The Combination Problem:
Three Layers, One Benchmark

Claude and MJC

March 12, 2026

Abstract

The Universal Model's standard forward pass $f_0(t)_j = \max_i \min(t_i, p_{ij})$ operates in the $(\max, \min)$ tropical semiring. Applied to order-6 n-gram event spaces on enwik9, it produces 5.0 bpc—while Kneser–Ney interpolation over the *same* statistical data yields 2.2 bpc. The 2.8 bpc *combination gap* is the central open problem in Universal Model research. We formalize three ways the gap can be pressured. The key empirical break was gap-weighted tropical blending: it reached 2.559 bpc at 100K and 2.244 bpc at 1M, beating KN at both tested scales while remaining native to support-space competition. A newer H3 normalized-conviction hybrid now pressures the picture further at higher orders, but this paper still treats gap-blend as the first decisive tropical-native challenge to the old framing.

We therefore formalize three ways the gap can be pressured:

Approach A replaces the forward pass. The tropical max discards sub-optimal evidence; weighted interpolation retains it. Replacing max with logsumexp or KN-style backoff interpolation closes the gap immediately, but abandons the algebraic properties (exact conditionals, monotonicity, fixed-point convergence) that make the UM a UM.

Approach B enriches the event space. Keep f_0 but add *sharpness observation* as an event, enabling the UM to select the most informative source via max-min at a higher level. The sharpest-LPP heuristic (selecting the source with the largest support gap $s_1 - s_2$) already reduces 4.9 to 2.0 bpc on the bigram model—a 60% closure of the gap—without changing f.

We then show that the real decision problem has three layers: the forward pass f, the event space E, and the competition policy. The correct resolution criterion is not elegance alone or prediction alone, but total lossless compressed size under a shared benchmark, with a working decoder and explicit model-cost accounting. MCP compatibility remains mandatory context, but not veto power.

1 The Gap

Definition 1 (Combination gap). *For a UM with event spaces $\{E_i\}$, per-LPP output distributions $\{d_i\}$, and output entropy H:*

$$\Delta_{comb} = H_{\max\text{-}\min} - H_{interp} = bpc(f_0) - bpc(f_{KN}). \tag{1}$$

Proposition 2 (The gap is structural, not statistical). *The gap does not decrease with more data. At 100K, 1M, and 10M bytes, $\Delta_{comb} \approx 2.8$ bpc consistently. The gap is between two* combination operators *applied to the same evidence, not between two levels of statistical precision.*

2 Approach A: Replace the Forward Pass

Definition 3 (Interpolating forward pass)**.** *Replace f_0 with the KN-interpolated forward pass:*

$$f_{KN}(t)_o = \sum_{k=1}^{K} \gamma_k(t) \cdot P_k(o \mid context_k), \tag{2}$$

where γ_k are backoff weights and P_k are per-order conditionals derived from the same count tables that f_0 uses.

Proposition 4 (What A gains)**.**

1. *Closes the gap immediately: 5.0 → 2.2 bpc.*
2. *Uses the same data (same LPPs, same counts).*
3. *Is a proven, well-understood algorithm (KN smoothing).*

Proposition 5 (What A loses)**.**

1. ***Exact conditionals** (graf 36): the tropical–integer bridge guarantees that f_0's conditionals equal exact Bayes. Interpolation introduces approximation at the combination step.*
2. ***Monotonicity**: f_0 is monotone ($t' \geq t \Rightarrow f_0(t') \geq f_0(t)$). Interpolation with adaptive backoff weights is not necessarily monotone.*
3. ***Fixed-point convergence** (graf 32): the UM's settling behavior depends on f_0 being a contraction in the tropical lattice. Interpolation does not preserve this.*
4. ***The UM identity**: if f is not $(\max, \min)$, the algebraic semantics (graf 30), the counting monad, and the compression–prediction duality (graf 31) all need re-derivation.*

3 Approach B: Enrich the Event Space

Definition 6 (Sharpness event space)**.** *For each LPP ℓ with output distribution d_ℓ, define the* sharpness *event:*

$$sharp(\ell) = s_1(\ell) - s_2(\ell), \tag{3}$$

where $s_1 \geq s_2 \geq \cdots$ are the sorted support values in d_ℓ. The sharpest LPP is $\ell^* = \arg\max_\ell sharp(\ell)$.

Proposition 7 (The sharpest-LPP result)**.** *Selecting the distribution from the sharpest LPP (highest $s_1 - s_2$) instead of combining all LPPs via max-min:*

Model	*max-min*	*sharpest-LPP*
Bigram (4K)	*4.883*	*1.955*
Bigram (64K)	*4.985*	*3.304*

The sharpest-LPP heuristic closes 60% of the combination gap on 4K and 34% on 64K, without changing f_0.

Proposition 8 (What B gains)**.**

1. *Preserves all tropical algebraic properties.*
2. *Uses max-min at a* meta-level*: selecting among LPPs is itself a max operation over sharpness events.*

3. *Is consistent with the UM's self-modeling: sharpness is an observable property of the model's own predictions, hence a legitimate event.*

Proposition 9 (What B loses)**.** *1.* ***Selection discards sub-optimal sources****: the sharpest LPP may be confidently wrong. Interpolation hedges; selection does not.*

2. ***The settling negative result****: iterative settling (repeatedly applying sharpness selection) has no effect because max-min is idempotent. One-shot selection works; iteration does not. This suggests the approach is a heuristic, not a principled fixed point.*

3. ***The 64K degradation****: sharpest-LPP closes only 34% of the gap at 64K vs. 60% at 4K. As models grow more complex, the heuristic weakens.*

4. ***No interpolation of evidence****: two LPPs may each be partially informative. Selection picks one; interpolation combines both. The combined evidence is strictly better than either alone (data processing inequality).*

4 Tension, Not Yet Resolution

Proposition 10 (A and B pressure different layers)**.** *Let f_A be the interpolating forward pass and f_B be the sharpest-LPP selector. Then:*

1. *f_A primarily changes Layer 1 (the combination rule itself).*

2. *f_B primarily changes Layer 3 (which source is allowed to speak).*

3. *Because they operate at different layers, naive comparisons of their bpc alone do not identify which change carried the gain.*

So the practical question is not whether one can write down a hybrid, but whether a controlled ablation shows that changing f is necessary after competition and event-space effects are isolated.

5 The Ablation Matrix

Both the performance-first and the theory-first positions demand controlled evidence before committing. The following matrix provides it.

Definition 11 (Ablation matrix)**.** *Run order-6 n-gram UM on enwik9 prefixes with five policies, reporting the same benchmark columns per policy:*

Proposition 12 (Resolution conditions)**.** *1. If row (b) closes ≥80% of the gap between (a) and (c) at equal model cost, source selection suffices and Approach B wins.*

2. *If row (b) leaves ≥1.0 bpc gap to row (c) at equal model cost, and row (d) does not close it, Approach A is needed.*

3. *If row (d) matches row (c) within 0.1 bpc, the entropy-weighted blend is a viable compromise: it uses sharpness (Approach B's observable) as the interpolation weight (Approach A's operation).*

Policy	frozen bpc	model bytes	AC bytes	total bytes	decode exact?	train cost	infer cost
(a) max-min (f_0)	?	?	?	?	Y	?	?
(b) sharpest-LPP	?	?	?	?	Y	?	?
(c) KN-interp	?	?	?	?	Y	?	?
(d) entropy-weighted blend	?	?	?	?	Y	?	?
(e) gap-blend ($w_k = 2^{s_1 - s_2}$)	?	?	?	?	Y	?	? (f) H3 normalized conviction ($w_k = 2^g$
?	?	?	?	Y	?	?	

Table 1: The ablation matrix. Every row must report total bytes under exact round-trip decoding, plus training and frozen inference cost. MCP compatibility is audited alongside the table, not hidden inside it.

4. *If row (f) consistently beats row (e) while preserving the same decoder and model-cost accounting, then normalized conviction supersedes plain gap-blend as the leading higher-order tropical-native competition law in the tested regime.*

Remark 13 (The MCP audit)**.** *For each row that changes f or otherwise alters the UM algebra, the paper must list:*

- ***Survives**: graf holds as stated under this policy.*
- ***Restated**: graf holds with a modified hypothesis (e.g., "under f_0" becomes "under f_{KN}").*
- ***Fails**: graf's conclusion is false under this policy.*

This forces the algebraic cost to be explicit, preventing the performance-first camp from ignoring theory and the theory-first camp from comparing theory to numbers dishonestly.

Remark 14 (Why the matrix settles the debate)**.** *The matrix answers both camps' demands simultaneously. The performance-first position ("show me the numbers") gets the full compression table. The theory-first position ("show me the algebraic cost") gets the MCP audit. No one can hide behind one criterion while ignoring the other.*

The debate is currently under-instrumented. Filling in the ?'s converts rhetoric into data. The resolution will come from the numbers, not from argument.

6 The Three-Layer Decomposition

The A/B framing above collapses three independent layers:

The latest ablation order sweep (orders 4, 6, 8 on 1M bytes) shows ent-blend still beats H3 at order 4 (2.424 vs 2.479) while H3 leads gap-blend at orders 6 and 8 (2.234 vs 2.241; 2.183 vs 2.291). This is the numerical basis for calling H3 "the leading higher-order hybrid so far." The JSON record is in ‘docs/archive/20260312/ablation-data.json‘ and the sweep is captured in ‘hybrid$_blend_sweep$‘.

Definition 15 (Three layers of the combination problem)**.** *1.* ***Layer 1: The forward pass** f**.*** *The core algebraic operation. Currently $f_0 = (\max, \min)$. Changing this changes the UM's mathematical identity.*

2. ***Layer 2: The event space** E**.*** *What is observable. Adding sharpness, word identity, or other meta-events enriches E without changing f.*

3. ***Layer 3: The competition policy*** π. *How to select among multiple sources (LPPs) when each provides a distribution over the output. Sharpest-LPP is one policy; flat max-min aggregation is another; interpolation is a third; and gap-weighted tropical blending is a fourth; H3 normalized conviction is a fifth.*

Proposition 16 (Sharpest-LPP is primarily a Layer 3 change)**.** *The sharpest-LPP heuristic (4.9→2.0 bpc) changes only the competition policy, not the forward pass or the event space. It selects one LPP's output distribution instead of combining all via* f_0*. The core* $(\max, \min)$ *algebra is preserved within each LPP.*

Remark 17 (The real question)**.** *The three layers decompose the problem but do not resolve it. They reveal a deeper question:* ***what is the resolution criterion when performance and algebraic identity conflict?***

- *The* performance criterion *says: minimize total compressed bytes (frozen bpc + model cost + AC overhead). If interpolation wins, use interpolation; algebraic properties are desirable but not constraining.*

- *The* identity criterion *says: any solution must be expressible within the UM's algebraic framework. Competition policies that operate outside* f_0 *must be shown to be derivable from* f_0 *applied to richer event spaces; otherwise the UM is merely a data structure, not a mathematical theory.*

These criteria conflict when the best-performing competition policy is not derivable from f_0 *on any event space.*

7 Resolution: Two Tracks, One Codebase

The combination gap admits two valid closures depending on what object is under study. This is not a compromise; it is a genuine ontological distinction.

7.1 Track U: The Universal Model

[Constitutive identity] The tropical max-min forward rule f_0 is not a design choice but an identity condition of the Universal Model. Changing f does not improve the UM; it builds a different system.

Under Track U, the combination problem is pursued within the tropical semiring: richer event spaces, sharpness observation, ES epistemics, and tropical-native competition rules such as gap-weighted blending are admissible tools. Results on this track are UM theorems—they validate MCP grafs 30, 32, 37 and inherit the full algebraic apparatus.

- Competition/selection remains the live bottleneck until the ablation matrix shows otherwise.

- MCP compatibility is *binding*, not advisory, for claims about the UM itself.

- Changing f triggers *relabeling*: no silent inheritance of existing algebraic results.

7.2 Track C: The Compressor

[Total-byte acceptance] When theory and compression conflict, the winner is $\arg\min_{\text{method}}$ total lossless bytes under a shared benchmark, exact round-trip decoding, and honest model-cost accounting.

Under Track C, the combination gap is closed by whatever policy minimizes total bytes: KN interpolation, entropy-weighted blending, gap-weighted tropical blending, or hybrid schemes (Approach A). Results on this track are engineering wins—they do not inherit MCP support.

- MCP matters as a design prior and explanatory constraint, but does not have veto power over a method that wins on total bytes.
- Every Track C result must state explicitly that it is a compressor-only result, not a UM theorem.
- The MCP grafs that depend on f_0 (algebraic semantics, fixed-point structure, tropical-GCD bridge) are *not invalidated* by Track C results—they simply apply to a different object.

7.3 The Boundary

Theorem 18 (Incompatibility)**.** *No single resolution can simultaneously:*

1. *close the full combination gap by using only rigid flat max-min aggregation,*
2. *preserve one unchanged MCP interpretation layer, and*
3. *maintain that every compressor win is automatically a UM theorem.*

At least one of these must be abandoned. Gap-weighted blending shows that the old theorem was too strong: the tropical semiring still has room to move, but rigid max-min no longer exhausts its admissible competition laws.

The two-track resolution abandons (3): the project is building two systems that share code, data, and most of their event-space infrastructure. Track U owns the algebra. Track C owns the benchmark. Neither claims the other's results.

8 Discussion

The combination problem is not a technical detail. It is the question of whether the Universal Model's mathematical identity (the tropical semiring) is compatible with its empirical requirements (competitive prediction). The answer, as crystallized by adversarial pressure between the algebraic and performance camps, is: *they are studying different objects.*

A compressor that replaces max-min with interpolation and achieves lower total bytes has not "solved the combination problem for the UM." It has built a better compressor that is not a UM. Conversely, a UM that closes the gap via richer event spaces and sharpness selection has not "beaten interpolation." It has validated a mathematical theory at the cost of engineering simplicity.

The 2.8 bpc gap marks the boundary between these two research programs. It is the price of the tropical semiring. Track U pays it willingly, betting that the algebraic structure will yield deeper understanding. Track C refuses to pay it, betting that total bytes is the only metric that matters. Both bets are legitimate. Neither subsumes the other.

The next real artifact is not another paper about which approach is "right," but a filled ablation sheet that quantifies exactly what each track gains and what it costs.

References

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

[2] Claude and MJC. *Settling: Sharpness Preference and ES Epistemics*. Hutter archive, 19 Feb 2026.

[3] Claude and MJC. *The Timing Resolution*. Hutter archive, 19 Feb 2026.

[4] Claude and MJC. *N-gram UM Achieves KN-6 Parity*. Hutter archive, 19 Feb 2026.

[5] Claude and MJC. *The Tropical–Integer GCD Bridge*. Hutter archive, 12 Feb 2026.

[6] Graf. *The Conviction–Accuracy Tradeoff: Why Support-Gap Blending Beats Entropy Weighting*. Hutter archive, 12 March 2026.

[7] Graf. *Conviction Depth: Fewer Wins, Bigger Wins*. Hutter archive, 12 March 2026.

[8] Claude and MJC. *Algebraic Semantics of the Universal Model*. Hutter archive, 12 Feb 2026.