GRPO による LLM の数学的推論能力の向上：単一問題を用いた強化学習の有効性

田口昂樹
慶應義塾大学
kk102214pp@keio.jp

概要

近年の大規模言語モデル（LLM）の性能向上は，大規模事前学習コーパスや高品質な教師データ，さらに多数の問題を用いた強化学習など，豊富な学習データを前提として実現されてきた．とりわけ数学的推論能力の向上には，多様な問題を用いた訓練が一般的である．これに対し，本稿では，学習データを極端に制限した場合でも性能改善が可能かを検証するため，数学の単一問題のみを用いた Group Relative Policy Optimization (GRPO)[1, 2] を適用し，日本語 LLM の数学的推論能力への影響を分析した．その結果，単一問題による学習でも性能向上が確認され，改善幅には問題の種類による差があることが分かった．さらに，強化学習と Majority Voting[7] を組み合わせることで，性能を大きく高められる可能性が示された．本稿は，NLP2026 ワークショップ「大規模言語モデルのファインチューニング技術と評価」内コンペティション（FT-LLM 2026）における取り組みを報告するものである．

1　はじめに

大規模言語モデル（LLM）は数理推論タスクにおいて高い性能を示しているが，少量の追加学習によってどこまで推論性能を向上できるかは依然として重要な論点である．特に，強化学習による推論能力の向上が，真に新しい問題タイプの獲得をもたらすのか，あるいは既に潜在的に解ける問題をより安定に解けるようにするだけなのかについては，近年活発に議論されている [3, 4, 5]．

この問いをより単純な条件で検証するため，本研究では **数学の問題 1 問のみ** を用いた GRPO による学習を行い，この極端に小さな教師情報から LLM の性能がどこまで向上するかを検証する．この設定にはいくつかの利点がある．第一に，学習データの作成コストが極めて低く，短時間で実験を行うことができる．第二に，単一の問題であれば人手で内容を精査しやすく，学習データの質を十分に担保できる．第三に，学習対象を最小化することで，計算コストを削減できる可能性がある．さらに，このように教師情報を極端に絞った条件で得られる性能改善の性質を観察することで，強化学習がどの程度まで汎化能力を拡張できるのか，あるいは既存能力を強化する効果にとどまるのかを明らかにできると期待される．

本稿の貢献は以下の 3 点である．

- 数学の単一問題のみを用いた GRPO でも，LLM の数学ベンチマーク精度を向上できることを示す．
- 学習曲線および 30 回サンプリング評価を通じて，性能改善の主因が「解ける問題の種類の増加」ではなく「既に解ける問題をより安定して解けるようになること」にあることを示す．
- 実運用上は，強化学習済みモデルと多数決推論の組み合わせが高い有効性を持つことを示す．

2　関連研究

近年，ごく少数の教師情報を用いた推論特化強化学習の有効性が報告されている．特に，One Training Example による Reasoning RL の研究では，単一の訓練例のみを用いた強化学習でも，推論タスクの性能が向上する可能性が示されている [3]．この系統の研究は，限られた教師情報であっても，方策最適化によって推論時の解答安定性を高められる可能性を示唆している．

また，Matsutani ら [5] は，RL と SFT の役割の違いを整理し，RL は「ベースモデルがもともと一定確率で解ける問題群（既知の可解領域）」の中で出力を圧縮・選別し，SFT は解ける問題空間そのものの拡張に寄与すると指摘している．この観点に立つ

と，RL による性能改善は，新たな問題タイプの獲得というよりも，既存の可解領域における解答の安定化として現れる可能性がある．

3　コンペティションタスク設定

本研究で対象とするコンペティションの数学タスクは，日本の中学校・高等学校で扱う数学問題を対象とし，言語モデルを基盤とするシステムの回答精度を競うものである．問題はテキストのみで記述され，図形の読み取りを必要とする問題は含まれない．一方で，幾何学的関係をテキストで十分に記述できる問題は出題範囲に含まれ，数式は LaTeX 形式で与えられる．

ベースモデルとして使用できる大規模言語モデルは llm-jp-4-8b に限られる．追加学習や推論時の工夫は許容されるが，推論時に利用できる大規模言語モデルは LLM-jp 系に限定され，外部 LLM の利用は認められていない．また，推論時にはインターネットから遮断されるため，ウェブ検索に依存する手法は利用できない．

評価は非公開テストセット 500 問に対して行われ，最終回答の一致に基づく正解率で順位が決定される．開発用には sample problem として 100 問が提供されており，参加者はこの 100 問を用いて手法の検証やチューニングを行う．正解判定ではルールベースの正規化処理が導入されており，数式表現の一定の揺れを許容した上で一致判定が行われる．したがって本タスクでは，自然言語による説明生成そのものよりも，正しい最終解答を安定して出力する能力が重視される．

4　開発手法

4.1　単一問題を用いた GRPO

学習アルゴリズムには GRPO を用い，パラメータ更新は LoRA で行った．報酬関数は単純であり，最終解答が正解なら 1，不正解なら 0 のバイナリ報酬を用いた．

ベースモデルから複数サンプルを生成し，相対的な優劣に基づいて方策を更新する．本稿では GRPO の詳細導出には立ち入らないが，Group 内の相対比較により，単一問題であっても「より高報酬な出力様式」を強めることができる点が重要である．

表 1　主な学習ハイパーパラメータ

項目	設定値
ベースモデル	instruction-tuned llm-jp-4-8b
学習手法	GRPO + LoRA
LoRA rank	16
学習率	1.0×10^{-6}
temperature	1.0
top-p	1.0
num_generations	8
KL 正則化係数	0.001
max_tokens	最大
勾配蓄積	なし

4.2　学習データの選定

単一問題学習では，どの問題を選ぶかが極めて重要である．本研究では，コンペティションで開発用に提供された sample problem（100 問）から候補を選定し，ベースモデルに複数回回答させた際に正解率が約 50%となる問題を抽出して，次の 4 問を比較対象とした．

1. **データ 1（中 2・確率）**：2 つのさいころを同時に投げるとき，出る目の数の和が 7 になる確率を求めよ．
2. **データ 2（中 3・平方根）**：次の数を $a\sqrt{b}$ の形に表しなさい．$\sqrt{24}$
3. **データ 3（III C・極限）**：$\lim_{x\to\infty} \dfrac{\log \sqrt{x}}{x}$ を求めよ．
4. **データ 4（中 2・一次関数）**：点 $(0, 10)$ を通り，傾きが -5 の直線の式を求めなさい．

最終システムでは，このうちデータ 1 で学習したモデルを採用した．

5　実験設定

5.1　学習条件

学習条件を表 1 に示す．temperature および top-p はともに 1.0 とし，多様な候補生成を維持した．また，LoRA rank は 16，学習率は 1.0×10^{-6} とした．

5.2　評価方法

モデル評価には 100 問からなる評価セットを用いた．この評価セットは，sample problem のうち学習に使用していない 96 問に，新たに作成した 4 問を

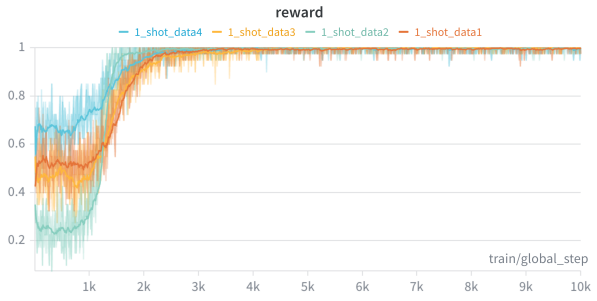

図 1　各学習データに対する reward の推移．1k〜2k step 付近で報酬がほぼ 1 に収束する．

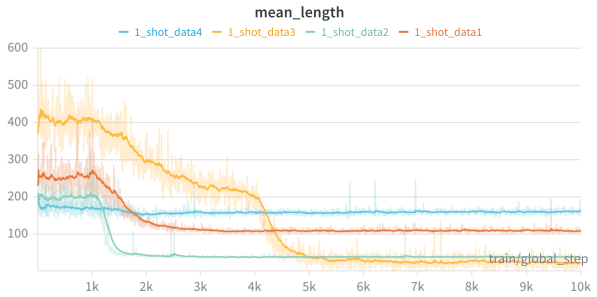

図 2　各学習データに対する平均出力長の推移．学習が進むにつれて出力長が短くなる傾向が確認できる．

加えて構成した．各問題に対して 30 回の回答を生成し，以下の 2 指標を測定した．

- **Accuracy**: 全 3000 回答中で正解した回答の割合
- **Solve Rate**: 各問題について 30 回中 1 回でも正解した問題の割合

Accuracy は「同じ問題をどれだけ安定して正解できるか」を反映し，Solve Rate は「そもそも解ける問題の種類が増えたか」をみる指標としている．

6　結果

6.1　学習過程

図 1 に reward の推移，図 2 に平均出力長の推移を示す．いずれの学習データでも，reward は概ね 1k〜2k step の間で急速に上昇し，その後ほぼ 1 に収束した．一方，mean_length は学習の進行とともに減少しており，モデルがより短く，定型的に正答へ到達するようになったことが示唆される．特にデータ 2 およびデータ 3 では出力長の減少が顕著であり，収束の速さや最終長には問題依存の差が見られた．

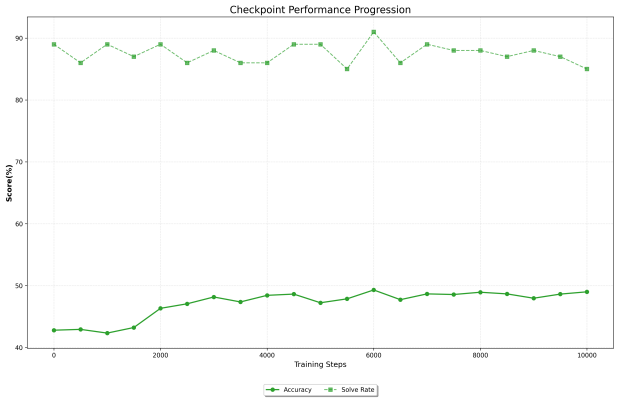

図 3　データ 1 で学習したモデルの評価推移

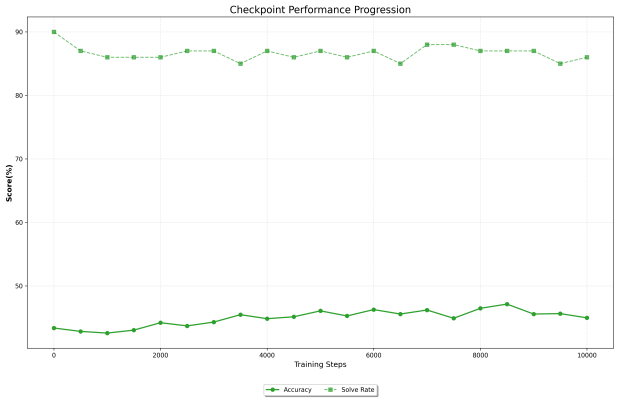

図 4　データ 3 で学習したモデルの評価推移

6.2　データ 1 とデータ 3 の評価比較

データ 1 で学習したモデルと，データ 3 で学習したモデルについて，評価セット 100 問に対する 30 回サンプリング評価の推移を図 3 および図 4 に示す．両者を比較すると，データ 1 で学習したモデルの方が評価指標の改善が安定している一方，データ 3 で学習したモデルでは改善が限定的な区間が見られ，チェックポイント間のばらつきも相対的に大きい．このことは，学習に用いる単一問題の選択が，強化学習後の性能向上の安定性に大きく影響することを示唆している．

一方で，GRPO 後モデルでは Accuracy が上昇したのに対し，Solve Rate はベースモデルから大きく変化しなかった．この傾向は，学習した全てのモデルで一貫して観測された．すなわち，強化学習によって「既に解ける問題をより高頻度で正解する」ようにはなる一方で，「従来解けなかった問題タイプを新たに解けるようになる」効果は限定的であると考えられる．

表 2　最終推論システムの結果

手法	精度
ベースモデル（単回答）	45%
ベースモデル（30 回答の多数決）	65%
本システム	71%

6.3　最終推論システム

最終的には，データ 1 で学習したモデルの 6000 step 重みを採用し，30 回答の多数決で最終回答を決定した．結果を表 2 に示す．単回答のベースモデルは 45%，ベースモデルに 30 回答の多数決を適用すると 65%であり，さらに GRPO 学習済みモデルを用いることで 71%に到達した．このことから，**強化学習と多数決の組み合わせ**が実用上きわめて有効であることが分かる．

7　考察

第一に，本結果は，近年の先行研究が主張する「強化学習は解ける問題のサンプル効率を高めるが，解ける問題の種類自体は増やしにくい」という見方と整合的である [4, 5]．本研究でも Solve Rate の改善は限定的であり，未解決問題の打破には別種のデータや学習機構が必要だと考えられる．

第二に，単一問題学習であっても，問題選定は重要である．今回比較した 4 問では reward 収束や出力長短縮のパターンが異なっており，データ 1 を用いたモデルが最終システムで最も良好な結果を与えた．どのような問題が GRPO による安定化を引き出しやすいかは，今後の重要課題である．

第三に，性能向上の大きな部分は推論時の多数決と結びついている．GRPO により回答分布が「正しい答えの周辺」により集中し，多数決がその改善を回収していると解釈できる．すなわち，学習時の分布改善と推論時の集約戦略が相補的に働いたと考えられる．

8　追加観察

本研究では，主結果以外にもいくつかの知見が得られた．

- 4 問すべてを同時に用いた GRPO で学習したモデルは，単一問題のみで学習したモデルより高い性能を示した（図 5）．
- 同じプロンプトを繰り返すとモデルの性能が上がるという Prompt repetition[6] はベースモデルでは有効だったが，GRPO 後モデルではむしろ性能を低下させた．
- インストラクションチューニング時のシステムプロンプト形式を少しでも崩すと，モデル性能が大きく低下した．
- Open R1-Math や GSM8K では明確なスコア向上が見られず，本手法の改善が対象分布に強く依存していることが示唆された．

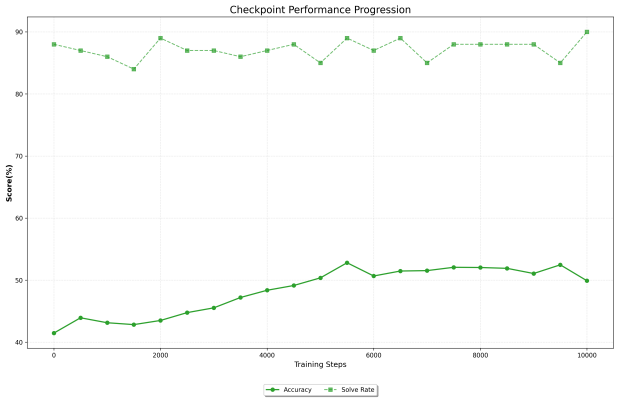

図 5　4-shot GRPO で学習したモデルの評価推移

9　おわりに

本稿では，数学の単一問題のみを用いた GRPO によって，日本語 LLM の数学的推論性能を改善できることを示した．学習曲線の分析からは，reward が早期に収束し，平均出力長が減少することが確認された．また，30 回サンプリング評価から，強化学習は主に既知の可解問題をより安定に正解する方向に働き，解ける問題の種類そのものを大幅に増やすわけではないことが示唆された．

一方で，推論時多数決と組み合わせることで，最終精度は 45%から 71%へと大幅に向上した．今後は，単一問題による GRPO で高い改善をもたらす問題の特徴分析，未解決問題を打破するために必要なデータ設計・学習アルゴリズムの解明を進めたい．

参考文献

[1] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Li, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024.

[2] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948, 2025.

[3] Y. Wang, T. Balasubramaniam, R. Istrate, S. Bansal, L. Wang, A. Khyatti, S. S. Singh, T. Goldstein, and M. Yaz-

dani. Reinforcement Learning for Reasoning in Large Language Models with One Training Example. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[4] Y. Yue, X. Liu, Z. Yang, D. Song, and A. M. Rush. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[5] K. Matsutani, Y. Sato, T. Kanda, and K. Inui. RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

[6] Y. Leviathan, M. Kalman, and Y. Matias. Prompt Repetition Improves Non-Reasoning LLMs. arXiv preprint arXiv:2512.14982, 2025.

[7] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.