The GCD Is Almost Always One: Empirical Validation of the KN Ring Structure

Claude and MJC

February 2026

Abstract

We test the central prediction of the kn-quotient-v2 paper: that the per-row GCD $g(c)$ of the KN count table equals 1 for almost all contexts. On 100M bytes of enwik9, we compute exact GCDs for 4,473 sampled context rows across orders 2–6 and sweep the discount D from 0.50 to 1.00. Results: $g(c) = 1$ for 98.3% of contexts. The optimal discount is $D^* = 0.85$ (1.9537 bpc), slightly below the traditional $D = 0.9$ (1.9563 bpc, +0.003 bpc). The integer discount $D = 1.0$ is significantly worse (2.011 bpc, +0.055 bpc). The per-row GCD discount $D = g(c)$ is a negative result: +0.138 bpc worse than global $D = 0.85$, because the mean g when $g > 1$ is 4,585, destroying count discrimination. These results confirm that the ring structure correctly describes the algebra, but the optimal statistical operation is fractional, not integer.

1 Experiment Design

We build a KN-6 model on 100M bytes of enwik9 using the same 128M-entry hash table as our v16 scorer. The hash table stores counts, type counts, and totals for contexts of orders 1–6.

GCD sampling. Every 100,000 bytes, for each active order 2–6, we scan all 256 possible output bytes for the current context, retrieve each count from the hash table, and compute the GCD of all nonzero counts. This gives 4,473 sampled context rows.

Discount sweep. We simultaneously compute KN-6 predictions for 11 discount values $D \in \{0.50, 0.55, \ldots, 1.00\}$ and record the bits per character for each.

2 Results

2.1 GCD Distribution

The prediction holds: **98.3% of contexts have $g(c) = 1$.** Only 78 of 4,473 rows have $g > 1$.

2.2 GCD by Order

Lower orders have higher $g = 1$ fractions. At order 2, nearly every context has at least one output appearing exactly once. At order 6, the fraction drops to 95.6%—still overwhelming, but the 4.4% with $g > 1$ are the high-frequency contexts where all observed outputs have appeared multiple times.

$g(c)$	Count	Percentage	Cumulative
1	4,395	98.26%	98.26%
2	41	0.92%	99.17%
3	15	0.34%	99.51%
4	4	0.09%	99.60%
5	7	0.16%	99.75%
7	3	0.07%	99.82%
9	2	0.04%	99.87%
≥ 15	6	0.13%	100.00%

Table 1: GCD distribution across 4,473 sampled context rows.

Order	Samples	$g = 1$	Fraction $g = 1$
2	992	990	99.80%
3	974	974	100.00%
4	939	925	98.51%
5	847	817	96.46%
6	721	689	95.56%

Table 2: Fraction of contexts with $g = 1$ by order.

2.3 Discount Sweep

The optimal discount is $D^* = 0.85$, beating the traditional $D = 0.9$ by 0.003 bpc. The curve is flat near the optimum: the range $D \in [0.80, 0.90]$ spans only 0.003 bpc.

The integer discount $D = 1.0$ pays a penalty of 0.058 bpc —substantial, because it zeroes out every count-1 output. Since 98.3% of contexts have at least one count-1 output, $D = 1$ kills the lowest-count continuation in nearly every row.

3 Interpretation

3.1 Why $g = 1$ Almost Everywhere

A context has $g > 1$ only when *every* observed output appears at least twice (and all counts share a common factor). In natural language, this is rare because the long tail of rare continuations ensures at least one output appears exactly once. Zipf's law guarantees this: the distribution of outputs given a context is heavy-tailed, and the tail always includes singletons.

The exceptions are high-frequency, low-diversity contexts: short contexts like single bytes, where a few common bigrams dominate and all have high counts. At order 6, 4.4% of sampled contexts have $g > 1$—these are the very common 5-byte contexts ("the ", "of th", etc.) where every continuation is well-observed.

3.2 The Discount–GCD Gap, Quantified

The kn-quotient-v2 paper predicted that the discount–GCD gap is small because $g = 1$ almost everywhere. We can now quantify all three components of the gap:

D	bpc	MB	Δ from $D = 0.85$
0.50	1.9920	23.7	+0.038
0.60	1.9733	23.5	+0.020
0.70	1.9606	23.4	+0.007
0.80	1.9541	23.3	+0.000
0.85	**1.9537**	**23.3**	—
0.90	1.9563	23.3	+0.003
0.95	1.9641	23.4	+0.010
1.00	2.0112	24.0	+0.058

Table 3: Discount sweep on 100M bytes.

1. **Subtraction vs. division.** For $g = 1$ rows, both operations subtract 1 (or D). The gap comes from the 1.7% with $g > 1$: for these, division by g preserves ratios while subtraction of D distorts them. Impact: < 0.002 bpc (estimated from the fraction and typical count sizes).

2. **Global vs. per-row.** The optimal global $D = 0.85$ vs. per-row $g(c) \in \{1, 2, 3, \ldots\}$. For the 98.3% with $g = 1$, the global $D = 0.85$ under-discounts (keeps $1 - 0.85 = 0.15$ residual for count-1 outputs). For the 1.7% with $g > 1$, the global D over-discounts relative to the GCD. Impact: the flat optimum (< 0.003 bpc over $D \in [0.80, 0.90]$) shows this is small.

3. **Fractional vs. integer.** $D = 0.85$ (fractional) vs. $D = 1$ (the smallest integer discount). Impact: 0.058 bpc —the largest component. The 0.15 residual that $D = 0.85$ leaves for count-1 outputs is important: it allows them to contribute to the prediction, biased toward the backoff distribution. $D = 1$ kills them entirely.

3.3 The 0.15 Residual

The most interesting finding is that $D^* \approx 0.85$, leaving a 0.15 residual for count-1 outputs. In ring terms, this residual has no interpretation as integer division—it is a fractional operation. But it has a clear statistical interpretation: a count of 1 is weak evidence. Subtracting 0.85 leaves 0.15, which says "this output has appeared once in this context, which is weak evidence that it might appear again." Subtracting 1.0 says "this output has appeared once, which is no evidence at all"—too aggressive.

The optimal D is the point where the benefit of sharpening the distribution (by reducing count-1 outputs) balances the cost of losing weak evidence. At 100M bytes on enwik9, this balance point is $D = 0.85$.

3.4 Type Count Distribution

The sampled contexts have a wide range of type counts (distinct outputs): 50% have ≤ 21 distinct outputs, with the distribution extending to $\sim$120. Low-type contexts (2–5 outputs) account for 18% of samples—these are the most constrained contexts (specific word positions, XML tags). High-type contexts (50+) account for about 50%—these are the flexible contexts where many continuations are possible.

4 Discount Scaling: D^* Tracks HT Saturation

A fine-grained sweep (D from 0.70 to 0.95 in 0.01 steps) on full enwik9 (1B bytes) reveals that the optimal discount shifts with scale:

Data	D^*	bpc	HT fill
10M	0.85	2.175	8%
100M	0.83	1.954	37%
200M	0.83	1.894	55%
500M	0.82	1.757	92%
1B	0.87	1.682	100%

Table 4: Optimal discount shifts with HT saturation.

The pattern: D^* decreases as data grows ($0.85 \to 0.82$) while the hash table has room, then *jumps back up* ($0.82 \to 0.87$) when the HT saturates at 100%. Explanation: when the HT is unsaturated, counts are reliable and less smoothing is optimal. When saturated, hash collisions inject noise into counts, and more smoothing (higher D) compensates.

At 1B, the curve is extremely flat: $D \in [0.85, 0.88]$ all give 1.6817–1.6818 bpc. The improvement of $D^* = 0.87$ over $D = 0.90$ is only 0.0003 bpc (∼40 KB)—not practically significant.

5 Per-Row GCD Discount (Negative Result)

The ring-native operation from kn-quotient-v2 is: divide each count by $g(c)$, the row GCD, then apply $D = 1$ to the reduced counts. We test this directly, comparing five strategies on 100M bytes:

Strategy	bpc	Δ from $D = 0.85$
$D = 0.85$ (global optimal)	1.9537	—
$D = 0.90$ (traditional)	1.9563	+0.003
$D = 1.00$ (integer)	2.0112	+0.058
$D = g(c)$ (per-row GCD)	2.0921	+0.138
Hybrid ($D = 0.85$ if $g = 1$, else $g(c)$)	2.0323	+0.079

Table 5: Per-row GCD discount is a negative result.

The per-row GCD discount is **worse than any global discount**, including $D = 1$. It loses 0.138 bpc—more than twice the penalty of global $D = 1$ (0.058 bpc). Even the hybrid (which only uses $g(c)$ for the 8% of evaluations where $g > 1$) loses 0.079 bpc.

Why it fails. The mean g when $g > 1$ is 4,585. These are ultra-common short contexts ("the ", "of ", etc.) where every output has been observed thousands of times. Dividing by $g = 4{,}585$ reduces all counts to single digits, then $D = 1$ wipes out most of the distribution. The resulting prediction is almost entirely backoff.

Note the discrepancy: 98.3% of *sampled contexts* have $g = 1$, but only 92.1% of *context-order evaluations* do. High-frequency short contexts are evaluated much more often, amplifying the $g > 1$ population from 1.7% to 7.9% of prediction weight.

5.1 Implication for the Ring Framework

The ring structure correctly describes the algebra: contexts are integers, order projection is modular reduction, and division by g is the natural ring operation. But the optimal *statistical* operation is fractional ($D = 0.85$), not integer ($D = g$).

This is not a failure of the ring framework—it is a *separation result*: the algebraic structure and the statistical optimum are different objects. The ring provides the scaffold; the prior (the 0.15 residual for count-1 outputs) provides the regularization that integer arithmetic cannot express.

6 Conclusion

The empirical results confirm the ring-theoretic prediction and quantify its limits:

1. $g(c) = 1$ for 98.3% of contexts (ranging from 95.6% at order 6 to 100% at order 3).
2. The optimal discount $D^* = 0.85$ is close to but below 1, beating $D = 0.9$ by 0.003 bpc and $D = 1.0$ by 0.058 bpc.
3. The discount–GCD gap is dominated by the fractional vs. integer component (0.058 bpc), not by subtraction vs. division (< 0.002 bpc) or global vs. per-row (< 0.003 bpc).
4. **Per-row GCD discount is harmful** (+0.138 bpc). The ring-native integer operation is algebraically clean but statistically suboptimal.

The UM-native model would use integer division by $g(c)$ per row, which is exact for $g > 1$ rows but equivalent to $D = 1$ for $g = 1$ rows. The 0.058 bpc penalty of $D = 1$ shows that the fractional discount captures real statistical value that pure integer arithmetic misses. The per-row GCD experiment confirms this: even using the algebraically "correct" per-row operation makes things worse (+0.138 bpc), because the high-frequency $g > 1$ contexts need gentle discounting, not aggressive division by their (large) GCD.

The resolution: the 0.15 residual is a *prior*—it encodes the belief that a count-1 event is weak but nonzero evidence. The ring structure provides the framework; the prior provides the regularization. Integer arithmetic alone is insufficient.

References

[1] Claude and MJC, "Kneser–Ney on the Integers: The Ring Structure," 2026.

[2] Claude and MJC, "Integer Factorization of Events," 2026.

[3] Claude and MJC, "Match Models and the Combination Problem," 2026.