Word-Level KN Mixture: Scaling from 100K to 1M

Claude and MJC

2026-03-13

Abstract

We test whether word-level structure improves external KN-6 compression at two scales (100K and 1M bytes of enwik9). The word-KN6 mixture adds a word-conditional distribution, mixing it with character-level KN-6 via log-linear interpolation. The gain decreases with scale (+0.111 at 100K, +0.060 at 1M) as KN-6's high-order contexts implicitly capture some word structure. The gain concentrates at word boundaries and in high-frequency words.

1 Method

The `word-kn6` command runs KN-6 on character bytes while simultaneously tracking word boundaries (spaces delimit words). At each position inside a word, the model computes:

- KN-6 distribution (character n-gram, order 6)
- Word-conditional distribution (empirical distribution of bytes following the current prefix within the current word)

The mixture uses log-linear interpolation: the log-probabilities are combined with a learned mixing weight α, optimized online.

2 Results

2.1 Overall Scaling

Scale	KN-6 alone	+Word mix	Gain	Mixing α
100K	2.719	2.608	+0.111	−1.011
1M	2.397	2.338	+0.060	—

The gain halves from 100K to 1M. KN-6 improves by 0.322 bpc (2.719 → 2.397) while the mixture improves by 0.270 (2.608 → 2.338). The word component's marginal value shrinks because KN-6's 5gram and 6gram contexts already capture word-internal regularities at larger scale.

2.2 Per-Category Breakdown

Category	KN-6 (100K)	Gain (100K)	KN-6 (1M)	Gain (1M)
Top-100 words	1.877	+0.242	1.434	+0.113
Other words	3.121	+0.102	2.804	+0.072
Non-words	2.422	+0.000	2.290	+0.000

Non-word positions (numbers, markup, whitespace between words) get zero gain by construction—the word mixture only activates inside words. Top-100 words capture the majority of the per-position gain because they have the most training data for the word-conditional distribution.

2.3 Per-Order Breakdown (1M)

The order-by-order results show where word structure helps most:

Order	KN-6	Mix	Gain	Positions
1	2.450	2.373	+0.077	136K
2	2.284	2.283	+0.001	113K
3	1.705	1.685	+0.020	89K
4	1.455	1.453	+0.002	70K
5	1.408	1.404	+0.004	55K
6	1.386	1.384	+0.002	43K
7+	1.378+	1.375+	+0.002	30K−

Order 1 (first byte of word after space) gets the largest gain (+0.077) because KN-6 has the weakest context there (only the space and 5 preceding bytes). Higher orders have diminishing gains because KN-6 already captures most word-internal structure.

3 Implications

1. **Word structure is worth ∼0.06 bpc at 1M scale.** This is modest but real, and concentrated at word boundaries.

2. **The gain decreases with scale** because KN-6's high-order contexts implicitly learn word patterns. At 10M+, the marginal value of explicit word tracking may be negligible for KN-6.

3. **For the UM**, word-level structure is more valuable than for KN-6, because the UM's cascading threshold bottleneck limits high-order context coverage. Word events provide orthogonal information that the UM cannot derive from byte n-grams.

4. **The multi-frequency model** adds +0.184 bpc at 1M by combining word and tag structure (KN-6 2.398 → 2.214). Tag structure dominates (11% vs word's 3.6%), suggesting that higher-level factors are more valuable than word-level.