Multi-Retro: Extending the UM Crossover via Repeated Retroactive Passes

Claude and MJC

March 2026

Abstract

The UM KN chain with retroactive training beats external KN-6 below ∼150K but loses above due to the frozen scoring penalty. We test whether multiple retroactive passes extend this crossover point. At 100K ($\tau = 1$, order 6): 1 pass = 2.628, 3 passes = 2.605, 5 passes = 2.600 bpc (KN-6 = 2.719). At 200K: 3 passes = 2.599 (KN-6 = 2.511, gap +0.088). The improvement is real but diminishing: each additional pass gives ∼5–10 ms/position. Multi-retro does not fundamentally change the scaling picture.

1 Multi-Retro Protocol

The standard UM KN chain runs three passes:

1. **Online**: learn structure + weights from data.
2. **Retroactive**: freeze structure, re-accumulate counts over the full dataset. Each neuron sees 2× counts.
3. **Frozen**: freeze all weights, score.

Multi-retro repeats step 2 k times before step 3. After k retroactive passes, each neuron has accumulated $(k + 1)\times$ counts (1 from online + k from retro). More counts should produce better-calibrated KN distributions.

2 Results at 100K

Retro passes	Last retro bpc	Frozen bpc	Δ vs 1-pass
1	2.785	2.628	—
3	2.627	2.605	−0.023
5	2.612	2.600	−0.028

External KN-6 at 100K: 2.719. All UM results beat KN-6. The retro-to-frozen gap narrows with more passes (0.157 → 0.012), suggesting the distribution is converging.

3 Results at 200K

Retro passes	Frozen bpc	Ext. KN-6	Δ
1	2.610	2.511	+0.099
3	2.599	2.511	+0.088
10	TBD	2.511	TBD

Multi-retro reduces the gap at 200K from +0.099 to +0.088 with 3 passes, but the crossover is not fundamentally moved.

4 Scaling Curve ($\tau = 1$, order 6, 3 retro passes)

Size	UM frozen	Ext. KN-6	Δ
100K	2.605	2.719	−0.114
200K	2.599	2.511	+0.088
300K	TBD	2.568	TBD
500K	TBD	2.489	TBD
1M	TBD	2.397	TBD

5 Diminishing Returns

Each retro pass gives diminishing improvement:

Pass	100K retro bpc	Δ from prev
1	2.785	—
2	2.652	−0.133
3	2.627	−0.025
4	2.617	−0.010
5	2.612	−0.005

The series is geometric: each pass adds ∼40% of the previous gain. At 100K, infinite passes would converge to ∼2.595 bpc (extrapolating). The frozen score tracks the retro score with a small offset (∼0.012 at convergence).

6 Implications

Multi-retro is a cold-start correction. Like the single retroactive pass, multiple passes help most when data is scarce. The gain at 100K (−0.028 for 5 passes) is meaningful; at 200K (−0.011 for 3 passes) it is marginal.

The crossover is structural. The UM's frozen scoring penalty grows with data size because the global distribution becomes a worse local fit on non-stationary data. Multi-retro doesn't fix this: it improves counts but doesn't change the scoring mechanism.

Path forward: online scoring. The fundamental fix is to score online (not frozen). Online KN distributions use local counts and don't suffer the frozen penalty. The remaining gap between online UM and external KN is the global unigram base (the UM's byte counts are global, not position-local).