PUBLIC VERIFICATION RECORD

# Public Verification Record No.1

**Measuring the average quality of a publicly distributed quantized model against its original checkpoint under identical conditions**


| | |
|---|---|
| **Issued by** | Treasure Planning LLC / 6 August 2026 |
| **Version** | `1.0` |
| **Status** | This is a public document. No non-disclosure agreement is required. |


## 1. Why we publish this record

Quantized (compressed) models are widely distributed so that large language models can run on limited compute. **However, published records comparing quality against the original uncompressed checkpoint under identical conditions are rare.**

We surveyed 228–232 publicly available quantization repositories across nine major overseas `MoE (Mixture-of-Experts)` model families. **Only 9** of them published a quantitative comparison against the original under identical conditions. **The remaining 96% had no such comparison.**

> **We counted a repository as "has comparison" only when all three of the following held.**
>
> 1. It compares the original and the quantized version
> 2. The comparison uses the same benchmark and the same scoring
> 3. Numbers are given for both sides

Statements about successful loading, sample generations, inference speed, `VRAM` usage, or file size were not counted as quality comparisons.

**This is the first of a series of records intended to fill that gap, one model at a time.**

# 2. Subject of measurement

## 2.1 Quantized version (the subject)

| Item | Value |
|---|---|
| Repository | `QuantTrio/Qwen3-Coder-30B-A3B-Instruct-AWQ` |
| `revision` | `c58857a7f41c0920f73d1b56678640f9c02017d7` |
| License | `Apache-2.0` |
| Quantization | `AWQ / 4-bit / group_size 128 / version gemm / zero_point true` |
| Not quantized | `.mlp.gate` (router) only |
| Structure | 48 layers / 128 experts / hidden 2048 |

**sha256 of the artifacts we obtained**

```
9ceb3350115f42be1f5156e6a7121029cfdc837f97f0875b29db4139c4c19920 model-00001-of-00006.safetensors
de1b0798c44741db570d2b1161003a1fde3cc3e52ded204fec343fae8c488a97 model-00002-of-00006.safetensors
25b05653e45380e785f303277d9d6a9c22a96c4a87d7c4310f072a177b09f1d8 model-00003-of-00006.safetensors
c0a75553f408cdacefbedcc875930ce67d7ea9f6ed6546b0f25037195a6c7e95 model-00004-of-00006.safetensors
2b036efba3962a1cebd0e4b5a5a5d20fdadf1b72a15e625818c9d8031a17c781 model-00005-of-00006.safetensors
3eef55ae738a8357ec441a3c03e1ecfe0bd71742f1ec6ab2aa3ef0251226f0d2 model-00006-of-00006.safetensors
d7884724adb315656e2202b300d6ac9cd10e5774fc65d758722de2ee9486fee4 config.json
3dac122ffd7c08d932b184229883862e932e5b4d154c0648da755beb8b629d76 model.safetensors.index.json
19564a48c4f71a2a1b937cce34c737a1e662b171c5f5d7edf641a15cd896f07d tokenizer.json
```

> **Note on identity**
>
> If the repository is updated, the `revision` alone no longer identifies the artifact. **The sha256 values above uniquely define what this record measured.**

## 2.2 Original (the baseline)

| Item | Value |
|---|---|
| Repository | `Qwen/Qwen3-Coder-30B-A3B-Instruct` |
| `revision` | `b2cff646eb4bb1d68355c01b18ae02e7cf42d120` |
| Precision | `bfloat16` |

### 2.3 Confirming common lineage

Before measuring, we mechanically verified that the quantized version shares a common lineage with the original.

| Item | Value |
|---|---|
| `config` structure | Match |
| `tokenizer` vocabulary hash | Match |
| Non-quantized weights (6 tensors sampled) | Bit-for-bit match |

> **One field excluded from comparison**
> We excluded exactly one field: `intermediate_size` (original 6144 / quantized 5472). All 48 layers of this model are `MoE` (`mlp_only_layers` is empty and `decoder_sparse_step` is 1), so no dense `FFN` exists and **no tensor in the checkpoint uses this value.** The exclusion and its reason are preserved in the measurement record.

## 3. Measurement conditions

| Item | Value |
|---|---|
| Evaluation data | Our own held-out corpus |
| sha256 of evaluation data | `acc364043186af7d3785b83c85aba2cbe393c8c3fe1c19b89802a9a8c491652a` |
| Sequences × sequence length | `53 × 1024` |
| Tokens measured | `54,219` |
| Compute precision | `float16` |
| GPU | `NVIDIA H100 NVL (95,830 MiB)` |
| `torch / CUDA / driver` | `2.12.0+cu130 / 13.0 / 595.71.05` |
| Date of measurement | 6 August 2026 |

## 4. Metric

We used **`Δbpb (delta bits per byte)`**: the difference in the number of bits the original and the quantized version each require for the same text.

A larger value means the quantized version needs more bits than the original, that is, **its average prediction has degraded.**

```
Δbpb = ( NLL of quantized version − NLL of original ) / ln 2
```

Differences are taken between corresponding token positions.

# 5. Result

| Item | Value |
|---|---|
| **Δbpb (point estimate)** | **0.05363** |
| 95% confidence interval | **[0.04071, 0.06796]** |
| Does the interval cross zero? | **No** |

**Finding**

Because the confidence interval does not include zero, **this quantized version is significantly degraded in average quality relative to the original**, under the conditions stated above.

## 5.1 Reference: size

| Item | Approximate |
|---|---|
| Original (bfloat16) | approx. 61 GB |
| Quantized version | approx. 16 GB |
| Ratio | approx. 3.8× |

**Note on size figures**

These are approximations from directory sizes, not a strict byte-level accounting.

## 5.2 Reference: the distributor's own statement

The README of the repository in question carries the following statement by the distributor:

> This model suffers from significant loss under 4-bit quantization,
> please use with caution.

**This measurement puts a number on the degree of loss that statement describes.**

# 6. Method

## 6.1 Confidence interval

We used a **sequence-level paired bootstrap**.

1. For each of the 53 sequences, compute the mean difference between the original and the quantized version
2. Resample the 53 sequences with replacement and compute the mean
3. Repeat step 2 `10,000` times to form a distribution
4. Take the 2.5th and 97.5th percentiles as the interval

## 6.2 Why the resampling unit is the sequence

> Tokens within the same sequence are correlated. Resampling tokens independently would ignore that correlation and **produce an interval narrower than it should be.** Using the sequence as the unit avoids this.

## 6.3 Settings

| Item | Value |
|---|---|
| Iterations | `10,000` |
| Random seed | `42` |
| Significance level | `0.05` |
| Resampling unit | sequence (53) |
| Implementation | `brain/linea_eval_schema.py` |
| `md5` | `38b1c2913f796051dfae1ccbd342d05d` |

## 6.4 Contract tests

The implementation ships with contract tests that mechanically check the following.

- Identical data yields a point estimate and an interval of exactly zero
- A constant offset is reproduced in both the point estimate and the interval
- The same random seed reproduces results regardless of call order
- The sequence-level interval is wider than a token-level interval
- Invalid input (dimension, shape, out-of-range values) is rejected

## 7. On reproducibility

A third party can reproduce this measurement given all three of the following.

| | | |
|---|---|---|
| 1 | The quantized checkpoint with the sha256 above | Public |
| 2 | The original checkpoint with the `revision` above | Public |
| 3 | The evaluation data | Not public |

**Limit on reproducibility**

**Because item 3 is not published, a third party cannot currently reproduce the exact figures in this record.**

We record the sha256 of the evaluation data so that, if we publish it in future, identity can be confirmed. Different evaluation data will naturally yield different figures. The figures in this record hold under the conditions stated above.

## 8. Limitations

**The following are things this record does not claim.**

**This is not a comparison of quantization methods.**
We compared one quantized version against its original. We did not compare it against other quantization methods or other settings.

**This is not a claim that our technology is superior.**
We performed no repair in this measurement. We compared an original against a quantized version produced by a third party.

**This is not a criticism of the distributor.**
The distributor states the loss in their own `README`, and this record puts a number on its degree. The repository is published under `Apache-2.0` and is widely used.

**This does not generalize.**
One model, one quantized version, one evaluation dataset. We say nothing about other models, other quantized versions, or other evaluation data.

**Average quality only.**
What we measured is the difference in average quality. Nothing else is in scope.

## 9. How to use this record

This is a public document. Quotation and redistribution are free.

- As input when deciding whether to adopt a quantized version
- As material for checking the assumption that "compression does not cost quality"
- As a reference for the method when performing similar measurements

**Condition on quotation**

**When quoting figures from this record, please carry the limitations in Section 8 with them.** Extracting the numbers alone would assert things this record does not assert.

## 10. Issuer

**Treasure Planning LLC**

Yuji Nose, Representative Partner

Matsudo Building 13F, 1307-1 Matsudo, Matsudo-shi, Chiba 271-0092, Japan

nose@treasure-kikaku.jp

We measure how compression changes the quality of MoE large language models and record the measurement conditions and their scope as a certificate.

**Notices**

Public Verification Record No.1 / 6 August 2026 / Version 1.0