

## PUBLIC VERIFICATION RECORD

## 公開検証記録 No.1

量子化済み公開モデルの平均品質を、原本と同一条件で測定した記録

発行 トレジャー企画合同会社 / 2026年8月6日  
版 1.0  
位置づけ 本書は秘密保持契約を必要としない公開資料です。

## 1. なぜこの記録を出すか

大規模言語モデルを限られた計算資源で動かすため、量子化（圧縮）されたモデルが広く配布されています。しかし、**圧縮前の原本と同一条件で品質を比較した記録が公開されている例は、ほとんどありません。**

当社が海外の主要な MoE（Mixture-of-Experts）9系統について、公開されている量子化リポジトリ 228～232件を調査したところ、原本との同一条件の定量比較が公開されていたのは9件のみでした。残る 96% には比較がありません。

## 「比較あり」と判定するための3条件

- 原本と量子化版を比較している
- 同一のベンチマーク・同一の採点で行っている
- 双方に数値が記載されている

起動確認、生成例、推論速度、VRAM 使用量、ファイル容量の記載は品質比較に数えていません。

本書は、その空白を1件ずつ埋める記録の第1号です。

## 2. 測定対象

### 2.1 量子化版（測定対象）

| 項目       | 内容                                                           |
|----------|--------------------------------------------------------------|
| リポジトリ    | QuantTrio/Qwen3-Coder-30B-A3B-Instruct-AWQ                   |
| revision | c58857a7f41c0920f73d1b56678640f9c02017d7                     |
| ライセンス    | Apache-2.0                                                   |
| 量子化方式    | AWQ / 4bit / group_size 128 / version gemm / zero_point true |
| 非量子化     | .mlp.gate(router) のみ                                         |
| 構成       | 48層 / 128 expert / hidden 2048                               |

#### 取得した実体の sha256

```
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
3dac122ffd7c08d932b184229883862e932e5bd154c0648da755beb8b629d76 model.safetensors.index.json
19564a48c4f71a2a1b937cce34c737a1e662b171c5f5d7edf641a15cd896f07d tokenizer.json
```

#### 同一性に関する注記

リポジトリが更新された場合、revision だけでは同一性を特定できません。上記の sha256 が本書の測定対象を一意に定めます。

### 2.2 原本（比較基準）

| 項目       | 内容                                       |
|----------|------------------------------------------|
| リポジトリ    | Qwen/Qwen3-Coder-30B-A3B-Instruct        |
| revision | b2cff646eb4bb1d68355c01b18ae02e7cf42d120 |
| 精度       | bfloat16                                 |

## 2.3 同一系譜であることの確認

測定に先立ち、量子化版が原本と同一の系譜であることを機械的に照合しました。

| 項目                | 内容       |
|-------------------|----------|
| config 構造         | 一致       |
| tokenizer の語彙ハッシュ | 一致       |
| 非量子化層の重み（6テンソル抽出） | ビット単位で一致 |

### 比較対象から除外した1件

intermediate\_size（原本 6144 / 量子化版 5472）のみを比較対象から除外しています。本モデルは全48層が MoE であり（mlp\_only\_layers が空、decoder\_sparse\_step が 1）、dense FFN が存在しないため、この値を使用するテンソルは checkpoint に含まれません。除外した事実と理由は測定記録に保存しています。

## 3. 測定条件

| 項目                    | 内容                                                               |
|-----------------------|------------------------------------------------------------------|
| 評価データ                 | 当社保有の holdout コーパス                                               |
| 評価データの sha256         | acc364043186af7d3785b83c85aba2cbe393c8c3fe1c19b89802a9a8c491652a |
| 系列数 × 系列長             | 53 × 1024                                                        |
| 測定トークン数               | 54,219                                                           |
| 計算精度                  | float16                                                          |
| GPU                   | NVIDIA H100 NVL (95,830 MiB)                                     |
| torch / CUDA / driver | 2.12.0+cu130 / 13.0 / 595.71.05                                  |
| 測定日                   | 2026-08-06                                                       |

## 4. 指標

$\Delta\text{bpb}$  (delta bits per byte) を用いました。同一のテキストに対して、原本と量子化版がそれぞれ何ビットを要するか之差です。

値が大きいほど、量子化版が原本より多くのビットを必要とする、すなわち平均的な予測が悪化していることを意味します。

$$\Delta\text{bpb} = ( \text{量子化版の負対数尤度} - \text{原本の負対数尤度} ) / \ln 2$$

同一のトークン位置どうしを対応させて差を取っています。

## 5. 結果

| 項目                       | 内容                 |
|--------------------------|--------------------|
| $\Delta\text{bpb}$ (点推定) | 0.05363            |
| 95% 信頼区間                 | [0.04071, 0.06796] |
| 区間がゼロを跨ぐか                | 跨がない               |

### 判定

信頼区間がゼロを含まないため、この量子化版は原本に対して平均品質が有意に劣化していると判定できます。

### 5.1 参考：容量

| 項目            | 内容      |
|---------------|---------|
| 原本 (bfloat16) | 約 61 GB |
| 量子化版          | 約 16 GB |
| 比             | 約 3.8 倍 |

### 容量値に関する注記

ディレクトリ容量からの概算値です。厳密なバイト単位の集計ではありません。

### 5.2 参考：配布者自身の記載

当該リポジトリの README には、配布者自身による次の記載があります。

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

本測定は、この記載が示す劣化の程度を数値で示したものです。

## 6. 方法

### 6.1 信頼区間の算出

系列単位の paired bootstrap により信頼区間を算出しました。

1. 53系列それぞれについて、原本と量子化版の差の平均を求める
2. 53系列を復元抽出で選び直し、その平均を計算する
3. 手順2を10,000回繰り返して分布を作る
4. 分布の2.5パーセンタイルと97.5パーセンタイルを区間とする

### 6.2 リサンプリングの単位を「系列」とした理由

同一系列内のトークンは互いに相関しています。トークンを独立に抽出すると、この相関を無視することになり、**信頼区間が実際より狭く出ます**。系列を単位にすることでこれを避けています。

### 6.3 設定

| 項目        | 内容                               |
|-----------|----------------------------------|
| 反復回数      | 10,000                           |
| 乱数シード     | 42                               |
| 有意水準      | 0.05                             |
| リサンプリング単位 | 系列 (53本)                         |
| 実装        | brain/linea_eval_schema.py       |
| md5       | 38b1c2913f796051dfae1ccbd342d05d |

### 6.4 契約テスト

この実装には契約テストが付属しており、次を機械的に検査しています。

- 同一データ同士では点推定も区間もゼロになる
- 定数差を与えると点推定と区間がその定数になる
- 同じ乱数シードなら呼び出し順に依存せず再現する
- 系列単位の区間が、トークン単位の区間より広くなる
- 不正な入力（次元・形状・範囲外の値）を拒否する

## 7. 再現について

本書の測定は、次の3つが揃えば第三者が再現できます。

- |   |                            |     |
|---|----------------------------|-----|
| 1 | 上記 sha256 の量子化版 checkpoint | 公開  |
| 2 | 上記 revision の原本 checkpoint | 公開  |
| 3 | 評価データ                      | 非公開 |

### 再現性に関する限定

評価データが公開されていないため、現時点で本書と同一の数値を第三者が再現することはできません。

評価データの sha256 を記録しているのは、将来これを公開した場合に同一性を確認できるようにするためです。別の評価データを用いれば、当然に異なる数値が得られます。本書の数値は上記の評価条件におけるものです。

## 8. 限定事項

以下は、本書が主張していないことです。

**量子化手法の優劣を比較したものではありません。**

本書は1つの量子化版を原本と比較しただけであり、他の量子化手法や他の設定との比較は行っていません。

**当社の技術が優れているという主張ではありません。**

本測定において当社は修復を行っていません。原本と、第三者が作成した量子化版を比較しただけです。

**配布者を批判するものではありません。**

配布者自身が README で劣化を明記しており、本書はその程度を数値化したものです。当該リポジトリは Apache-2.0 で公開され、多くの利用があります。

**一般化できません。**

1モデル・1量子化版・1評価データによる測定です。他のモデル、他の量子化版、他の評価データについては何も述べていません。

**平均品質のみを扱っています。**

本書が測定したのは平均的な品質の差です。それ以外の観点は対象外です。

## 9. この記録の使い方

本書は公開資料です。引用・転載は自由です。

- ・ 量子化版を採用する際の判断材料として
- ・ 「圧縮しても品質は落ちない」という前提を確認する材料として
- ・ 同種の測定を行う際の方法の参照として

### 引用時の条件

本書の数値を引用する場合は、第8章の限定事項を併記してください。数値のみを切り出すと、本書が主張していないことを主張したことになります。

## 10. 発行者

トレジャー企画合同会社

代表社員 能勢 優治

〒271-0092 千葉県松戸市松戸1307-1 松戸ビル13F

nose@treasure-kikaku.jp

当社は、MoE 構成の大規模言語モデルについて、圧縮による品質の変化を測定し、測定条件と範囲を証明書として記録する事業を行っています。

### 脚注・権利関係

公開検証記録 No.1 / 2026-08-06 / 版 1.0

モデル名は各社の商標または登録商標です。記載各社と当社の間に提携・推奨等の関係はありません。