspondence between the Universal Model and statistical mechanics. Dataset positions are microstates. Patterns (partial event descriptions) are macrostates. The SN pattern strength is the Boltzmann entropy $S = \log_2 \Omega$. The UM's binary-ES softmax is the Boltzmann distribution with inverse temperature $\beta = \ln 2$. The forward pass computes a chain of partition functions, one per layer. The free energy $F = -\beta^{-1} \ln Z$ at the output layer determines the bpc. Factoring (projecting onto fewer event spaces) is coarse-graining: it increases the macrostate entropy while preserving the predictive partition function. We derive the second law in this setting: factoring never decreases entropy, and the hidden layer is the optimal coarse-graining that maximizes mutual information with the output. The mapping goes both ways: $E \to \mathbb{N}$ (counting microstates gives Boltzmann entropy) and $\mathbb{N} \to E$ (the partition function determines which events dominate).

1 The Microstate Space

Definition 1 (Microstate)**.** *A* microstate *is a complete specification of the system at one dataset position. At position t, the microstate is:*

$$\mu_t = (t,\ x_t,\ h^{\pm}_{0,t}, \ldots, h^{\pm}_{127,t},\ y_{t+1}) \in \mathcal{M} \tag{1}$$

where $x_t \in E_{in} = \{0, \ldots, 255\}$ is the input byte, $h^{\pm}_{j,t} \in \{h^+_j, h^-_j\}$ is the binary hidden event for neuron j, and $y_{t+1} \in E_{out}$ is the actual next byte.

The total microstate space has N elements (one per dataset position):

$$|\mathcal{M}| = N = 1024 \tag{2}$$

Each microstate is unique: position t has a specific input, a specific hidden state, and a specific output. There are no degenerate microstates (984 of 1024 have distinct binary hidden states; the remaining 40 share states with other positions but differ in t, x_t, or y_{t+1}).

Observation 1 (The microstate space IS the dataset)**.** *$\mathcal{M}$ is not an abstract space of possible configurations—it is the actual dataset. Each microstate is a* fact*: "at position 42, the input was* `a`*, neuron 7 was positive, ..., and the next byte was* `e`*." The microstates are given, not sampled. Statistical mechanics usually works with ensembles (imagined copies of the system); here we work with the actual data.*

2 The Macrostate Space

Definition 2 (Macrostate)**.** *A* macrostate *is a partial specification: a constraint on which microstates are consistent. Formally, a macrostate M is a subset of $\mathcal{M}$:*

$$M \subseteq \mathcal{M}, \qquad \Omega(M) = |M| \tag{3}$$

where $\Omega(M)$ is the number of consistent microstates.

Macrostates arise from *patterns*—conjunctions of events:

Macrostate	Events specified	Ω	$S = \log_2 \Omega$
$M_\emptyset$ (no constraint)	none	1024	10.00
“input is a”	$x_t = \mathsf{a}$	45	5.49
“h_7^+”	$h_{7,t} = +1$	537	9.07
“input a, h_7^+”	$x_t = \mathsf{a}, h_7^+$	24	4.58
“input a, output e”	$x_t = \mathsf{a}, y = \mathsf{e}$	15	3.91
“input a, h_7^+, output e”	3 events	8	3.00
Full microstate at $t = 42$	all events	1	0.00

Theorem 1 (Macrostate entropy is SN strength)**.** *The Boltzmann entropy of macrostate M (in bits) equals the SN strength of the pattern that defines M:*

$$\boxed{S(M) = \log_2 \Omega(M) = \log_2 c(p_M) = s(p_M)} \tag{4}$$

where p_M is the pattern and $c(p_M)$ is its count in the dataset.

Proof. By definition, $\Omega(M) = |\{t : \mu_t \in M\}| = c(p_M)$. The SN strength is $s = \log_2 c$. □

3 The Partition Function

3.1 The binary-ES partition function

At each timestep, each hidden neuron j has a binary event space $\{h_j^+, h_j^-\}$ with accumulator difference D_j. The partition function for this 2-state system:

$$Z_j = 2^{A_j^+} + 2^{A_j^-} = 2^{A_j^+}(1 + 2^{-(A_j^+ - A_j^-)}) = 2^{A_j^+}(1 + 2^{-D_j}) \tag{5}$$

The Boltzmann probabilities:

$$P(h_j^+) = \frac{2^{A_j^+}}{Z_j} = \frac{1}{1 + 2^{-D_j}} = \sigma(D_j \ln 2) \tag{6}$$

Definition 3 (Binary-ES free energy)**.** *The free energy of the binary ES for neuron j:*

$$F_j = -\frac{1}{\beta} \ln Z_j = -\frac{1}{\ln 2} \ln(2^{A_j^+} + 2^{A_j^-}) = -\frac{1}{\ln 2}\Big(A_j^+ \ln 2 + \ln(1 + 2^{-D_j})\Big) \tag{7}$$

where $\beta = \ln 2$ (inverse temperature).

The free energy determines the “cost” of the hidden event. When $|D_j| \gg 0$ (deep saturation), $F_j \approx -A_j^+$ (free energy ≈ negative accumulator). The entropy contribution $S_j = -P_+ \log P_+ - P_- \log P_- \to 0$: no uncertainty, no cost.

When $D_j = 0$, $S_j = 1$ bit (maximum uncertainty), and the free energy includes a full bit of entropy cost.

3.2 The output partition function

The output event space has 256 possible bytes. The partition function:

$$Z_{\text{out}} = \sum_{o=0}^{255} e^{A(o,t)} \tag{8}$$

The output probability:

$$P(o \mid h_t) = \frac{e^{A(o,t)}}{Z_{\text{out}}} \tag{9}$$

The output free energy:

$$F_{\text{out}} = -\frac{1}{\beta} \ln Z_{\text{out}} = -\frac{1}{\ln 2} \ln \sum_{o} e^{A(o,t)} \tag{10}$$

3.3 bpc as free energy difference

Theorem 2 (bpc = free energy gap)**.** *The bpc at position t for output y is:*

$$bpc(t) = \log_2 \frac{1}{P(y \mid h_t)} = \log_2 Z_{out} - \frac{A(y,t)}{\ln 2} = \frac{F_{out} - E(y)}{\ln 2/\beta} \tag{11}$$

where $E(y) = -A(y,t)$ is the "energy" of the actual output. The bpc is the free energy gap between the partition function (total evidence for all outputs) and the evidence for the actual output.

Proof. $P(y) = e^{A(y)}/Z$, so $-\log_2 P(y) = \log_2 Z - A(y)/\ln 2$. □

Corollary 1 (Average bpc = average free energy gap)**.**

$$bpc = \frac{1}{N} \sum_t bpc(t) = \frac{1}{N} \sum_t \left(\log_2 Z_{out}(t) - \frac{A(y_{t+1},t)}{\ln 2} \right) \tag{12}$$

4 Factoring as Coarse-Graining

4.1 The factoring operation

Definition 4 (Factoring)**.** *A* factoring *of the microstate space $\mathcal{M}$ is a projection $\pi : \mathcal{M} \to \mathcal{M}'$ that maps microstates to macrostates by "forgetting" some event spaces. For example:*

$$\pi_{input} : (t, x_t, h_{0,t}, \ldots, h_{127,t}, y_{t+1}) \mapsto x_t \tag{13}$$

projects onto the input event space alone.

Factoring increases entropy:

Theorem 3 (Factoring inequality (second law))**.** *For any factoring π:*

$$S(\pi(\mu)) \geq S(\mu) \tag{14}$$

with equality iff π is injective on $\mathcal{M}$ (no two microstates map to the same macrostate).

Proof. $\pi(\mu)$ is a macrostate consistent with μ and possibly other microstates. $\Omega(\pi(\mu)) \geq 1 = \Omega(\mu)$. Hence $S(\pi(\mu)) = \log_2 \Omega(\pi(\mu)) \geq 0 = S(\mu)$. □

This is the second law of thermodynamics in the UM setting: coarse-graining (forgetting details) never decreases entropy. Compression (the RNN's hidden layer) is a factoring: it projects the full microstate onto 128 binary events, increasing the macrostate entropy.

4.2 The factoring hierarchy

The RNN's architecture defines a hierarchy of factorings:

1. **Full microstate**: $\mathcal{M}$ with 1024 elements. $S = 0$ per position.
2. **Hidden state factoring**: $\pi_H : \mu_t \mapsto (h_{0,t}, \ldots, h_{127,t})$. The 128 binary events form a macrostate with $\Omega(h) = |\{t : h_t = h\}|$ consistent positions. Average $S_H = \frac{1}{N}\sum_t \log_2 \Omega(\pi_H(\mu_t)) \approx$ 0.06 bits.
3. **Redux factoring**: $\pi_{20} : \mu_t \mapsto (h_{j_1,t}, \ldots, h_{j_{20},t})$ for the top 20 neurons. Coarser: more positions per macrostate, higher entropy. But bpc *improves* by 0.15—the extra 108 neurons were noise.
4. **Input factoring**: $\pi_x : \mu_t \mapsto x_t$. 52 macrostates. $S = 4.74$ bits (the marginal entropy).
5. **Trivial factoring**: $\pi_\emptyset : \mu_t \mapsto *$. One macrostate containing all positions. $S = 10.0$ bits.

Observation 2 (The hidden layer is a near-optimal coarse-graining)**.** *The hidden state factoring achieves $S_H \approx 0.06$ bits—nearly as specific as the full microstate (984 distinct states out of 1024 positions). Yet the output prediction (0.079 bpc) is far from zero, meaning the hidden state is not fully exploited by W_y.*

The redux factoring (20 neurons) has higher S_H but lower bpc: it is a better *coarse-graining because it discards the 108 noise neurons that increase entropy without increasing mutual information with the output.*

The optimal coarse-graining maximizes $I(macrostate; y_{t+1})$, not minimizes S_H. The second law guarantees S_H increases under coarser factoring; the art is choosing which details to forget.

5 The Two Directions of the Thermodynamic Map

5.1 $E \to \mathbb{N}$: Counting microstates gives entropy

Starting from events:

$$\underbrace{e \in E}_{\text{event}} \xrightarrow{\text{count}} \underbrace{c(e) = \Omega(M_e) \in \mathbb{N}}_{\text{microstate count}} \xrightarrow{\log} \underbrace{S(e) = \log_2 c(e)}_{\text{entropy (= SN strength)}} \tag{15}$$

This is the Boltzmann direction: observe the system, count consistent configurations, take the log. Every SN strength in the UM is an entropy computed this way.

For a pattern p over multiple events:

$$c(p) = |\{t : \text{all events in } p \text{ hold at position } t\}| \tag{16}$$

The joint count gives the joint entropy. The conditional entropy:

$$S(y \mid x) = \log_2 c(x, y) - \log_2 c(x) = \log_2 \frac{c(x, y)}{c(x)} \tag{17}$$

Wait—this is the conditional log-count, which is $-\log_2 \lambda(y|x)$ (negative log-luck). The conditional entropy $H(Y|X) = \mathbb{E}[-\log P(Y|X)]$ is the average over positions.

5.2 $\mathbb{N} \to E$: The partition function determines dominant events

Starting from numbers:

$$\underbrace{A(e) \in \mathbb{R}}_{\text{accumulator}} \xrightarrow{Z=\sum e^{A(e)}} \underbrace{Z}_{\text{partition function}} \xrightarrow{P(e)=e^{A(e)}/Z} \underbrace{e^* = \arg\max P(e)}_{\text{dominant event}} \tag{18}$$

This is the statistical mechanics direction: given the energies (accumulators), compute the partition function, identify which events dominate. The dominant event is the one with the most microstate support (lowest energy, highest count).

The model *predicts* events by computing which macrostate has the most microstates: $\hat{y} = \arg\max_o P(o) = \arg\max_o e^{A(o)} = \arg\max_o c(o \mid \text{context})$. Prediction IS finding the macrostate with maximum entropy (conditioned on the observed context).

5.3 The bidirectional thermodynamic map

Theorem 4 (Bidirectional thermodynamic map)**.** *The UM's forward pass implements the map $E \to \mathbb{N} \to E$ (events to numbers to events) via:*

1. *$E \to \mathbb{N}$: input events determine accumulators through pattern strengths (which are log-microstate-counts).*

2. *$\mathbb{N} \to \mathbb{N}$: accumulators combine additively (pattern composition = microstate intersection counting).*

3. *$\mathbb{N} \to E$: the partition function determines the output event probabilities (dominant macrostates).*

The backward pass (attribution/gradient) implements $E \to \mathbb{N} \to E$ in reverse:

1. *$E \to \mathbb{N}$: the output event y determines the gradient g_t (evidence for the prediction).*

2. *$\mathbb{N} \to \mathbb{N}$: the gradient propagates backward through the Jacobian (reverse partition function).*

3. *$\mathbb{N} \to E$: the backward gradient identifies which input events contributed to the prediction.*

6 The Free Energy Landscape

6.1 Per-pattern free energy

Each pattern p with strength $s(p) = \log_2 c(p)$ contributes to the free energy of the system:

$$F(p) = -\frac{s(p)}{\beta} = -\frac{\log_2 c(p)}{\ln 2} = -\log_e c(p) = -\ln \Omega(p) \tag{19}$$

This is $F = -k_B T \ln \Omega$ with $k_B T = 1/\beta = 1/\ln 2$.

The *total* free energy at the output:

$$F_{\text{out}} = -\frac{1}{\ln 2} \ln \sum_o e^{A(o)} = -\frac{1}{\ln 2} \ln Z_{\text{out}} \tag{20}$$

6.2 The free energy decomposition

Proposition 1 (Free energy decomposes by pattern layer)**.** *The output free energy decomposes as:*

$$F_{out} = F_{prior} + \Delta F_{W_y} + \Delta F_{softmax} \tag{21}$$

where:

- $F_{prior} = -\frac{1}{\ln 2} \ln \sum_o e^{b_y[o]}$*: the prior free energy from biases alone.*
- ΔF_{W_y}*: the free energy change from hidden-state evidence (*$W_y \cdot h_t$ *contributions).*
- $\Delta F_{softmax}$*: the normalization correction.*

The bpc at each position is:

$$\text{bpc}(t) = -\frac{A(y_{t+1}, t)}{\ln 2} - F_{\text{out}}(t) = \frac{E(y_{t+1}) + F_{\text{out}}(t)}{1/\ln 2} \tag{22}$$

This is the standard statistical mechanics relation: the surprise of an event equals its energy minus the free energy, divided by kT.

6.3 The free energy landscape of the weight space

The training loss $\mathcal{L} = \text{bpc}$ is the average free energy gap over all positions. Training by BPTT navigates the 82,304-dimensional weight space to minimize this average.

Observation 3 (The free energy landscape has the thermodynamic structure)**.** *In the* $E \to \mathbb{N}$ *direction:*

- *The analytic construction (counting) reaches a point in weight space with* $\mathcal{L} = 1.89$ *bpc. This point is determined by the data's microstate counts: it sits at the "thermodynamic equilibrium" of the* $E \to \mathbb{N}$ *map.*
- *BPTT training reaches a different point with* $\mathcal{L} = 4.97$ *bpc. This point is a local minimum of the free energy landscape in* $\mathbb{N}$*-space.*

The analytic construction finds a lower free energy because it computes in E*-space (convex) rather than navigating* $\mathbb{N}$*-space (non-convex). The trained model is stuck in a metastable state.*

7 The Entropy Budget

7.1 Shannon entropy from microstates

The Shannon entropy of the output distribution at position t:

$$H_t = -\sum_o P(o \mid h_t) \log_2 P(o \mid h_t) \tag{23}$$

The bpc at position t (for the actual output y_{t+1}):

$$\text{bpc}(t) = -\log_2 P(y_{t+1} \mid h_t) = \log_2 Q_{\text{out}}(t) \tag{24}$$

The relationship:

$$\text{bpc}(t) \geq H_t \quad \text{(with equality iff output is uniform over support)} \tag{25}$$

The average bpc is the cross-entropy between the model's distribution and the data's distribution. When the model is perfect, bpc $= H$ (the true entropy rate of the source).

7.2 The entropy budget by layer

The total entropy $S_{\text{total}} = \log_2 N = 10$ bits distributes across layers:

Layer	Entropy budget	Meaning
Input (marginal)	4.74 bits	How rare is the input byte?
Hidden layer	∼0.06 bits	How much uncertainty at the bottleneck?
Residual	5.20 bits	Information not used for prediction
Output (bpc)	0.079 bits	How surprised is the model?
Total	10.00 bits	$\log_2 1024$

The input carries 4.74 bits of the 10-bit budget. The hidden layer adds 0.06 bits of uncertainty (nearly zero: the bottleneck is tight). The output extracts 0.079 bits of surprise (the model's prediction error). The remaining 5.20 bits are the *redundancy* in the representation: information present in the hidden state but not needed for prediction.

7.3 Mutual information as the useful entropy

Proposition 2 (MI = total entropy minus residuals)**.** *The mutual information between the hidden state and the output:*

$$I(h_t; y_{t+1}) = H(y_{t+1}) - H(y_{t+1} \mid h_t) = 4.74 - 0.079 = 4.66 \textit{ bits} \tag{26}$$

This is the useful entropy: the amount by which the hidden state reduces uncertainty about the output. The model uses 4.66 of the 4.74 bits of output entropy (98.3% efficiency).

The redux (20 neurons) achieves $I \approx 4.66 + 0.15 = 4.81$ bits of useful entropy—*more* than the full model. The extra 108 neurons carry information that is correlated with the output *noise* rather than the output *signal*.

8 Bidirectional Factoring

8.1 $E \to \mathbb{N}$: Fine-to-coarse (the second law)

Factoring from finer to coarser macrostates:

$$\text{microstate} \xrightarrow{\pi_H} \text{hidden state} \xrightarrow{\pi_{20}} \text{20-neuron redux} \xrightarrow{\pi_{\text{pair}}} \text{2-offset pair} \xrightarrow{\pi_x} \text{input byte} \xrightarrow{\pi_\emptyset} * \tag{27}$$

Each arrow is a factoring that increases entropy. The second law guarantees monotonicity:

$$0 \le S_H \le S_{20} \le S_{\text{pair}} \le S_x \le 10 \tag{28}$$

The $E \to \mathbb{N}$ direction *compresses*: each factoring replaces a detailed description with a coarser one. The count of consistent microstates grows.

8.2 $\mathbb{N} \to E$: Coarse-to-fine (refinement)

The reverse direction *refines*: given a coarse macrostate, add events to narrow the set of consistent microstates.

$$* \xrightarrow{+x_t} \text{input byte} \xrightarrow{+h_{j_1}} \text{1 neuron} \xrightarrow{+h_{j_2}} \text{2 neurons} \xrightarrow{\cdots} \text{hidden state} \xrightarrow{+t} \text{microstate} \tag{29}$$

Each arrow adds an event, decreasing entropy. The added event provides information that distinguishes among microstates.

This is the $\mathbb{N} \to E$ direction of the partition function: given the current macrostate (with its partition function Z), compute which additional event would most reduce the free energy. The model's prediction is this computation: given the hidden state (a macrostate), which output event has the lowest energy (highest count)?

8.3 The bidirectional factoring at the hidden layer

The hidden layer is the meeting point of the two directions:

- $E \to \mathbb{N}$ **(encoding)**: the input history $(x_0, \ldots, x_t)$ is factored through W_x and W_h into 128 binary hidden events. This is coarse-graining: the exponentially many input histories map onto ∼984 distinct macrostates.
- $\mathbb{N} \to E$ **(decoding)**: the hidden state h_t is projected through W_y to predict the output event. This is refinement: the 128 bits of hidden state narrow the 256-byte output space.

The hidden layer is an *information bottleneck*: it retains only the information about the input history that is relevant for predicting the output, discarding the rest. The thermodynamic analogue: the hidden layer is the *macrostate description* that maximizes $I(\text{macrostate}; \text{output})$ subject to the constraint $H(\text{macrostate}) \leq 128$ bits.

9 The Boltzmann Distribution in Detail

9.1 Deriving the UM softmax from maximum entropy

The maximum entropy principle states: among all distributions consistent with the known constraints (the accumulated evidence), choose the one with maximum entropy.

For the binary ES of neuron j, the constraint is the accumulator difference D_j:

$$\mathbb{E}[\sigma_j] = P(h_j^+) - P(h_j^-) \quad \text{subject to} \quad \mathbb{E}[\sigma_j] \cdot D_j = \text{evidence} \tag{30}$$

The maximum entropy distribution is the Boltzmann distribution:

$$P(h_j^+) = \frac{e^{\beta D_j/2}}{e^{\beta D_j/2} + e^{-\beta D_j/2}} = \sigma(\beta D_j) \tag{31}$$

with $\beta = \ln 2$ (the UM's base-2 convention).

Theorem 5 (UM softmax = maximum entropy)**.** *The UM's binary-ES softmax computes the maximum entropy distribution consistent with the accumulated log-support. The inverse temperature $\beta = \ln 2$ is determined by the convention that pattern strengths are measured in bits (base-2 logarithms).*

9.2 Temperature and the base of the logarithm

The "temperature" of the UM is:

$$T = \frac{1}{k_B \beta} = \frac{1}{\ln 2} \approx 1.443 \tag{32}$$

(with $k_B = 1$). This is the conversion factor between nats and bits.

At $T = 1/\ln 2$:

- A 1-bit pattern strength ($s = 1$) corresponds to an energy difference of $\ln 2$ nats, which gives an odds ratio of $2 : 1$.
- A 10-bit pattern strength gives odds $1024 : 1$.
- The sat-rnn's mean margin $|D_j| = 60.5$ gives odds $2^{60.5} : 1 \approx 10^{18} : 1$—the neurons are "frozen" at astronomical temperature ratios.

9.3 Phase transitions and saturation

In thermodynamics, a phase transition occurs when a small change in temperature causes a discontinuous change in the macrostate. In the UM:

Observation 4 (Saturation is a frozen phase)**.** *The sat-rnn operates deep in the "frozen" phase: $\beta|D_j| \gg 1$ for 98.9% of neuron-steps. In this regime, each binary ES is effectively deterministic—one state dominates exponentially.*

The 1.1% of neuron-steps with $|D_j| < 1.0$ are "critical": near the phase boundary where both states are comparably likely. These are the fragile transitions where the mantissa (thermal noise) can flip the outcome.

The mantissa's contribution (-0.095 bpc: it degrades *prediction) is the thermal noise of the frozen phase: fluctuations at critical points that inject randomness into an otherwise deterministic system.*

10 Conclusion

The Universal Model is a statistical mechanical system. Dataset positions are microstates. Patterns define macrostates. The SN strength is the Boltzmann entropy. The forward pass computes partition functions layer by layer. The bpc is the free energy gap. Factoring is coarse-graining. The second law holds: coarser macrostates have higher entropy. The hidden layer is an information bottleneck that maximizes mutual information with the output.

The mapping goes both ways. $E \to \mathbb{N}$: counting microstates gives entropy and determines pattern strengths. $\mathbb{N} \to E$: the partition function identifies dominant events and generates predictions. The RNN's forward pass is the forward thermodynamic computation (macrostate → partition function → prediction). The attribution chain is the reverse (prediction → partition function → responsible microstates).

References

[1] Michaeljohn Clement. CMP. 2026. `https://cmpr.ai/cmp.pdf`

[2] Claude and MJC. The SN Export Gap. 7 Feb 2026.

[3] Claude and MJC. The Pattern-Chain UM. 8 Feb 2026.

[4] Claude and MJC. The Hidden Quotient. 8 Feb 2026.

[5] Claude and MJC. The Factor Map. 9 Feb 2026.

[6] Claude and MJC. From Counting to Construction. 11 Feb 2026.

[7] Claude and MJC. E onto $\mathbb{N}$. 11 Feb 2026.

[8] Claude and MJC. The Quotient Chain. 11 Feb 2026.