Threshold Scaling Degradation in the UM KN Chain

Claude and MJC

2026-03-13

1 The Hypothesis

At 100K, lowering the threshold τ from 4 to 1 improves the UM's frozen KN score: τ=4 gives 2.749 bpc, τ=2 gives 2.656 bpc, τ=1 gives 2.605 bpc (3 retro passes). The natural hypothesis: τ=1 should scale better because it creates neurons for every observed context, matching external KN's coverage.

This hypothesis is wrong. At larger scales, τ=1 *degrades* while external KN improves.

2 Results

Scale	UM τ=1	UM τ=4	KN-6 frozen	Gap τ=1	Gap τ=4
100K	2.605	2.749	1.265	+1.340	+1.484
300K	2.778	2.813	1.272	+1.506	+1.541
500K	2.920	2.890	1.291	+1.629	+1.599
1M	3.101	3.054	1.335	+1.766	+1.719

Table 1: UM frozen KN (order 6, 3 retro passes) vs external KN-6 frozen. Both τ settings degrade with scale; τ=4 better from 500K onwards.

Key observations:

1. UM τ=1 *degrades* with scale: 2.605 → 2.778 → 2.920 → 3.101.
2. UM τ=4 also degrades: 2.749 → 2.813 → 2.890 → 3.054.
3. At 500K, τ=4 overtakes τ=1 (2.890 vs 2.920): the crossover persists at 1M (3.054 vs 3.101).
4. KN-6 frozen is roughly flat: 1.265 → 1.272 → 1.291 → 1.335.
5. The gap *grows* for both: τ=1 gap 1.340 → 1.766; τ=4 gap 1.484 → 1.719.

3 Root Cause: Sparse Context Dilution

With τ=1, every observed context creates a neuron. At order 6, this creates 5gram contexts that fire only 1–2 times in the data. The KN discount from these sparse contexts steals probability mass and redistributes it according to the backoff chain.

At 100K, most contexts are sparse anyway, so τ=1's extra coverage helps. But at 300K+, the dataset has enough repetition that τ=4 contexts are well-calibrated, while τ=1 contexts remain sparse because the number of unique 5/6-grams grows faster than observations.

The UM's cascading threshold makes this worse: a 5gram neuron requires its 4gram parent to exist and fire. But at τ=1, parents are created immediately, so children are created at the first possible moment—giving them almost no observations for calibration.

4 The Ratio Problem

External KN-6 stores counts in a flat hash table. Every n-gram context has a count proportional to its frequency in the data. The KN discount D is calibrated to this frequency.

The UM stores counts in LPP entries. But because neurons are created through cascading threshold, high-order entries are *front-loaded*: many entries are created early (when the parent first appears) and then accumulate counts slowly. With τ=1, this front-loading is maximal—every context gets an entry at first observation.

The result: at position 300K, a 5gram context might have been observed 2 times, but the UM created it at position 1000. External KN would have all observations of that context (say 5 times) from a single pass.

5 Implications

1. τ=1 **is not the answer.** Matching external KN's coverage requires not just creating neurons but giving them sufficient observations for calibrated distributions.
2. **The retroactive pass helps but is insufficient.** 3 retro passes at 100K: 3.203 → 2.627 → 2.652 → 2.627 → 2.605. At 500K: 3.189 → 2.968 → 2.930 → 2.928 → 2.920. Diminishing returns that don't close the gap.
3. **The fundamental bottleneck is representation, not coverage.** The UM's cascading threshold creates contexts with wrong count ratios. More retro passes help but can't fix the structural mismatch.
4. **Two viable paths forward:**
   - Use external KN-6 directly for compression (abandon UM prediction).
   - Use UM for interpretability/structure, external KN for compression.

6 Comparison with Previous Claims

The `#order_scaling_corrected_paper` reported that τ=1 at 100K beats τ=4 (2.628 vs 2.749). This is true at 100K. But the result does not extrapolate: at 500K, τ=1 gives 2.920 while τ=4 gives 2.890, and at 1M the gap widens further (3.101 vs 3.054). The crossover occurs between 300K and 500K and persists at all larger scales. Both settings degrade, but τ=1 degrades faster due to sparse context dilution at high orders.

Combined with the scoring protocol correction (UM frozen 2.605 vs KN-6 frozen 1.265 at 100K), the UM's KN chain is fundamentally limited by its cascading threshold architecture. No τ setting closes the gap; they merely trade coverage for calibration.