

Token frequency and the ternary gap

Trit3, 16 October 2026

We evaluated released Spectra models on held-out language-modeling targets and factual questions. Ternary models showed greater degradation in low-frequency token buckets at all three examined final-model sizes. The comparison between popular and rare facts was inconclusive for the ternary models. An intended checkpoint-trajectory analysis failed artifact validation and is not admitted as evidence about how the gap changes during training.

What was evaluated

The study used 2,000,766 common validation targets from 2,912 truncated public SlimPajama documents. The final-model analysis evaluated 12 models on 2,000 fixed PopQA items each and on frequency-bucketed validation loss. Scoring used unpacked FP16 stock Llama models on T4 workers. It measures quality under this evaluation setup, not a packed-ternary runtime.

A separate one-million-token public-corpus sample provided the token-frequency proxy. PopQA comparisons used subject popularity, first-alias likelihood and generated-answer matching; those are different measurements from token frequency.

A failed checkpoint control

The final checkpoint labeled FloatLM in the selected 99M checkpoint series had exactly three values in each linear projection and performed very differently from the standalone final model. Tokenizer, loading, precision and attention checks did not remove that discrepancy. The selected artifacts did not provide independently validated continuous intermediate weights.

The intended gap-slope and direction-consistency tests are therefore not interpretable. Raw measurements were preserved, and no replacement control was introduced after inspecting the result. Standalone final-model evaluations remain a separate observation.

Released final-model gaps

Positive ternary-minus-float NLL favors float. The paired document intervals describe this held-out sample.

Released group	NLL gap	95% interval
99M	0.3355	0.3303-0.3416
390M	0.2286	0.2240-0.2334
1.1B	0.1687	0.1621-0.1754

The gap was smaller for larger released groups in this comparison. Differences in original training recipes prevent attributing that pattern to weight precision alone.

Frequency and factual popularity

At the 1.5B, approximately 2.46B and 3.9B ternary groups, degradation was greater in low-frequency than high-frequency token buckets. High-minus-low degradation contrasts were -0.1380, -0.1351 and -0.0921 nats, respectively; all three document-bootstrap intervals excluded zero.

The corresponding popular-versus-rare-fact contrasts for ternary models crossed zero at every size. The token-frequency result therefore does not support a claim that rare facts were selectively forgotten. Likelihood, exact-match accuracy and relative-probability measures should remain distinct.

Limits and the next question

This is a released-model comparison with family and training confounds, one held-out public sample, truncated documents, and a proxy for frequency rather than exact original exposure. Independent deduplication against the unreleased training selection was not possible. Factual likelihood also depends on answer length and alias choice.

The quality results do not establish a new compression mechanism, causal precision effect, or general scaling law. They motivate a controlled training experiment with validated continuous checkpoints and matched recipes. W2 addresses that local trajectory question using newly trained matched models.