What calibration tells mixed-precision quantization

MANTIS DISCOVERY

MANTIS · MIT CSAIL

OCTOBER 2026

01

The question

Mixed-precision quantization gives each part of a network its own number of bits, and a few calibration runs decide which parts get more. Which measurements from those runs actually lead to better allocations?

02

How we are approaching it

- Compare scalar loss profiles, reconstructed output vectors and direct comparisons, under the same budget of forward passes.
- Evaluate every allocation exhaustively on 25 small trained networks, to separate approximation bias, sampling variance and options left off a shortlist.

03

A toy allocation problem

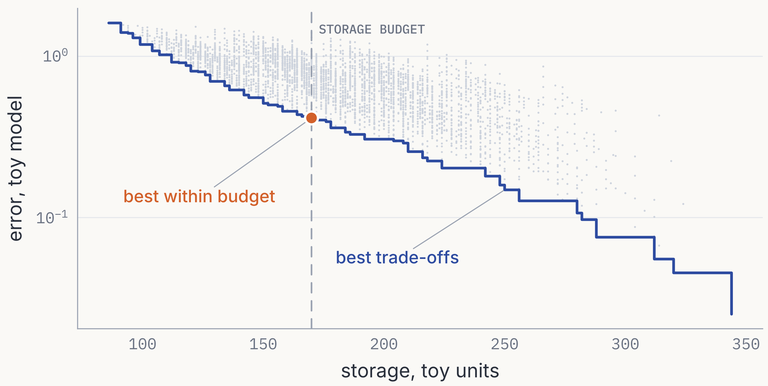


FIGURE 1. This is a toy version, not from the study. The grey points are all 4,096 ways to give six tensors 2, 3, 4 or 8 bits. The blue line marks the best trade-offs, and the orange point is the best choice under the storage budget.

04

Allocation regret



FIGURE 2. Scoring outputs instead of adding up scalar losses reduces normalized allocation regret across the 25 networks. The bars show the reduction with round-to-nearest quantization and after a GPTQ update.

05

What we have so far

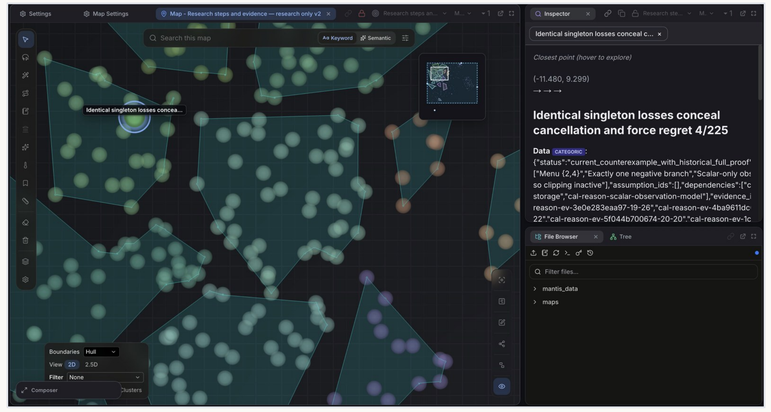
With round-to-nearest quantization, scoring outputs cuts normalized allocation regret by 61.0% compared with adding up scalar losses.

After a GPTQ update the cut is 6.7%, and its interval includes zero.

Cross-tensor Fisher terms can help under rounding but add sampling error after reconstruction.

06

In Mantis



500 RECORDS · THE OPEN RECORD

Identical singleton losses conceal cancellation and force regret 4/225.

The record is a counterexample: in a three-branch example every single-tensor loss looks the same, and the best rule that sees only scalar profiles still has a worst-case regret of 4/225.

ALSO ON RECORD

A two-candidate example shows that the screening threshold 2η cannot be reduced.

07

Open questions

- The study covers small networks and quantizers. Do the findings carry over to larger models?
- Lower discrepancy from the full-precision teacher did not always mean better accuracy or speed. What should calibration aim for instead?