H3 Pipeline Integration: Conviction-Weighted Meta-Combination and the Scale Convergence

Vector

2026-03-12

Abstract

We show that blending per-order Kneser–Ney chain probabilities with $2^{g/H}$ weights (H3 meta-combination) beats standard KN-6 interpolation on enwik9 at 100K (-0.079 bpc) and 1M (-0.045 bpc), but *converges to parity at 10M* ($+0.000$ bpc). Three integration architectures were tested; only meta-combination of complete KN chains succeeds. The advantage is scale-dependent: H3 adds value when orders differ in quality (small data), but vanishes when all orders are well-calibrated (large data). This suggests H3 is a *cold-start correction* rather than a fundamental improvement to KN.

1 Background

The H3 combination rule $w_k = 2^{g_k/H_k}$ was discovered via ablation experiments on the pure ngram UM model (§20260312 papers). There, it beat KN interpolation at orders ≥ 6 but trailed at orders ≤ 4. The open question: does it transfer to the real KN-6 pipeline?

2 Three Integration Attempts

We tested three architectures for integrating H3 into the KN-6 pipeline. All use the same hash table and learning.

2.1 Flat blend of raw distributions (FAILED)

Materialize the raw KN distribution at each order k: $p_k(b) = \max(c_{kb} - D, 0)/tc_k$, normalize, compute g_k and H_k, weight by $2^{g_k/H_k}$, average.

Result: +0.713 bpc at 100K (catastrophic). The unigram's weight $w_0 \approx 1$ dilutes everything.

2.2 Conviction-modulated backoff (FAILED)

Replace KN's fixed $\lambda = D \cdot ty/tc$ with $\lambda_{H3} = \lambda_{KN}/(1 + \text{conviction})$.

Result: +0.486 bpc at 1M. Reducing λ removes probability mass that KN needs for smoothing.

2.3 Meta-combination of KN chains (SUCCESS at small scale)

Compute $p_k^{\text{KN}}(\text{actual})$ = standard KN interpolation from order 0 up to order k. Then blend across orders:

$$p_{H3} = \frac{\sum_{k=0}^{6} w_k \cdot p_k^{\text{KN}}}{\sum_{k=0}^{6} w_k}, \quad w_k = 2^{\lfloor g_k / H_k \rfloor}$$

3 Results

Method	**100K**	**1M**	**10M**
KN-6 interpolation	2.719	2.398	2.178
H3 flat blend	3.433	—	—
H3 modulated backoff	—	2.883	—
H3 meta-combination	**2.641**	**2.353**	2.178
Δ (H3 − KN-6)	−0.079	−0.045	+0.000

Table 1: H3 advantage shrinks with scale and vanishes at 10M.

The full pipeline at 10M (KN-6 + sparse + match) reaches 1.954 bpc. H3 provides no additional gain when combined with sparse and match.

4 The Scale Convergence

The advantage curve $(-0.079, -0.045, +0.000)$ is monotonically decreasing. This is not a sampling artifact—it reflects a real structural phenomenon:

- **Small data**: orders differ sharply in quality. Order 6 may have 2 observations for a context while order 3 has 200. H3 correctly downweights the sparse high-order prediction. KN's fixed λ chain cannot adapt to this.
- **Large data**: all orders are well-populated. KN's interpolation chain, tuned by the discount D, is near-optimal. H3's conviction weights become approximately uniform, adding no information. The 256× compute overhead buys nothing.

5 Interpretation: Cold-Start Correction

H3 meta-combination is a *cold-start correction*: it helps when the model hasn't seen enough data for KN's fixed discount to work well. As data grows, the correction vanishes.

This is consistent with the ablation results. The ablation tested at 100K–1M, where H3 won. The crossover between KN and H3 happens between orders 4 and 6 at 100K—exactly where context sparsity becomes severe.

6 Architectural Lessons

1. **Don't modify KN's internals.** The interpolation chain is not a trust knob. λ carries discounted mass; reducing it creates coverage holes.

2. **Compose complete predictors.** Meta-combination of KN chains works; internal modification doesn't.

3. **Conviction is a cold-start signal.** Useful when predictors differ in quality; irrelevant when all are mature.

4. **The discount IS the combination rule.** At scale, KN's $D = 0.9$ already implements near-optimal order combination. H3 is redundant.

7 Implications

The path to beating KN-6+sparse+match on enwik9 does not go through H3 meta-combination. The sparse and match models already provide independent signals that compose well with KN's interpolation. H3 provides a smoothing benefit at small scale that KN's chain already provides at large scale.

The combination problem (§20260312) remains open. The tropical tax is real (2.9 bpc max-min vs. KN interpolation), and H3 does not solve it at scale—it only masks it during cold-start.