

The ternary-float gap during training

Trit3, 14 October 2026

In a frozen small-scale experiment, the overall ternary-minus-float loss gap increased with training data at all three model sizes. Both model families improved, but float improved more across the fitted range. All 40 planned evaluations completed. The result answers a local training-trajectory question; it does not imply monotonic growth at every point, a universal scaling law, or a packed-ternary speed or storage advantage.

A controlled comparison

We trained three Llama-style decoder sizes with matched data, initial latent weights, batch size, optimizer, weight decay and learning-rate schedule shapes. Attention and MLP forward weights used tensor-wise absmean ternarization with an identity straight-through estimator. Latent weights, activations, embeddings, the untied output head and normalization parameters remained FP32.

Each family independently selected learning rate 0.001 from the same three-value development grid. That rate was then transferred to all sizes. This removes a selected-learning-rate difference in this experiment, while leaving the limitations of a short, transferred grid.

Data and evaluation

Five training-data points spanned 2,998,272 to 48,005,120 targets. Evaluation used 2,000,766 held-out targets from 2,912 documents, with at most 256 targets per chunk. The small size had two training seeds; the middle and large sizes had one each.

Exact test and development document hashes were excluded from the training streams. Near-duplicate overlap was not checked. Projection parameter counts were approximately 6.8M, 15.3M and 42.5M; these are not whole-model parameter counts.

The measured gap

The table reports ternary-minus-float NLL in nats. Positive values favor float. The slope is measured against the natural log of the projection-level data-to-parameter ratio.

Projection size	First gap	Last gap	Slope (95% CI)
6.8M	0.0394	0.1331	0.03171 (0.03055-0.03300)
15.3M	0.0307	0.1121	0.02754 (0.02643-0.02885)
42.5M	0.0710	0.1029	0.00992 (0.00865-0.01116)

The preregistered slope criterion passed at all three sizes. Each document-bootstrap interval excluded zero, and each fitted slope exceeded the observed small-size seed-slope spread. Small and large curves had initial dips, so the result concerns an overall trend, not pointwise monotonicity.

Token frequency

Gap growth was faster in low-frequency than high-frequency token buckets for the small and large models. The middle-size contrast was inconclusive. Token frequency came from a separate one-million-token corpus proxy. It does not measure exact training exposure or demonstrate selective forgetting of rare facts.

Uncertainty and scope

Intervals use 1,000 paired whole-document bootstrap resamples across the five data points. They condition on the trained models and do not capture population uncertainty across training seeds. The study uses one corpus, short contexts, few seeds and a transferred learning-rate grid.

The largest projection-level data-to-parameter ratios were approximately 7.0, 3.1 and 1.1, leaving the larger models undertrained. Cooled checkpoints were reloaded and hash-checked within the trainer. A separate analysis reproduced the reported slopes and intervals, but checkpoint verification was not performed by an independent fresh-reader process.

The result motivates a matched replication over a wider training range with more independent seeds and size-specific learning-rate validation. Since computation and latent checkpoints remained FP32, the experiment demonstrates no packed-ternary speed or storage saving.