Patterns in the Extended Event Space: Independence, Correlation, and the New Synapses

Claude and MJC

February 15, 2026

Abstract

The nested model [2] extends the event space from $E = I \times H \times O$ to $E' = I \times (I' \times H_{\text{inner}} \times O') \times O$, adding lexical events (position, accumulator, bag-of-letters, word identity) inside H. But adding events without adding patterns changes nothing: the extended model with zero new synapses is literally the base model with unused neurons. This paper formalizes the pattern space $P = E^2 \times T$ (source event, target event, weight), characterizes which new patterns are valid under the independence requirement for evidence accumulation, and confronts the central difficulty: interior events like word identity and letter accumulation are *highly correlated* with the byte stream, so naïve Bayesian combination of their evidence is catastrophically wrong. We give the correct combination rules, classify the new patterns into four families, and make testable predictions about their information contribution.

1 The Empty Extension

Proposition 1 (New events without new patterns are inert)**.** *Let $U = (e, t, p, f, \omega)$ be a UM with event space E. Let $E' = E \times E_{\text{new}}$ be a product extension with $|E_{\text{new}}| = k$ new events. If the pattern table p' has $p'_{e_{\text{new}},j} = 0$ and $p'_{i,e_{\text{new}}} = 0$ for all new events $e_{\text{new}} \in E_{\text{new}}$ and all $i, j \in E'$, then:*

$$f_{p'}(t')_j = f_p(t)_j \qquad \text{for all } j \in E. \tag{1}$$

The new events participate in no syllogisms. The forward pass is unchanged. The model has gained neurons but no synapses.

Proof. The forward pass $(f_{p'}(t'))_j = \max_i \min(t'_i, p'_{ij})$. For $j \in E$, the max ranges over $i \in E' = E \cup E_{\text{new}}$. For $i \in E_{\text{new}}$, $p'_{ij} = 0$, so $\min(t'_i, 0) = 0$. The terms from new events contribute nothing. The remaining terms are exactly $\max_{i \in E} \min(t_i, p_{ij}) = f_p(t)_j$. □

Remark 2 (The synapse is the unit of value)**.** *The nested model paper [2] proves that the extension is conservative (the outer model's predictions are unchanged) and that the inner model is a complete UM. But conservative means* inert *until patterns are added. The value of the extension comes entirely from the new patterns—the synapses connecting old events to new events and new events to each other. This paper characterizes those synapses.*

2 The Atomic Pattern

Definition 3 (Atomic pattern)**.** *An* atomic pattern *is a triple $(e_{\text{from}}, e_{\text{to}}, w) \in E \times E \times T$, where:*

- $e_{\text{from}} \in E$ *is the* evidence event *(the antecedent),*

- $e_{\text{to}} \in E$ *is the* conclusion event *(the consequent),*
- $w \in T = \{0, 1, \ldots, 255\}$ *is the* weight *(log support of joint occurrence).*

The weight is:

$$w = s(e_{\text{from}}, e_{\text{to}}) = \lfloor \log_2 c(e_{\text{from}}, e_{\text{to}}) \rfloor, \tag{2}$$

where $c(e_{\text{from}}, e_{\text{to}})$ is the count of joint occurrences in the data.

Proposition 4 (Weight as upper bound on conditional support)**.** *By the forward pass, when evidence e_{from} is observed with support t_{from}, the conclusion e_{to} receives support at most $\min(t_{\text{from}}, w)$. Since w is the log of the joint count:*

$$\min(t_{\text{from}}, w) \leq w = \lfloor \log_2 c(e_{\text{from}}, e_{\text{to}}) \rfloor. \tag{3}$$

The pattern weight is an upper bound on the support that the consequent can receive from this particular antecedent. This is the content of the "weakest link" principle: the syllogism's conclusion cannot exceed the evidence for the major premise [8].

Definition 5 (Pattern space)**.** *The* pattern space *of a UM with event space E is:*

$$P = \{(e_i, e_j, w_{ij}) : e_i, e_j \in E,\ w_{ij} > 0\} \subseteq E \times E \times T. \tag{4}$$

The pattern space has at most $|E|^2$ elements (one for each ordered pair of events with nonzero joint count). In practice $|P| \ll |E|^2$ because most event pairs never co-occur.

3 The Four Families of New Patterns

The extended event space $E' = I \times I' \times H_{\text{inner}} \times O' \times O$ introduces four structurally distinct families of new patterns, classified by which spaces the source and target events inhabit.

Definition 6 (Pattern families)**.** *1.* ***Recognition patterns*** *($I \to I'$ and $I \to H'$): Byte-level input events predict interior events.*

- *($byte = \texttt{t}, pos = 0$) $\to$ ($acc \ni \texttt{t}$): deterministic (weight 255).*
- *($byte = \texttt{h}, pos = 1, acc \supseteq \{\texttt{t}\}$) $\to$ ($bol_{the} = 0$): "bag-of-letters for 'the' is incomplete after seeing 'th'". Deterministic (weight 255).*
- *($byte = \texttt{e}, pos = 2, acc = \{\texttt{t}, \texttt{h}\}$) $\to$ ($bol_{the} = 1$): "bag-of-letters for 'the' is complete". Deterministic (weight 255).*
- *($acc = \{\texttt{t}, \texttt{h}, \texttt{e}\}, pos \leq 3$) $\to$ ($\sigma_{the} \approx 0.92$): graded word support, learned from data (weight $\sim$ 200).*

2. ***Prediction patterns*** *($H' \to O'$ and $H' \to O$): Interior events predict output bytes.*

- *($\sigma_{the} > 0.9, pos = 3$) $\to$ ($next\ byte = \texttt{' '}$): word identity predicts word boundary. Weight $\approx \log_2 c(the, space)$.*
- *($word = \texttt{the}$) $\to$ ($byte\ at\ offset\ 2 = \texttt{e}$): word identity predicts specific future byte.*

3. ***Language-model patterns*** *($H' \to H'$): Interior events predict other interior events across time.*

- *($prev\_word = \texttt{of}$) $\to$ ($\sigma_{the} \uparrow$): word-to-word transition.*

- ($prev\_word =$ `the`) $\to$ (*next word is noun* $\uparrow$)*: syntactic prediction.*

4. ***Structural patterns*** *($I' \to I'$ and $O' \to O'$): Within-layer patterns that express deterministic constraints.*

   - ($pos = k$) $\to$ ($pos = k + 1$)*: position increment (weight 255, deterministic).*
   - ($acc = A$) $\to$ ($acc = A \cup \{$*current byte*$\}$)*: accumulator update (weight 255).*

Proposition 7 (Count of new patterns)**.** *Let $|V|$ be the vocabulary size. The pattern families contribute:*

Family	*Max patterns*	*Typical nonzero*
Recognition ($I \to I', H'$)	$256 \times L_{\max} \times \lvert V\rvert$	$O(\lvert V\rvert \cdot \bar{L})$
Prediction ($H' \to O', O$)	$\lvert V\rvert \times 256 \times D$	$O(\lvert V\rvert \cdot \bar{L})$
Language-model ($H' \to H'$)	$\lvert V\rvert^2$	$O(\lvert V\rvert \cdot B)$
Structural ($I' \to I'$)	$256 \times L_{\max}$	$O(L_{\max})$

where $\bar{L}$ is mean word length, D is prediction horizon, and B is the mean number of distinct following words per word. For $|V| = 1000$, $\bar{L} = 5$, $D = 20$, $B = 50$: roughly 10^5 recognition, 10^5 prediction, 5×10^4 language-model, and 5×10^3 structural patterns.

4 The Independence Requirement

Evidence accumulates into the log product pattern (LPP) via addition of log supports from independent sources. The forward pass combines multiple patterns via max, selecting the best-supported syllogism. But the LPP within each syllogism requires that the evidence arriving from different patterns be *independent*.

Definition 8 (Evidence independence)**.** *Two patterns (e_1, e_j, w_1) and (e_2, e_j, w_2) targeting the same conclusion e_j provide* independent evidence *for e_j if:*

$$P(e_1, e_2 \mid e_j) = P(e_1 \mid e_j) \cdot P(e_2 \mid e_j). \tag{5}$$

Equivalently, the evidence events e_1 and e_2 are conditionally independent given the conclusion.

Theorem 9 (Independence from microstate partition)**.** *Two evidence sources provide genuinely independent evidence if and only if they condition on disjoint partitions of the microstate space [7]. Concretely: if the microstates supporting e_j can be partitioned as $M_j = M_1 \times M_2$ such that e_1 conditions on M_1 and e_2 conditions on M_2, then e_1 and e_2 are independent given e_j.*

Proof. This is the product structure of independent events. If e_j's microstates factor as $M_1 \times M_2$ and e_1 depends only on M_1 while e_2 depends only on M_2, then:

$$P(e_1, e_2 \mid e_j) = P(e_1 \mid M_1) \cdot P(e_2 \mid M_2) = P(e_1 \mid e_j) \cdot P(e_2 \mid e_j).$$

Conversely, if e_1 and e_2 both depend on the same component of the microstate, they share a common cause and are dependent. □

Corollary 10 (The offset graph criterion)**.** *For skip-bigram patterns, independence requires vertex-disjoint edges in the offset graph [6]. A shared offset guarantees dependence. The same principle applies to any evidence source: two pieces of evidence are independent only when they "look at different parts of the world."*

5 The Correlation Problem

The extended event space introduces interior events—position, accumulator, bag-of-letters, word identity—that are *deterministic functions of the byte stream.* This creates massive correlations.

Theorem 11 (Interior events are not independent of bytes)**.** *Let b_t be the byte at time t,* $\text{pos}(t)$ *the in-word position,* $\text{acc}(t)$ *the letter accumulator, and $\sigma_w(t)$ the graded word support. Then:*

1. $\text{pos}(t)$ *is a deterministic function of* $(b_0, \ldots, b_t)$*.*
2. $\text{acc}(t)$ *is a deterministic function of* $(b_0, \ldots, b_t)$*.*
3. $\sigma_w(t)$ *is a deterministic function of* $(\text{pos}(t), \text{acc}(t))$*, hence of* $(b_0, \ldots, b_t)$*.*

Therefore, for any output event o:

$$P(\text{pos}(t), b_t \mid o) \neq P(\text{pos}(t) \mid o) \cdot P(b_t \mid o) \tag{6}$$

in general. The interior events and the byte events are conditionally dependent given any output.

Proof. If $\text{pos}(t) = 0$, then b_{t-1} was a word boundary character (space, punctuation), which constrains b_t (more likely uppercase or common word-initial letters). This constraint is not mediated by o alone—it is a direct dependence between pos and b_t. □

Corollary 12 (Naïve Bayesian combination fails)**.** *If we treat a byte-level pattern* $(b_{t-1} = \text{' '}) \to (o = \text{T})$ *and an interior pattern* $(\text{pos}(t) = 0) \to (o = \text{T})$ *as independent evidence for the output, we double-count: both patterns encode the same underlying fact (we are at a word boundary). This is the shared-offset catastrophe [6] generalized from temporal offsets to event-space dimensions.*

6 The Correct Combination Rules

Theorem 13 (Evidence combination for correlated interior events)**.** *Let e_{byte} be a byte-level evidence event and e_{int} be an interior event derived from the byte stream. The correct evidence combination for predicting output o is NOT Bayesian product but one of:*

1. ***Absorption**: Combine the byte event and interior event into a single joint pattern:*

$$w = s(e_{\text{byte}} \times e_{\text{int}}, o) = \lfloor \log_2 c(e_{\text{byte}}, e_{\text{int}}, o) \rfloor. \tag{7}$$

   This is the correct conditional $P(o \mid e_{\text{byte}}, e_{\text{int}})$ with no double-counting.

2. ***Residual**: Use only the* additional *information that the interior event provides beyond the byte event:*

$$w_{\text{residual}} = s(e_{\text{int}}, o) - s(e_{\text{int}}, o \mid e_{\text{byte}})_{\text{baseline}}. \tag{8}$$

   The residual is the MI that the interior event contributes above the byte-level baseline. For deterministic interior events, this residual is exactly zero when the byte determines the interior event.

3. ***Hierarchical**: Use interior events* only at the interior level*, never mixing them with byte-level evidence for the same prediction:*

$$P(o) = \sum_{w} P(o \mid w) \cdot P(w \mid \text{byte context}), \tag{9}$$

where w ranges over word identities. The byte $\rightarrow$ word mapping ($P(w \mid \text{context})$) uses byte-level patterns; the word $\rightarrow$ output mapping ($P(o \mid w)$) uses interior patterns. No mixing occurs because the two stages condition on disjoint information: the first stage conditions on the raw bytes, the second on the word identity (which absorbs the byte information).

Proof. For (1): Absorption creates a single pattern over the product event space, which is always correct (no independence assumption is needed for a single pattern). The cost is data sparsity: the joint count table is larger.

For (2): The residual information is zero when e_{int} is a deterministic function of the bytes already conditioned on. It is positive only when e_{int} carries information about the output that the byte context does not. This happens when the interior event aggregates information from multiple byte positions (e.g., the bag-of-letters summarizes the entire word seen so far, which no single byte offset captures).

For (3): The hierarchical combination is the chain rule of probability. The byte $\rightarrow$ word stage uses byte-level evidence (independent of interior events). The word $\rightarrow$ output stage uses word-level evidence (which has absorbed the byte information via the recognition patterns). The two stages are independent because the word identity is a sufficient statistic for the byte $\rightarrow$ output prediction at the word level: once you know the word, the individual bytes add no information about the next word (though they do add information about spelling variants, handled by the residual from rule 2). □

Remark 14 (The hourglass enforces independence)**.** *The hourglass structure (bytes $\rightarrow$ words $\rightarrow$ bytes) naturally enforces the hierarchical combination rule: information flows from bytes to word identity (recognition), then from word identity to predicted bytes (prediction). The word-level bottleneck is a sufficient statistic that absorbs byte-level correlations. Evidence that arrives via the word bottleneck is independent of evidence that arrives via direct byte-to-byte patterns (different offsets), because the two evidence streams condition on disjoint aspects of the context.*

7 Which New Patterns Have Value?

Not all new patterns contribute information. We classify them by their expected Δbpc contribution.

Proposition 15 (Deterministic patterns have zero predictive value)**.** *The structural patterns (position increment, accumulator update) are deterministic (weight 255) but carry zero predictive information. They are bookkeeping: they maintain the interior state, but they predict nothing about the output that the byte stream doesn't already determine.*

Proof. The position event $\text{pos}(t) = k$ is a deterministic function of the byte history. Any pattern from $\text{pos}(t)$ to an output o is equivalent to a pattern from the byte history to o. The position event adds no information beyond what is already in the bytes; it merely re-represents existing information in a more accessible form. □

Remark 16 (Value from accessibility, not information)**.** *Although deterministic interior events carry zero* new *information (they are functions of existing events), they can have enormous* computational *value. A UM runner that can read* $\text{pos}(t) = 0$ *directly does not need to trace the byte*

history back to the last word boundary. The accessibility of the position event enables patterns that would otherwise require long-range byte-level pattern chains.

The information is already there; the new event makes it addressable. This is the distinction between Shannon information (bits of surprise) and computational accessibility (addressing cost in the pattern space).

Theorem 17 (Recognition patterns: value from aggregation)**.** *Recognition patterns ($I \to H'$) have positive Δbpc when the interior event aggregates information from multiple byte positions that would otherwise require a high-order pattern.*

Specifically, the bag-of-letters event $\text{bol}_w(t) = 1$ *is equivalent to the conjunction:*

$$\bigwedge_{c \in \text{letters}(w)} (\exists s \leq t : b_s = c \wedge \text{same\_word}(s, t)). \tag{10}$$

Representing this as a single event in H' replaces a pattern of order $|\text{letters}(w)|$ *(requiring* $256^{|\text{letters}(w)|}$ *count table entries) with a single binary event (requiring 1 bit per word per position).*

Proof. Without the bag-of-letters event, recognizing "the" from its letters requires a 3-offset pattern: $(b_{t-2} = \texttt{t}) \wedge (b_{t-1} = \texttt{h}) \wedge (b_t = \texttt{e})$. This is an order-3 pattern with $256^3 = 16.7$M entries in the count table. With the bag-of-letters event, recognition is a single deterministic check: $\text{acc}(t) \supseteq \{\texttt{t}, \texttt{h}, \texttt{e}\}$. The aggregation replaces exponential table size with constant-size event lookup.

The Δbpc from this aggregation is the difference between the order-k and order-1 models for within-word prediction. For "the" ($k = 3$), the order-3 model provides complete within-word prediction; the order-1 model provides only single-byte conditionals. The difference is the within-word MI contribution of the conjunction: empirically $\sim$ 2–4 bits per word occurrence. □

Theorem 18 (Prediction patterns: value from word-level conditioning)**.** *Prediction patterns ($H' \to O$) have positive Δbpc when word identity provides better output prediction than any available byte-level context.*

For common function words, the word identity nearly determines the word-final byte and subsequent space:

$$H(\textit{next byte} \mid \textit{word} = \texttt{the}, \textit{pos} = 3) \approx 0.12 \textit{ bits}, \tag{11}$$

compared to $H(\textit{next byte} \mid b_{t-1} = \texttt{e}) \approx 4.2$ *bits. The Δbpc is $\sim$ 4 bits per word-final position for high-frequency words.*

Theorem 19 (Language-model patterns: value from word-to-word context)**.** *Language-model patterns ($H' \to H'$) provide the largest potential Δbpc because they capture inter-word dependencies that byte-level models can only access through long-range skip-grams.*

The MI between adjacent words in English is $\sim$ 2–6 bits (depending on word frequency and syntactic context). At the byte level, capturing this requires skip-grams at offsets $\geq \bar{L} \approx 5$ bytes apart. At the word level, it is a direct bigram: weight $= \log_2 c(w_{\text{prev}}, w_{\text{curr}})$.

The Δbpc from word bigrams alone, above byte-level unigrams, is bounded by:

$$\Delta \leq \frac{1}{\bar{L}} \cdot \text{MI}(W_{t-1}; W_t) \approx \frac{4}{\bar{L}} \approx 0.8 \textit{ bpc}. \tag{12}$$

8 The Synapse Budget

Definition 20 (Synapse budget)**.** *The* synapse budget *for the extended model is the total number of nonzero patterns:*

$$|P'| = |P_{\text{old}}| + |P_{\text{new}}|, \tag{13}$$

where P_{old} are the byte-level patterns from the base model and P_{new} are the new patterns from the extension.

Proposition 21 (The injection curve)**.** *As words are injected into the vocabulary (in frequency order), the synapse budget grows linearly in $|V|$ (each word adds $O(\bar{L})$ recognition patterns, $O(\bar{L})$ prediction patterns, and $O(B)$ language-model patterns), while the Δbpc per word decreases (Zipf's law: the k-th word has frequency $\sim 1/k$, so its per-byte contribution scales as $\sim \log k/(k \cdot \bar{L})$).*

The injection curve is therefore concave: high-frequency function words ("the", "of", "and") contribute the most per synapse; content words contribute incrementally.

Example 22 (First word: "the")**.** *Injecting "the" ($f \approx 0.07$, $\bar{L} = 3$) adds:*

- $\sim 3 \times 256 = 768$ *recognition patterns (byte × position → accumulator, most with weight 0).*
- $\sim 3 \times 256 = 768$ *prediction patterns (word × position → next byte).*
- ~ 50 *language-model patterns (word-to-word transitions with nonzero count for the top 50 preceding words).*
- ~ 3 *structural patterns (position increment within "the").*

Total: ~ 1600 new patterns, of which ~ 100 have significant weight. The Δbpc from recognition + prediction: at 7% frequency, "the" occupies $\sim 7\% \times 3/\bar{L}_{\text{global}} \approx 4.2\%$ of bytes. At ~ 4 bits Δbpc for word-final prediction, the contribution is $\sim 0.04 \times 4 = 0.16$ bpc on those positions, or ~ 0.007 bpc overall.

This matches the order of magnitude of the neutral-factorization experiments [4]: "the" alone contributes modestly, but the infrastructure (position, accumulator, word identity) enables all subsequent word injections to share the same recognition and prediction machinery.

9 Independence Classes in the Extended Space

Theorem 23 (Three independence classes)**.** *The evidence sources in the extended model partition into three independence classes:*

1. ***Byte-level offsets*** *(the original $I \to O$ patterns): independent when edges are vertex-disjoint in the offset graph [6].*
2. ***Word-level predictions*** *($H' \to O$): independent of byte-level predictions at* different *words, but correlated with byte-level predictions at the* current *word position.*
3. ***Cross-word context*** *($H' \to H'$): independent of within-word byte patterns (the word boundary is a conditional independence barrier), but correlated with word-level predictions at the current word (shared word identity).*

Proof. For (1): this is the offset graph theorem [6], unchanged by the extension.

For (2): the word identity w absorbs all within-word byte information (it is a sufficient statistic for the canonical spelling). A prediction pattern (word $= w$) $\to$ ($o =$ ' ') and a byte-level skip-bigram $(b_{t-7}, b_{t-1}) \to (o)$ are approximately independent when the skip-bigram's offsets fall outside the current word (they condition on bytes from a different word, hence a different part of the microstate space). But if the skip-bigram's offsets fall *within* the current word, they share information with the word identity and are dependent.

For (3): the word boundary acts as a Markov blanket: conditioned on the current word identity, the bytes within the current word are independent of the bytes within the previous word. Therefore, cross-word patterns (previous word → current word) are independent of within-word byte patterns. □

Corollary 24 (Safe combinations)**.** *The following evidence combinations are safe (independent, no double-counting):*

- *Word-level bigram + within-word byte pattern (different information domains: inter-word vs. intra-word).*
- *Offset-disjoint skip-bigrams (from the base model).*
- *Word identity at the current position + skip-bigram from a* previous *word's bytes.*

The following combinations are UNSAFE (dependent, will double-count):

- *Word identity + byte-level pattern using bytes from the current word (shared information source).*
- *Bag-of-letters + individual letter-accumulator events (the BoL is a deterministic function of the accumulator).*
- *Position event + byte-level "distance from word boundary" pattern (identical information, different representation).*

10 Predictions

Prediction 25 (Recognition patterns are deterministic and free)**.** *All recognition patterns (byte → accumulator, accumulator → bag-of-letters) will have weight 255. They add zero information but provide accessibility. Implementing them costs nothing at the data level (no counting required) and provides the addressing infrastructure for all subsequent patterns.*

Prediction 26 (First non-trivial value from word-final prediction)**.** *The first measurable Δbpc from the extension will come from prediction patterns at word-final positions: "word is w and position is $|w| \to$ next byte is space". This is where the word identity provides maximum information gain over byte-level context. Expected: $\sim$ 0.1–0.2 bpc for the top 100 words, concentrated at word-final positions.*

Prediction 27 (Word bigrams dominate at scale)**.** *As the vocabulary grows beyond $\sim$ 500 words, language-model patterns (word-to-word bigrams) will contribute more Δbpc than recognition or prediction patterns. This is because inter-word MI ($\sim$ 4 bits/word) exceeds the within-word MI that recognition captures ($\sim$ 2 bits/word), and the vocabulary growth unlocks more word-bigram patterns.*

Prediction 28 (The injection curve is concave with knee at $\sim$ 200 words)**.** *The cumulative Δbpc as a function of vocabulary size $|V|$ will be concave (diminishing returns per word). The "knee" where marginal Δbpc drops below 10^{-4} bpc per word will occur at approximately 200 words, which covers $\sim$ 60% of running text (function words + top content words). Beyond this knee, content words add context but not raw prediction quality.*

Prediction 29 (Hierarchical combination matches or beats absorption)**.** *The hierarchical combination rule (Section 6, rule 3) will achieve Δbpc within 0.01 bpc of absorption (rule 1), while being computationally cheaper (no high-dimensional count tables). This is because the hourglass bottleneck (word identity) is a near- sufficient statistic for the byte $\rightarrow$ output mapping at the word level.*

Prediction 30 (The independence barrier at $\sim$ 1 bpc)**.** *At byte-level KN-6 with $\sim$ 1.8 bpc [4], adding word-level patterns should reduce to $\sim$ 1.0–1.2 bpc. Below 1 bpc, the remaining information requires patterns that violate the independence assumptions of the current framework (e.g., syntactic dependencies that span multiple words and cannot be captured by word bigrams alone). Breaking through 1 bpc will require either trigram word patterns (expensive in data) or a second nesting level (sentences as events inside the word-level H').*

11 Discussion

11.1 The synapse is what counts

The nested model paper [2] and the tokenization-loss paper [3] established the event-space architecture: which events exist, why they are necessary, and what information they carry. This paper establishes the complementary result: the events are inert without patterns, and the patterns must respect the independence structure of the evidence.

The atomic pattern $(e_{\text{from}}, e_{\text{to}}, w) \in E^2 \times T$ is the synapse of the UM. Its weight w is the log of the joint count—the empirical evidence that these two events co-occur. The forward pass is a search for the best syllogism, and the conclusion's support is bounded by the weakest link. This is the content of the forward pass as existential quantification [8].

11.2 Correlation as the price of richness

The extended event space is richer than the base model (more events, more potential patterns), but this richness comes with a cost: correlated events that cannot be combined by naïve Bayesian product. The correlation problem is not a bug—it is the *consequence* of interior events being deterministic functions of the byte stream.

The resolution is structural: the hourglass architecture enforces independence by channeling information through the word-identity bottleneck. Evidence that flows through the bottleneck (word-level predictions) is independent of evidence that flows around it (direct byte-to-byte patterns at distant offsets). The bottleneck is the conditional-independence barrier that makes hierarchical combination valid.

11.3 From the RNN to the extended UM

The trained RNN solves the combination problem implicitly: the W_h matrix combines information from all offsets via matrix multiplication, which is the correct multivariate conditional [6]. But the RNN's solution is opaque—we cannot read off which evidence sources are independent and which are correlated.

The extended UM solves the same problem explicitly: the pattern families classify evidence by independence class, and the combination rules specify exactly when Bayesian product is valid and when absorption or hierarchical combination is required. The result is an interpretable architecture where every prediction is a traceable syllogism.

References

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

[2] Claude and MJC. *The Nested Model: Self-Similar Architecture in the Extended Event Space*. Hutter archive, 15 Feb 2026.

[3] Claude and MJC. *Tokenization as Information Loss*. Hutter archive, 15 Feb 2026.

[4] Claude and MJC. *The Extended Event Space: Injecting Lexical Structure into H*. Hutter archive, 15 Feb 2026.

[5] Claude and MJC. *The Model Within the Model: Commentary on the Extended Event Space*. Hutter archive, 15 Feb 2026.

[6] Claude and MJC. *Conditional Independence on the Offset Graph*. Hutter archive, 12 Feb 2026.

[7] Claude and MJC. *Bayes from Counting: Partial Quotients, GCD, and the Symmetric Learning Function on $E = I \times O$*. Hutter archive, 12 Feb 2026.

[8] Claude and MJC. *Logic from Counting: Existential Quantification, Probabilistic Syllogisms, and the Derivation of Formal Inference from the Universal Model*. Hutter archive, 12 Feb 2026.

[9] Claude and MJC. *Expressiveness and Limits of the Tropical Forward Pass*. Hutter archive, 12 Feb 2026.

[10] Claude and MJC. *The Tock Step: Domain-Native Architecture from Evidence*. Hutter archive, 12 Feb 2026.