日本語 LLM のローカル強化学習と学習過程の検証

Local Reinforcement Learning of Japanese Large Language Models and Verification of the Training Process

山口大空，曽我真人
Sora Yamaguchi, Masato Soga
和歌山大学システム工学部
Faculty of Systems Engineering Wakayama University
Email: s276278@wakayama-u.ac.jp

あらまし：近年，大規模言語モデル(LLM)は広く利用されているが，商用モデルの多くは学習過程が非公開であり，挙動の検証が困難である．本研究では，管理可能なローカル環境において，日本語 LLM を対象に人間の選好を反映した強化学習(RLHF)を段階的に実施し，生成モデルの挙動変化を検証した．報酬モデル(RM)と 近接方策最適化(PPO)学習を用いた比較実験の結果，評価基準に沿った応答生成の改善が確認された一方，その効果は入力全体に一様には及ばず，生成が変化しないケースも多く観測された．
キーワード：大規模言語モデル(LLM)，　人間の選好を反映した強化学習(RLHF)，報酬モデル(RM)，近接方策最適化(PPO)，LoRA

1. はじめに

ChatGPT に代表される大規模言語モデル(LLM)は広く普及しているが，その多くは学習データや学習手順が非公開であり，モデルの挙動や学習結果を外部から検証することが難しいという課題を持つ．

本研究では，日本語 LLM を対象として，学習過程の検証が可能なローカル環境を構築し，人間の選好を反映した強化学習(RLHF)による学習を通じて，モデル挙動の変化を定量的に評価することを目的とする．

また，本研究では高齢者との回想会話を想定した認知症予防カウンセリングを応用先として設定し，その利用を見据えた対話方針および学習設計を行ったが，実運用や利用者評価は今後の課題とする．

2. 学習手法

一般に RLHF は，事前学習済み言語モデルに対して教師あり微調整(SFT)を行い，さらに人間の比較評価を学習した報酬モデル(RM)と，その報酬を用いた近接方策最適化(PPO)による強化学習によって生成モデルを更新する一連の枠組みを指す. 本研究では，SFT までを完了した日本語 LLM を出発点とし，RM および PPO による学習過程に着目してローカル環境での実装と評価を行った．

2.1 LoRA による学習

本研究では，計算資源の制約を考慮し，学習には LoRA を用いたパラメータ効率的学習を採用した．基盤モデルの重みは固定し，LoRA によって追加された低ランクパラメータのみを学習対象とすることで，学習による影響を局所的に制御できる構成とした．

2.2 報酬モデル(RM)の構築

RM は，人間による応答比較結果をもとに，生成された応答の好ましさをスカラー値として出力するモデルである．本研究では，同一プロンプトに対して生成された二つの応答を比較し，好ましい応答を chosen，もう一方を rejected とした pairwise 形式のデータを用いた．

学習では，chosen が rejected より高いスコアを出力するようにモデルを更新し，相対的な優劣関係を学習させた．構築した RM は，主に後述する PPO 学習において報酬として用いられる. また, 本研究では，PPO 学習を段階的に実施する構成としたため，それに対応して RM についても段階的な学習を行った．

なお，RM の学習においては，最終学習段階のモデルを一律に用いるのではなく，各学習段階における学習状況を踏まえ，報酬としての利用に適していると判断した学習途中の checkpoint を採用した．

2.3 近接方策最適化(PPO)による生成モデルの更新

生成モデルの更新には，PPO を用いた．各学習段階において，構築済みの RM を報酬関数として使用し，RM の示す評価基準に沿うよう生成モデルの出力分布を更新した．

PPO 学習は段階的に実施し，各段階で得られた生成モデルを評価したうえで，改善されたモデルのみを次段階に引き継ぐ構成とした．これにより，生成モデルの学習を安定させつつ，各学習段階における応答傾向の変化を段階的に確認できる構成とした．

3. 実験と評価

3.1 評価方法

本研究では，PPO によって得られた生成モデルの挙動変化を確認するため，学習に用いた RM を評価器として用いた．評価の目的は，PPO 学習によって生成モデルの応答傾向が，RM が表現する評価基準

に沿って実際に変化したかを確認することにある．

評価では，学習前および各 PPO 学習段階の生成モデルに同一プロンプトを入力し，RM が出力する報酬値の平均を指標として応答傾向の変化を比較した．また，評価結果の傾向を直感的に把握するため，RM による応答比較結果を win，loss，tie の三値に分類して併せて集計した．

3.2 評価結果

本研究では，PPO 学習によって生成モデルの応答傾向がどのように変化したかを確認するため，報酬モデル RM を用いた評価を行った．また，評価にあたってすべての生成を greedy デコードで実施しており，同一入力に対する base 出力は常に同一である．これにより，評価結果はサンプリングによる偶然的な差ではなく，PPO 学習による生成分布の変化そのものを反映したものとなっている．

表 1 に，学習前の生成モデル(base)および PPO-1 から PPO-4 までの各学習段階について，共通の評価基準として RM1 を用いて算出した評価結果を示す．なお，表中の win，loss，tie は，いずれも各 PPO 段階の生成モデルを学習前の base モデルと比較した結果であり，PPO 学習段階同士の直接比較ではない．

表 1 RM1 による PPO 学習段階ごとの評価結果

	mean_score	win	loss	*勝率	tie
base	-0.994	-	-	-	-
PPO-1	-0.983	205	196	51.12%	1599
PPO-2	-0.988	190	168	53.06%	1642
PPO-3	-0.970	198	156	55.92%	1646
PPO-4	-0.988	228	193	54.15%	1579

*ここでの勝率は base と比較して生成出力が変化した比較のみを対象とし，その中で win を 1，tie を 0.5，loss を 0 として算出した割合である．

表 1 に示すように，PPO-1 では学習前モデルと比較して報酬値の平均(mean_score)の上昇が確認され，RM が表現する評価基準に沿った応答生成が進んでいることが示された．このため，PPO-1 は次段階の学習に引き継ぐモデルとして採用した．

一方，PPO-2 では win の割合が増加する傾向が見られたものの，報酬値の平均は安定して向上せず，tie の数値も上昇していることから総合的な判断により，本研究では PPO-2 を次段階の学習および最終的な生成モデルとしては採用しなかった．

PPO-3 では，報酬値の平均が安定して向上するとともに，win および loss の傾向も改善しており，RM が表現する評価基準に沿った応答生成が最も一貫して確認された．この結果を踏まえ，本研究では PPO-3 を最終的な生成モデルとして採用した．

一方，PPO-4 では，追加的な PPO 学習を行ったものの，報酬値の平均および応答傾向の安定性の観点から，PPO-3 を上回る明確な改善は確認されなかった．そのため，PPO-4 は最終採用には至らなかった．

また，RM による応答比較を詳細に確認したところ，PPO-1 から PPO-4 のいずれの学習段階においても tie に分類された応答の大部分は，学習前後で生成内容自体が変化していなかった．

これらの評価結果とは別に，報酬値や win／loss に差が見られた生成結果については，RM の学習データを作成した著者自身が内容を直接確認し，順位付けを行った結果，RM 作成時に用いた判断基準と RM による採点結果との間に大きな乖離が生じていないことを補助的に確認している．

なお，本研究における生成モデルの評価は，PPO 学習段階に応じて複数の RM を用いて実施しているが，本章では，各学習段階を共通の基準で比較するため，すべての段階に共通して用いられている RM1 による評価結果のみを示している．また，表中の mean_score は RM1 が出力する相対的な報酬値の平均であり，値の正負そのものではなく，学習段階間の相対的な差を評価するための指標である．

4. 考察・まとめ

評価の結果，PPO 学習は RM が表現する評価基準に沿った方向への調整を可能にする一方で，その効果がすべての入力に一様に現れるわけではないことが確認された．

この結果は，PPO 学習が失敗していることを意味するものではなく，本研究における PPO による改善は，報酬モデルの評価基準に沿った方向付けとしては機能しているものの，生成結果全体に及ぼす影響は部分的にとどまっていると考えられる．

以上の結果から，本研究で用いた学習設計および評価条件の下では，RLHF による PPO 学習は，報酬モデルの評価基準に沿った方向への調整を可能にするものの，生成モデルの振る舞いが対話全体を大きく変化させる段階には至っていないことが明らかとなった．とくに，高齢者の認知症予防を目的とした対話支援への応用を想定した場合，本研究で確認された改善は，実際の会話体験や支援効果として安定的に現れる水準には達していないと考えられる．

参考文献

(1) NRC デイリートラッキング 生成 AI の利用経験 2025 3 月調査, https://www．nrc．co．jp/report/250414．html
(2) 国立情報学研究所 大規模言語モデル研究開発センター,「LLM-jp-3 シリーズ instruct2，instruct3 の公開」 https://llmc．nii．ac．jp/topics/llm-jp-3-instruct2-3/,
(3) EEdward Hu，Yelong Shen，Philip Wallis，Zeyuan Allen-Zhu，et al.，LoRA: Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2106.09685，2021
(4) Long Ouyang，Jeff Wu，Xu JiangJiang,et al.，Training Language Models to Follow Instructions with Human Feedback, arXiv preprint arXiv:2203.02155，2022
(5) Nisan Stiennon，Long Ouyang，Jeff Wu，et al.，Learning to Summarize with Human Feedback，arXiv preprint arXiv:2009.01325，2020
(6) John Schulman，Filip Wolski，Prafulla Dhariwal，Alec Radford，Oleg Klimov Proximal Policy Optimization Algorithms, arXiv preprint arXiv:1707.06347，2017.