

PUBLIC VERIFICATION RECORD

公開検証記録 No.2

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

項目	内容
発行	トレジャー企画合同会社 / 2026年8月9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。

当社は、海外の主要な MoE(Mixture-of-Experts)モデルについて公開されている 量子化リポジトリを全数調査し、原本との同一条件の定量比較が公開されている 割合を集計しています。大半のリポジトリには、そうした比較がありません。

母集団の件数、比較が公開されていた件数、判定の基準、集計に用いた手順は **調査方法**に掲載しています。数値は再集計によって 更新されるため、本書には転記していません。

本書は、その空白を1件ずつ埋める記録の第2号です。

1.1 本号の位置づけ

公開検証記録 No.1 は、同一の配布者による別のモデルの 量子化版を測定したものです。本号は、それと次の条件を揃えています。

揃えた条件	内容
配布者	同一(QuantTrio)

量子化方式	同一(AWQ / 4bit / group_size 128 / version gemm / zero_point true)
非量子化の範囲	同一(.mlp.gate のみ)
モデル構成	同一(48層 / 128 expert / hidden 2048 / moe_intermediate 768)
評価データ	同一(sha256 一致・後述)
測定トークン数	同一(54,219)
指標・統計手法	同一(Δ bbp / 系列単位 paired bootstrap)
異なる点	測定対象のモデルのみ

したがって、本号と No.1 の数値の差は、配布者・量子化設定・評価条件の違いによるものではありません。比較結果は第5.2章に示します。

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/Qwen3-30B-A3B-Thinking-2507-AWQ
revision	3723ded1261d119450306434d8f454fb217ac1fd
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 / topk 8 / hidden 2048 / moe_intermediate 768
checkpoint のキー総 数	56,115

取得した実体の sha256

```
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
```

sha256 の出所について

上記のうち .safetensors 4件は、配布元が Git LFS のメタデータとして 公開している値です。当社が実体から算出した値ではありません。この値は、同じ revision をダウンロードした第三者が sha256sum で照合できます。

config.json / model.safetensors.index.json / tokenizer.json の3件は、当社が実体を取得して算出した値です。

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/Qwen3-30B-A3B-Thinking-2507
revision	144afc2f379b542fdd4e85a1fcd5e1f79112d95d

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。

照合項目	結果
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

抽出して照合した6テンソルは次のとおりです。

```
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
```

比較対象から除外した1件

_name_or_path(量子化版のみに存在)を比較対象から除外しています。これは量子化作業時のパスを記録するメタデータであり、モデルの構造にも checkpoint のテンソルにも対応しません。除外した事実と理由は測定記録に保存しています。

この1件を除き、原本と量子化版の config は完全に一致します。

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	acc364043186af7d3785b83c85aba2cbe393c8c3fe1c19b89802a9 a8c491652a

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 80GB HBM3(81,559 MiB)
torch / CUDA	2.12.0+cu126 / 12.6
測定日	2026-08-07

評価データの sha256 は No.1 と同一です。両号は同一の評価データで測定しています。

4. 指標

Δbpb (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の差です。

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
Δbpb (点推定)	0.01882
95% 信頼区間	[0.01523, 0.02265]
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	概算
原本(bfloat16)	約 61 GB
量子化版	16,809,468,128 バイト(約 16.8 GB)
比	約 3.6 倍

容量値に関する注記

量子化版のバイト数は、配布元のメタデータに記載された .safetensors 4件の合計です。原本の値はディレクトリ容量からの概算であり、厳密なバイト単位の集計ではありません。

5.2 No.1 との比較

第1.1章に示したとおり、本号と No.1 は測定対象のモデル以外の条件を揃えています。

記録	測定対象	Δbpb	95% 信頼区間
No.1	QuantTrio/Qwen3-Coder-30B-A3B-Instruct-AWQ	0.05363	[0.04071, 0.06796]
No.2(本号)	QuantTrio/Qwen3-30B-A3B-Thinking-2507-AWQ	0.01882	[0.01523, 0.02265]

2つの区間は重なりません

No.2 の上限 0.02265 は No.1 の下限 0.04071 を下回ります。両者の差は、本測定の統計的な不確かさの範囲内では説明できません。

この2件は、配布者・量子化方式・group_size・非量子化の範囲・モデル構成・評価データ・測定トークン数・統計手法をすべて揃えています。異なるのは測定対象のモ

デルだけです。したがって、同じ配布者が同じ設定で作成した量子化版であっても、原本との平均品質の差はモデルによって異なることを、この2件は示しています。

なお、この2件は2つのモデルについての観測です。どの程度の差がどの範囲のモデルで生じるかについては、本書は何も述べていません(第8章)。

5.3 配布者の記載について

本号の測定対象について、配布元の README には量子化による品質の変化に関する記載を確認できませんでした(2026年8月7日時点、上記 revision)。

No.1 の測定対象では、配布者自身が README に「4bit 量子化のもとで著しい損失があるため注意して使用されたい」旨を記載していました。本号の対象には同種の記載がありません。

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

リサンプリング単位について

本実装は、渡された配列を単位としてリサンプリングします(戻り値のラベルは question)。本測定では系列単位に集約した 53本(1系列 1,023 トークン、 $53 \times 1,023 = 54,219$)を渡しているため、リサンプリング単位は系列です。

実装の md5 は No.1 と同一です。両号は同一の実装で算出しています。本号の作成にあたり、同じ手順で No.1 の数値($0.05363 / [0.04071, 0.06796]$)を再計算し、公開値と一致することを確認しました。

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。

#	項目	状態
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。
- **No.1** との差は、量子化手法の良し悪しを示すものではありません。両号の差は、同じ手法を異なるモデルに適用した結果の差です。どちらの量子化版がより良く作られているかを述べたものではありません。
- 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。当該リポジトリは Apache-2.0 で公開されています。本書は、公開されていなかった数値を測定して記録したものです。
- 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。No.1 と合わせても2件です。
- 制度上の認証ではありません。本書は当社が自らの手順で測定し、その条件と結果を記録した文書です。第三者機関による認証・認定・適合性評価のいずれでもなく、公的な資格や制度に基づく地位を示すものではありません。
- 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.2 / 2026-08-09 / 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。