PUBLIC VERIFICATION RECORD

Public Verification Record No.2

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

Item	Value
Issued by	Treasure Planning LLC / 9 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 survey the publicly available quantization repositories derived from major overseas MoE (Mixture-of-Experts) models and count how many publish a quantitative comparison against the original under identical conditions. **Most repositories publish no such comparison.**

The size of the population, the number of repositories that published a comparison, the criteria used to judge, and the counting procedure are published on our Methodology page. Because those figures are updated as we recount, they are not transcribed into this record.

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

1.1 How this record relates to No.1

Public Verification Record No.1 measured a quantized version of a different model produced by the same distributor. This record holds the following conditions identical to No.1.

Condition held identical	Value
Distributor	Same (QuantTrio)
Quantization	Same (AWQ / 4-bit / group_size 128 / version gemm / zero_point true)
Not quantized	Same (`.mlp.gate` only)
Model structure	Same (48 layers / 128 experts / hidden 2048 / moe_intermediate 768)
Evaluation data	Same (sha256 matches; see below)
Tokens measured	Same (54,219)
Metric and statistics	Same (Δbpb / sequence-level paired bootstrap)
What differs	**The model measured, and nothing else**

The difference between the figures in this record and those in No.1 therefore does not arise from the distributor, the quantization settings, or the evaluation conditions. The comparison is given in Section 5.2.

2. Subject of measurement

2.1 Quantized version (the subject)

Item	Value
Repository	`QuantTrio/Qwen3-30B-A3B-Thinking-2507-AWQ`
revision	`3723ded1261d119450306434d8f454fb217ac1fd`

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 / top-k 8 / hidden 2048 / moe_intermediate 768
Keys in the checkpoint	56,115

sha256 of the artifacts

```
948ce02f04cf58748d27d27f82b24ac60d4ebc5a0480b0d96a876b1bf0ff4568  model-00001-of-00004.safetensors
4aa48901b2a928545ead20cb869fead4f9015197ab98369a646dc3f4b7747558  model-00002-of-00004.safetensors
f5ec9a76ed137dcf4f8c157fc497d2269cc2dbc98a85cf350dd8daa933f3dbef  model-00003-of-00004.safetensors
0535d6f81184f9f78e6ad02941e4280aa7d55deb39a2f210ee42e25cc030f360  model-00004-of-00004.safetensors
fb919f95ea713dfe151d753f1a72b8aca4d91b37bf037556e70bb37e034687c2  config.json
206114c230a700259ec81a0e466ffca32462afb8c558f76ca1aea55e73bc331d  model.safetensors.index.json
19564a48c4f71a2a1b937cce34c737a1e662b171c5f5d7edf641a15cd896f07d  tokenizer.json
```

> **Where these sha256 values come from**
>
> The four `.safetensors` values are those published by the distributor as Git LFS metadata. We did not compute them from the artifacts ourselves. A third party who downloads the same revision can verify them with `sha256sum`.
>
> The three values for `config.json`, `model.safetensors.index.json` and `tokenizer.json` were computed by us from the retrieved artifacts.

> **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/Qwen3-30B-A3B-Thinking-2507`
revision	`144afc2f379b542fdd4e85a1fcd5e1f79112d95d`
Precision	bfloat16

2.3 Confirming common lineage

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

Check	Result
config structure	Match
tokenizer vocabulary hash	Match
Non-quantized weights (6 tensors sampled)	Bit-for-bit match

The six tensors sampled for comparison were:

```
model.embed_tokens.weight
model.layers.15.self_attn.q_norm.weight
model.layers.22.self_attn.q_norm.weight
model.layers.30.input_layernorm.weight
model.layers.38.input_layernorm.weight
model.layers.45.input_layernorm.weight
```

> **One field excluded from comparison**
>
> We excluded exactly one field: `_name_or_path`, which is present only in the quantized version. It records the path used when the quantization was performed; it corresponds to nothing in the model structure and to no tensor in the checkpoint. The exclusion and its reason are preserved in the measurement record.
>
> Apart from this one field, the config of the original and that of the quantized version match exactly.

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 80GB HBM3 (81,559 MiB)
torch / CUDA	2.12.0+cu126 / 12.6
Date of measurement	7 August 2026

The sha256 of the evaluation data is identical to that in No.1. Both records were measured on the same evaluation data.

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)	**0.01882**
95% confidence interval	[0.01523, 0.02265]
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	16,809,468,128 bytes (approx. 16.8 GB)
Ratio	approx. 3.6×

> **Note on size figures**
>
> The byte count for the quantized version is the sum of the four `.safetensors` files as recorded in the distributor's metadata. The figure for the original is an approximation from directory size, not a strict byte-level accounting.

5.2 Comparison with No.1

As stated in Section 1.1, this record and No.1 hold every condition identical except the model measured.

Record	Subject	Δbpb	95% CI
No.1	`QuantTrio/Qwen3-Coder-30B-A3B-Instruct-AWQ`	0.05363	[0.04071, 0.06796]

No.2 (this record)	`QuantTrio/Qwen3-30B-A3B-Thinking-2507-AWQ`	0.01882	[0.01523, 0.02265]

> **The two intervals do not overlap**
>
> The upper bound of No.2, 0.02265, lies below the lower bound of No.1, 0.04071. The difference between the two is not explained by the statistical uncertainty of these measurements.

These two records hold the distributor, the quantization method, the group size, the scope of what is left unquantized, the model structure, the evaluation data, the number of tokens measured, and the statistical procedure all identical. Only the model measured differs. These two records therefore show that **even for quantized versions produced by the same distributor with the same settings, the difference in average quality against the original varies by model.**

These are observations about two models. This record says nothing about how large such differences are, or across what range of models they occur (see Section 8).

5.3 On the distributor's statement

For the subject of this record, we found no statement in the distributor's README concerning the change in quality caused by quantization (as of 7 August 2026, at the revision given above).

For the subject of No.1, the distributor did state in their README that the model suffers significant loss under 4-bit quantization and should be used with caution. No comparable statement appears for the subject of this record.

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`

> **On the resampling unit**
>
> The implementation resamples whatever array it is given, element by element (the label it returns is `question`). For this measurement we passed 53 values already aggregated to the sequence level (1,023 tokens per sequence; 53 × 1,023 = 54,219), so the resampling unit is the sequence.

The md5 of the implementation is identical to that in No.1. Both records were computed with the same implementation. In preparing this record we recomputed the figures of No.1 (0.05363 / [0.04071, 0.06796]) with the same procedure and confirmed that they match the published values.

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.

#	Item	Availability
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.

- **The difference from No.1 does not indicate that one quantization is better made than the other.** The difference between the two records is the result of applying the same method to different models. It is not a statement about the quality of the work.
- **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 repository is published under Apache-2.0. This record measures and publishes a figure that had not been published.
- **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. Even taken together with No.1, these are two records.
- **This is not a certification under any institutional scheme.** This document records a measurement we performed by our own procedure, together with its conditions and result. It is not a certification, accreditation, or conformity assessment by a third-party body, and it confers no status under any public qualification or scheme.
- **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.2 / 9 August 2026 / Version 1.0