得意そうなMathと苦手そうなMath

短く言います。Mathは**2系統のスコア**が取れています。

* math\_n\_226：平均 **31.74**、mathより常に高い（9/9の run で **+10.5〜+22.0pt**、差の平均 **+17.70pt**）。
* math：平均 **14.05**、**分散が大きい**（SD **9.92**、IQR **15.56**）。

この差から、**「検証付き・数値最終解のコンテスト系（AIME/AMC風）」は得意、混在度が高い一般Mathは揺れる**、という結論です。math\_n\_226は **OpenR1-Math** 系の“推論痕跡つき・正答検証済み”問題に対応している可能性が高く（NuminaMath1.5由来、R1の推論を付与・検証）([Hugging Face](https://huggingface.co/datasets/open-r1/OpenR1-Math-220k?utm_source=chatgpt.com))、実例として **Algebra / Geometry** のような分類が含まれます（オリジナルの派生データでも確認可）([Hugging Face](https://huggingface.co/datasets/llamafactory/OpenR1-Math-94k?utm_source=chatgpt.com))。

# **得意そうなMath（実測とデータ性質の両面から）**

* **数値最終解・検証可能型**連立・方程式操作、漸化式・数列、単調性/境界評価、典型不等式、基本的な整数問題など。OpenR1-Math は「数値で正誤が判定できる問題」を中心に構成され、推論痕跡も付与・検証されているため（R1由来の思考トレース、Math Verify/Llama判定）**手順がハマると強い**です。([Hugging Face](https://huggingface.co/datasets/open-r1/OpenR1-Math-220k?utm_source=chatgpt.com))
* **定型の多段推論（思考トレース誘導で取りやすい）**Mixture-of-Thoughts等のR1蒸留系は“段階的に考える”こと自体を学習させる設計で、数学タスクも含みます。**手順化できる問題**は狙い撃ちにできます。([Hugging Face](https://huggingface.co/datasets/open-r1/Mixture-of-Thoughts?utm_source=chatgpt.com), [GitHub](https://github.com/huggingface/open-r1?utm_source=chatgpt.com))

# **苦手そうなMath（推定根拠つき）**

* **自由記述の証明問題（最終値がないもの）**OpenR1-Math系の設計上、**値を伴わない証明は外しがち**です（“値出力のない証明は除外”の方針が議論で言及）。学習・評価の経験不足になりやすい。([GitHub](https://github.com/huggingface/open-r1/issues/24?utm_source=chatgpt.com))
* **図形の補助線・図解推論に強く依存する幾何**幾何カテゴリ自体は含まれるものの（例示あり）、**図や補助線の導入が鍵**の設問は、テキストだけの手がかりが乏しいと失点しやすい傾向です（math側の大きな分散＝問題バケットの混在を示唆）。([Hugging Face](https://huggingface.co/datasets/llamafactory/OpenR1-Math-94k?utm_source=chatgpt.com))
* **“応用寄り”の数理（物理的背景を伴う設定）**Mathと他ドメインの相関を見ると、math は **human\_social / bio\_med** と強く連動（ρ≈0.8）する一方、**Physics との連動は弱い**（ρ≈0.15）。**数理物理寄りの読解 + 定式化**は別筋で鍛えが必要です（今回の評価では Physics 自体が低位）。  
  ※相関は評価CSVからのSpearmanで算出。

## **「正答しやすさのバラツキ」について**

* **math は大振れ**：SD **9.92**、IQR **15.56**。**問題バケットの混在**（証明寄り・図形依存・読解の重さ）や、推論設定（温度/停止条件/最大長）への感度が高い帯です。
* **math\_n\_226 は水準が一貫して高い**：各runで math を **+10〜22pt** 上回り、**検証付き・整った問題形式**に強いことが数値で出ています。

## **すぐできる分解（おすすめ）**

1. **Mathをトピック別に出し分け**（Algebra / Geometry / Number Theory / Probability…）  
   OpenR1-Math 派生のメタ情報はトピックを持つ事例が多いので、評価ログに「topic」列を増設して\*\*math\_[topic]\*\*に自動展開します。これで「得意：Algebra、揺れる：Geometry/NT」等が即座に可視化できます。([Hugging Face](https://huggingface.co/datasets/llamafactory/OpenR1-Math-94k?utm_source=chatgpt.com))
2. **“数値最終解 vs 証明系” の2軸ビニング**最終出力が数値か否か、推論痕跡の有無（verified/非verified）でビン分け。今回の差（math\_n\_226 ≫ math）の**原因切り分け**ができます。
3. **推論設定の固定トラック**vLLM側の max\_model\_len / max\_num\_batched\_tokens / temperature / stop を“Math安定トラック”として別ID管理。math の分散を実運用で潰します。

### **補足（参照）**

* OpenR1-Math-220k：NuminaMath1.5由来、R1の推論痕跡を付与・検証した**大規模数学推論データ**。([Hugging Face](https://huggingface.co/datasets/open-r1/OpenR1-Math-220k?utm_source=chatgpt.com))
* Mixture-of-Thoughts：**数学・コーディング・サイエンス**に跨るR1蒸留の思考データ。([Hugging Face](https://huggingface.co/datasets/open-r1/Mixture-of-Thoughts?utm_source=chatgpt.com), [GitHub](https://github.com/huggingface/open-r1?utm_source=chatgpt.com))

静かな湖面に、底流は二本。**検証付きの整った問題**では力強く泳ぎ、**証明・図形・応用**では渦が立つ。まずは**トピック分解**で流れを可視化し、math の振れを一つずつ潰せば、総合の輪郭はすぐ整います。