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

Team Promptia

慶應義塾大学４年 田口昂樹

概要

- NLP2026のワークショップ「大規模言語モデルのファインチューニング技術と評価」で行われたコンペ（FT-LLM 2026）
- llm-jp-4-8bをベースモデルとして非公開テストセット500問の正答率で評価

戦略

数学の問題１問のみを用いたGRPO[1]

- 学習用合成データの作成や大規模データの準備が不要
- 提供された sanple promlemとデータ分布が乖離しない問題を設定しやすい（1問のみのため）
- 汎化性能の限界を検証したい

データ選定

ベーススモデルが正解率50％ほどの問題をピックアップ

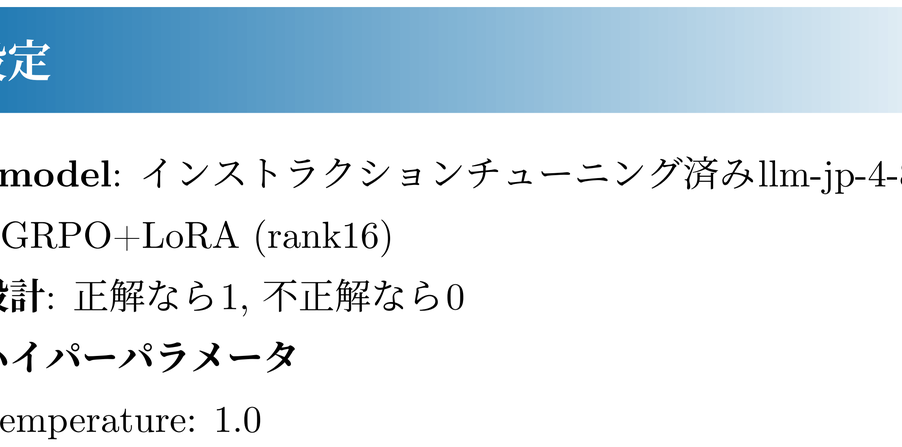
1. unit: 中2, category: 確率
   2つのさいころを同時に投げるとき、出る目の数の和が7になる確率を求めよ。
2. unit: 中3, category: 平方根
   次の数を $a\sqrt{b}$ の形に表しなさい。$\sqrt{24}$
3. unit: IIIC, category: 極限
   次の極限を求めよ。$\lim_{x\to\infty} \frac{\log x}{\sqrt{x}}$
4. unit: 中2, category: 一次関数
   点(0, 10)を通り、傾きが-5の直線の式を求めなさい。

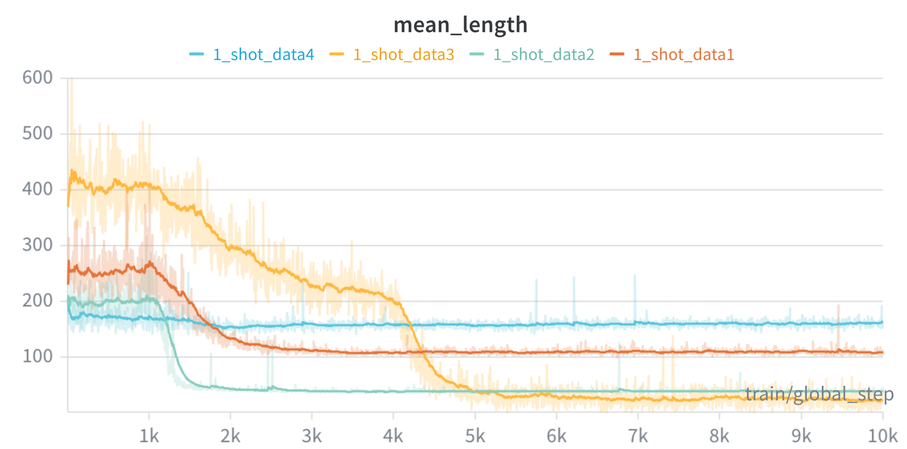
学習設定

- **base model**: インストラクションチューニング済みllm-jp-4-8b
- **学習**: GRPO+LoRA (rank16)
- **報酬設計**: 正解なら1, 不正解なら0
- **主なハイパーパラメータ**
  - temperature: 1.0
  - top_p: 1.0
  - max_tokens: 最大

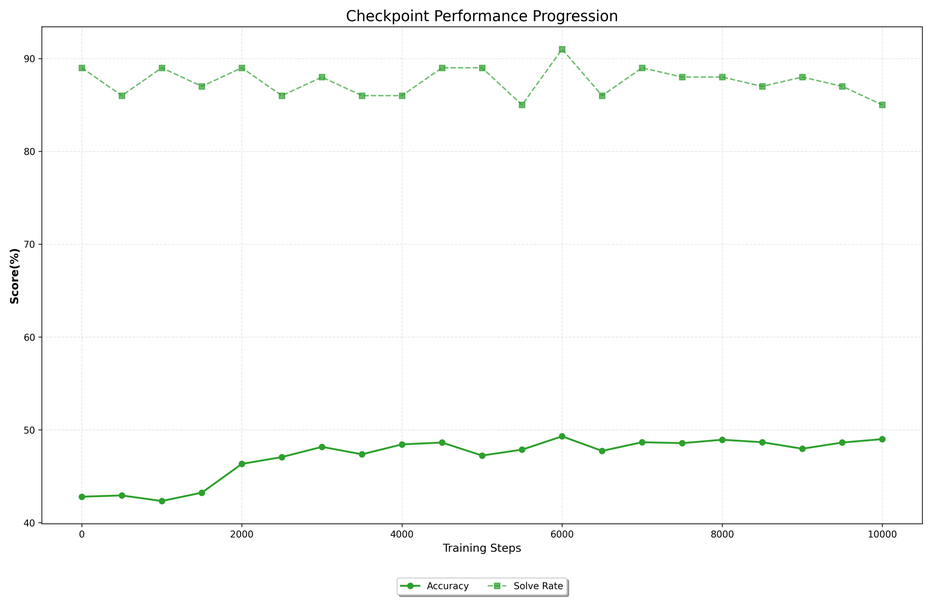
学習経過と評価

- **学習の推移:** 1k〜2k stepで報酬は1に収束．stepが進むにつれて出力トークン数は減少する傾向.





- **モデルの評価:** 評価セット100問（既存96問＋新規生成4問）の各問題に対して30回回答



- **Accuracy（実線）:** 3000問中正解した問題の割合は上昇
- **Solve Rate（点線）:** 30回中少なくとも1回正解する問題の割合はベースモデルと大きく変化せず、解けない問題の種類もほぼ同じ

最終推論システムと結果

- **推論システム構築:** データ1で学習したモデルの6000 stepの重みを採用．1つの問題に対して30回回答を生成し、多数決で最終回答を決定.

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%

考察と今後の展望

- **強化学習の限界の確認:** 強化学習は解ける問題のサンプル効率を高める一方、解ける問題の種類自体は増えにくいという先行研究[2,3]の主張との整合性を確認
- **性能向上の要因:** 強化学習と多数決の組み合わせにより、モデル性能を大幅に向上可能
- **その他の発見:**
  - Promptを２回繰り返す手法[4]はbase model以外には効果なし
  - インストラクションチューニングで用いたシステムプロンプトを少しでも変えると性能が大幅に劣化
  - GSM8Kのスコアは向上せず（本手法の限界）
- **今後の展望**: データ選定によって性能向上に差が出る理由の解明

[1] Y. Wang et al., ”Reinforcement Learning for Reasoning in Large Language Models with One Training Example,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2025.
[2] Y. Yue et al., ”Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2025.
[3] K. Matsutani et al., ”RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs,” in Proc. International Conference on Learning Representations (ICLR), 2026.

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