数学推論能力を向上させる大規模言語モデルのファインチューニングに関する研究

作成日 2025 年 3 月 5 日

折池雄太[*1] 恩田直登[*2] 平澤寅庄[*3] 園田翔[*4] 三内顕義[*5]

概要

本稿では，大規模言語モデル (LLM) を対象とした，数学推論タスクにおけるファインチューニング技術について報告する．とくに，問題文から解答に至る一連の推論過程を学習する手法として，SFT と強化学習を試みた事例を取り上げる．実際に進めたコンペティション (言語処理学会ワークショップ第 1 回「大規模言語モデルのファインチューニング技術と評価」) における取り組みや得られた知見をまとめ，今後の数学推論モデル開発に向けた展望を示す．

1 はじめに

大規模言語モデル（LLM）の性能向上により，様々な自然言語処理が可能になった．しかし，数学などの論理的思考を要する問題を解くことは依然として困難である．LLM は大量のテキストコーパスを用いた事前学習や教師ありファインチューニング（Supervised Fine-Tuning, SFT），さらに強化学習によって訓練される．この方法は自然言語生成には十分効果的であるが，段階的な論理的推論を必要とする数学的推論においてはまだ不十分な面がある．

数学の問題では，単に文脈に沿ったテキストを生成するだけでなく，正しい中間ステップを踏んだうえで最終的に正確な答えを導く必要がある．例えば，複雑な方程式を解く場合，LLM は各ステップで論理的に推論する必要があるが，SFT のみでは学習していない推論過程を導くことが困難だと考えられている．そこで，LLM の行動に報酬を与える強化学習を導入し，適切な推論過程を促す方法が注目されている．

本研究では，最終的な正解を報酬として利用する（accuracy reward に基づく）強化学習により，LLM の数学推論能力の向上を目指す。さらに，別のアプローチとして，推論ステップごとに報酬を付与する強化学習手法（step-by-step PPO）についても検討した結果を報告する。

2 関連研究

本研究に関連する研究として，LLM の数学推論能力を改善するための試みを記載する．

- Chain-of-Thought (CoT) プロンプティング (問題に対して最終回答を出す前に，中間的な推論ステップを生成するようモデルに促す手法) を用いると数学の正答率が向上する．(Wei et al., 2022)
- Process Reward Model (PRM) は，LLM が複雑なタスクを解く際の推論過程を監督する手法。従来の Outcome Reward Model（ORM）は最終結果の正誤のみを評価するが，PRM は各ステップの正誤をフィードバックすることで，LLM が論理的なミスを減らし，正確な推論を行うことを支援する．

[*1] AutoRes, folded.pond@gmail.com
[*2] オムロンサイニックエックス株式会社, n.onda.go20@gmail.com
[*3] オムロンサイニックエックス株式会社, tosho.hirasawa@sinicx.com
[*4] 理化学研究所, sho.sonoda@riken.jp
[*5] 京都大学・国立情報学研究所

(Uesato et al., 2022; Lightman et al., 2024)

- Math-Shepherd (Wang et al., 2024) では，PRM を用いた step-by-step PPO (推論ステップごとに強化学習を行う手法を) が提案された..
- DeepSeek-R1 (DeepSeek-AI, 2025) では，accuracy reward に基づく強化学習 (真の正解を報酬として利用する強化学習) が提案された.

3 取り組み概要

本節では，我々が取り組んだコンペティション（言語処理学会ワークショップ第 1 回「大規模言語モデルのファインチューニング技術と評価」）において実施した開発や議論をまとめる.

3.1 使用データセットとモデル

- **AI-MO/NuminaMath-CoT**：約 86 万の数学問題とその解に至る推論過程 (CoT) のペアを含む英語の大規模データセット.
- **llm-jp/llm-jp-3-13b**：コンペティションで指定されたベースモデル．英語と日本語のバイリンガルモデル.
- **Qwen/Qwen2.5-Math-7B**：数学向け事前学習済みモデル（作成したモデルとの比較実験用）．英語と中国語のバイリンガルモデル.

3.2 数学推論タスクへのファインチューニング

コンペティション用に作成したモデルは，llm-jp-3-13b を (1)SFT，(2) 強化学習の 2 段階で訓練したものである．このモデルの性能を比較するために，Qwen2.5-Math-7B-Instruct に対して (2) 強化学習を行った.

- **(1) SFT**：NuminaMath-CoT を用いて SFT を 2epoch で実施した．[*6]
- **(2) 強化学習**：DeepSeek-R1 で用いられている accuracy reward に基づく GRPO による強化学習を行った．その際ライブラリには Hugging Face で公開されているオープンソースソフトの open-r1 を用いた.

3.3 実験結果

上記手法により学習したモデルの性能比較のため，日本語訳した MATH のデータを用いて正答率を確認した.

- **llm-jp-3-13b + NuminaMath-CoT (SFT のみ)**：正答率 19% を確認.
- **SFT 後に GRPO 実施**：正答率 19% から学習の進行とともに 13% まで低下した.

[*6] NuminaMath-CoT は英語のデータセットである．今回のコンペティションは，日本語で記述された数学の問題を llm-jp-3-13b をベースとするモデルで解答するものであるため，理想的には，日本語訳をした NuminaMath-CoT で llm-jp-3-13b を SFT することが望ましいと考えられる．このデータセットの翻訳を検討したが，翻訳コストや翻訳精度の問題があったため，データセットを翻訳せず英語のまま llm-jp-3-13b に SFT で学習させた.

- **Qwen2.5-Math-7B-Instruct + GRPO**：学習初期に正答率が 65%→70% と上昇.

Qwen-Math は数学のデータセットで十分に事前学習しているため，llm-jp とは異なり，強化学習時に正答率が向上したと考えられる.

4 その他の取り組み

コンペティションに提出するモデルを検討するために行ったその他の取り組みについてまとめる.

4.1 KL ペナルティによる安定性

コンペティションに提出するモデルと比較するために Qwen2.5-Math-7B を学習させたところ，KL ペナルティの強度 β の違いにより，学習の安定性が異なりうる（β を 0 にすると，学習とともにモデルが崩壊する）ことが示唆された．なお，3 章での GRPO は $\beta = 0.04$ で実施している.

- **Qwen2.5-Math-7B-Instruct + GRPO ($\beta = 0.04$)**：学習初期に正答率が 65%→70% と上昇.
- **Qwen2.5-Math-7B + GRPO ($\beta = 0.00$)**：学習初期に正答率が 12%→28% と上昇したが，その後モデルが崩壊し正答率が 0 になった.

4.2 step-by-step PPO

accuracy reward に基づく強化学習とは別に，推論ステップごとに強化学習を行う手法を実施した．具体的には，モデルの出力を`\n\n` 単位で推論ステップとして区切り，各ステップについて PRM による報酬に基づく PPO を実行した．しかし，以下のような課題が生じた：

- **policy と PRM の tokenizer の一致**：llm-jp は数学に関するデータセットを用いた事前学習が十分に施されていないと予想し，まずは Qwen-Math ベースで step-by-step PPO を実行した後に，llm-jp ベースで step-by-step PPO を実行しようとした．前者は実行できたが，後者を実行しようとした際，PRM と policy で tokenizer を一致させる処理 [*7] に非常にコストがかかることが予想されたため，これを断念した.
- **報酬ハッキング**：Qwen-Math ベースで step-by-step PPO を実施したところ，PRM の精度の問題により不自然な短文（“So,\n\n”など）でも報酬が高くなることがあった．このような不適切な報酬に基づく強化学習はモデルの推論能力を悪化させる懸念がある.

5 結論

本稿では，コンペティション参加を通じて実践した数学推論能力向上のための LLM のファインチューニング事例を報告した．SFT のみでも一定の数学推論能力は得られるが，強化学習においては，報酬ハッキングや

[*7] policy を llm-jp/llm-jp-3-13b とする際は，tokenizer が一致した PRM を使用する必要がある．llm-jp をベースとした PRM の作成には，(1) 大量の数学データセットを用いた llm-jp の事前学習，(2) 十分な量の Process Supervision Dataset (trl-lib/prm800k や trl-lib/math_shepherd のような，推論ステップとその良否を含むデータセット) の用意，(3)Process Supervision Dataset を用いたベースモデルの SFT，という 3 つのプロセスが必要.

推論能力の低下，学習不安定化が課題であることを確認した．今後は，更なる数学推論能力の向上や，形式言語への拡張，堅牢な数理推論モデルの実現を目指したい．

6 謝辞

本研究は内閣府ムーンショット型研究開発事業（Moonshot プロジェクト）と JST さきがけ JPMJPR2125 の助成を受けたものです．

参考文献

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint: 2501.12948*, 2025.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*, 2024.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process- and outcome-based feedback. In *2nd MATH-AI Workshop at NeurIPS'22*, 2022.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.