REINFORCE++:プロンプトモデルと報酬モデルの両方に対して堅牢な効率的なRLHFアルゴリズム

江胡
janhu9527@gmail.com

ジェイソン•クライン•リュージャソンクレインラブ@gmail.com

ウェイ•シェン
shenwei0917@126.com

抽象的な

人間のフィードバックからの強化学習（RLHF)は、大規模言語モデル（LLM)を人間の価値観や嗜好に整合させる上で極めて重要です。ChatGPT/GPT-4のような最先端のアプリケーションでは、一般的に近似ポリシー最適化（PPO)が採用されていますが、批評ネットワークを含めると、計算オーバーヘッドが大幅に増加します。REINFORCE Leave One-Out （RLOO）、ReMax、Group Relative Policy Optimization（GRPO)などのREINFORCEベースの手法は、批評ネットワークを排除することでこの制約に対処しています。しかし、これらの手法は、正確な優位性推定において課題を抱えています。具体的には、各プロンプトへの応答ごとに優位性を個別に推定するため、より単純なプロンプトへの過剰適合や報酬ハッキングに対する脆弱性につながる可能性があります。これらの課題に対処するため、バッチの正規化された報酬をベースラインとして用いながら批評モデルを排除する、新たな手法であるREINFORCE++を紹介します。実証的評価により、REINFORCE++はプロンプトセットの切り捨てを必要とせずに、様々な報酬モデルにわたって堅牢なパフォーマンスを発揮することが実証されています。さらに、 REINFORCEベースの手法と比較して、 RLHFおよび長い思考連鎖（CoT)設定において優れた一般化を実現します。実装はhttps://github.com/OpenRLHF/OpenRLHFで入手可能です。

1 はじめに

人間からのフィードバックによる強化学習（RLHF)は、大規模言語モデル（LLM)を人間の価値観や嗜好に整合させるための重要な手法です（Vemprala et al., 2023; Achiam et al., 2023 ; Ouyang et al., 2022a; Shen & Zhang, 2024）。DPO （Rafailov et al., 2023)のような非RL代替手法が登場しているにもかかわらず、 ChatGPT/GPT-4 （Vemprala et al., 2023; OpenAI, 2023）、Claude （Anthropic, 2023）、 Gemini （Team et al., 2023)などの最先端のアプリケーションは、方策最適化においてRLアルゴリズム、特にPPOに依存し続けています。しかし、PPO (Schulman et al., 2017)は批判的ネットワークを必要とするため、計算オーバーヘッドとメモリ使用量が大幅に増加し、小規模クラスターにおける大規模モデルのアライメントが制限されます。この問題に対処するため、研究者らは批判的ネットワークを排除する様々なREINFORCEベースの手法を提案してきました。具体的には、ReMax (Li et al., 2023)、 REINFORCE Leave One-Out (RLOO) (Wu et al., 2024)、 Group Relative Policy Optimization (GRPO) (Shao et al., 2024) などが挙げられます。さらに、Deepseek-R1は、長形式CoT設定におけるREINFORCEベースの手法の有効性を実証しており(Guo et al., 2025)、ルールベース報酬を用いたGRPOを用いて、難易度の高いデータセットで最先端のパフォーマンスを達成しました。

批評ネットワークがなければ、REINFORCEベースの手法では、個々のトークンの利点を正確に推定することがしばしば困難になります。この限界に対処するために、様々なREINFORCEベースラインアプローチが提案されていますが、それぞれに重大な欠点があります。ReMaxは、各プロンプトに対して貪欲な探索を用いて応答を生成し、その報酬をベースラインとして用いるため、ベースライン計算のためだけにモデル応答を消費するという非効率的な方法を採用しています。RLOOとGRPOは、プロンプトごとに複数の応答を生成するという異なるアプローチを採用しています。RLOOは他の応答の平均報酬をベースラインとして用いるのに対し、GRPOはすべての応答にわたって正規化された報酬を用います。これらの手法は利点の推定精度を向上させますが、プロンプトごとに複数の応答を最適化するという手法は、リスクを高めます。

責任著者

報酬ハッキングの危険性。さらに、これらの手法はプロンプトごとに報酬のベースラインを個別に計算するため、最適化中に特定の訓練プロンプトにおいて過学習や不安定性が生じる可能性があります。そのため、これらの手法では、各タスクに固有のプロンプトセットを慎重に選定する必要があります。

これらの課題に対処するために、我々はPPO から批評家モデルを排除し、グローバル バッチの平均報酬をベースラインとして使用する新しい REINFORCE ベースの方法である REINFORCE++ を提案します。
このアプローチは、特定の訓練プロンプトへの過剰適合を防ぎ、 Bradley-Terryモデルとルールベース報酬モデルの両方において堅牢性を示します。特に、REINFORCE++はプロンプトセットの切り捨てを必要とせず、RLHFとlong CoT RLの両方の設定において高い汎化性能を実現します。

この論文の残りの部分は次のように構成されています。セクション2 では、さまざまな REINFORCE ベースの RLHF 手法の違いを分析し、プロンプト固有の報酬ベースラインの無効性と、RLOO および GRPO 手法における過剰適合の問題について示します。セクション3 では、REINFORCE++ を紹介し、その実装と利点を詳しく説明します。セクション 4 では、包括的な実験を通じて REINFORCE++ を評価します。最初に、 Bradley-Terry およびルールベース報酬モデルの両方を使用して、REINFORCE++ を他の RLHF 手法と比較し、その優れた、または同等のパフォーマンスを示します。次に、長文の Chain-of-Thought (CoT) 設定で、計算上の制約により、 REINFORCE++ と現在の最先端の REINFORCE ベースの手法である GRPO を比較することに焦点を当てます。厳選された少数のプロンプトでの実験を通じて、特定のプロンプトへの過剰適合における GRPO の弱点を明らかにします。その後、長文CoTタスクを用いた実験を行い、REINFORCE++が難易度の高いテストデータセットにおいて優れたパフォーマンスを発揮し、分布外（OOD)汎化能力が向上していることを示しました。第5章では、本手法の結論と今後の研究方向性について述べます。

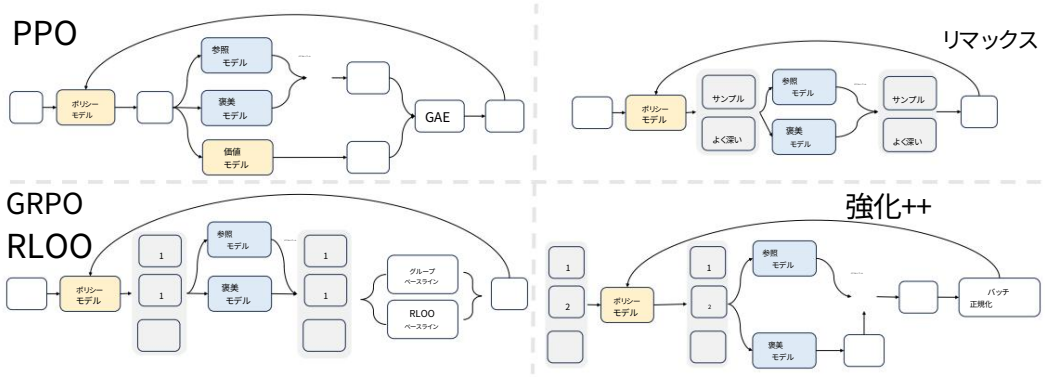

図 1: PPO、Remax、GRPO、RLOO、REINFORCE++ の比較。

2 背景と関連研究

ChatGPT/GPT-4、Claude、Geminiなどの最先端のアプリケーションでは、PPOなどの強化学習アルゴリズムを方策最適化に利用しています。特に、PPOは、以下の代理目的関数を最大化することでLLMを最適化します。

$$LPPO(\theta) = E_{q \sim P(Q),\, o \sim \pi_{\theta old}(O|q)} \frac{1}{|お|} \sum_{t=1}^{|お|} \min\left(st(\theta)A_t,\ Clip(st(\theta), 1-\ ,\ 1+\ )A_t\right) \tag{1}$$

どこ：

$$st(\theta) = \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\theta old}(o_t|q, o_{<t})} \tag{2}$$

PPOは、アクター•クリティックに基づく強化学習（RL)アルゴリズムであり、クリティックモデルが大量の学習リソースを消費します。そのため、多くの研究者が一連のPPOを提案してきました。

REINFORCEベースの手法（ReMax、RLOO、GRPOなど)は、批評モデルに伴う計算オーバーヘッドを回避しながら、比較的正確なトークンごとの優位性推定値を得ることができます。これらの手法は、各プロンプトのベースライン報酬を優位性推定値として計算する代替手法を設計しています。

具体的には、 PPOにおけるベースラインに対する行動のパフォーマンスを評価するために、アドバンテージが計算されます。まず、状態、行動、報酬、そしてそれに続く状態のサンプルを収集します。次に、一般化アドバンテージ推定（GAE)を用いてアドバンテージを計算します。GAEは、一連の時間ステップにおける時間差誤差δq,ot = rt + γV(ot+1) − V(ot)を統合します。

$$A_{q,o_t} = \sum_{l=0}^{\infty} (\gamma\lambda)^{l} \delta_{t+l} \tag{3}$$

ここで、 λはバイアスと分散のバランスをとるパラメータです。PPOは、正確なアドバンテージ推定値を取得し、未知のトークンに一般化するために、批評モデルを用いてアドバンテージ関数を学習し、正確なアドバンテージ推定値を取得し、未知のトークンに一般化する必要があります。

図1に示すように、ReMaxは貪欲なデコード法を用いて応答を生成し、その報酬をこのプロンプトのベースライン報酬として取得する。したがって、クエリqの利点は次のように表される。

$$A_{q,o_t} = r(o_{1:t}, q) - r(o^{*}_{1:t}, q) \tag{4}$$

どこ：

$$o^{*}_{1:t} = \underset{o'_{1}, \ldots, o'_{t}}{\operatorname{argmax}} \prod_{i=1}^{t} \pi_\theta(o'_i \mid q, o'_{<i}) \tag{5}$$

特に、ReMax はこの応答をモデルのトレーニングに使用しません。

RLOOとGRPOはどちらも、プロンプトに対して複数の応答をサンプリングします。RLOOは、現在のプロンプトに対する他のすべてのサンプルの平均報酬をベースラインとして採用します。これは式6に示されています。

$$A^{(i)}_{k-1\, q,o_t} = r(o^{(私)}_{1:t}, q) - \frac{1}{\ } \sum r(o^{(j)}_{1:t}, q) \tag{6}$$

一方、GRPO は、グループ相対的優位性推定法を採用しており、平均報酬を現在のプロンプトに対するすべてのサンプル応答の標準偏差で割ったものをベースライン報酬として使用します (式 7 に示されています)。

$$A^{(i)}_{q,o_t} = \frac{r(o^{(i)}_{1:t}, q) - \mathrm{mean}(\{r(o^{(j)}_{1:t}, q)\}_{j=1}^{k})}{\mathrm{std}(\{r(o^{(j)}_{1:t}, q)\}_{j=1}^{k})} \tag{7}$$

しかし、多様なデータセットを考慮すると、各プロンプトのベースライン報酬はRLHF設定において必須ではないと考えられます。各プロンプトのベースライン報酬は、各トレーニングプロンプトに対して比較的正確な優位性の推定を提供し、モデルがプロンプト下で最も高い報酬をもたらす応答を学習するのに役立ちますが、報酬ハッキングや過学習の問題を加速させてしまいます。RLHFには、従来の強化学習の問題とは2つの重要な違いがあることに留意してください。

- 従来の RL 問題では、RL ポリシーをトレーニングし、同じ環境でテストします。
  ただし、RLHF メソッドではプロンプト セットをトレーニングし、別のデータセット (分布外 (OOD) データセットを含む) でテストします。
- 従来の強化学習問題では、常にゴールデン報酬が存在します。対照的に、RLHF法では常に報酬モデルまたはルールベースの報酬が使用されるため、報酬ハッキングが発生する可能性があります。

したがって、報酬ハッキングと過剰適合の問題はどちらもモデルの性能を低下させる可能性があります。具体的には、RLOOやGRPOなどの手法を用いて、バッチ内のプロンプトに対する複数の応答を最適化すると、特定の単純なプロンプトにおいて最良の応答が過剰適合する傾向があり、最終的にはモデルの性能を低下させます。

一般化。さらに、単一バッチ内でプロンプトに対する複数の応答を最適化すると、モデルの出力の多様性が低下し、トークンレベルの優位性の分布の多様性が低下し、それらのトークンに過適合が発生します。対照的に、PPOではこの問題の影響は大きくありません。なぜなら、価値ネットワークは継続的に学習され、学習した優位性を維持するため、より一般化可能なトークン単位の優位性をサポートできるからです。

したがって、特定のプロンプトの過剰適合を回避し、トレーニング バッチ内のプロンプトの多様性を高めるために、 REINFORCE++ はプロンプトごとに 1 つの応答をサンプリングし、グローバル バッチ サイズ内でトークン単位の利点を正規化して、トレーニングの安定性を高めることができます。

3 方法

REINFORCE++ は PPO 目標を最適化し、式 1 で定義されたクリッピング戦略を採用します。
勾配推定値の分散をさらに低減し、学習プロセスをより安定的かつ効率的にするために、グローバル学習バッチの平均報酬をベースライン報酬として採用します。したがって、 REINFORCE++の利点は、強化学習における正規化された報酬の活用にあります。

$$A_{q,o_t} = r(o_{1:t}, q) - \beta \cdot \sum_{i=t}^{T} KL(i) \tag{8}$$

どこ：

$$KL(t) = \text{対数} \frac{\pi^{RL}_{\theta_{old}}(o_t \mid q, o_{<t})}{\pi_{SFT}(o_t \mid q, o_{<t})} \tag{9}$$

トークンレベルのKLペナルティの勾配は、 RLHFにおけるGRPOのk3損失に関して、理論的に偏りがないことが証明されています。さらに、この利点を全てのプロンプトについてグローバルバッチ全体で正規化します。

$$\text{質問}^{\text{規範}} = \frac{A_{q,o_t} - \text{平均}A_{q,o_t}}{\text{std}A_{q,o_t}}、 \tag{10}$$

REINFORCE++は、プロンプト固有の報酬ベースラインを削除し、トレーニングバッチ内の単一のプロンプトからの複数の応答を最適化しません。PPOと比較すると、REINFORCE++は批評モデルを削除し、割引係数を1に設定します。アルゴリズムの詳細な実装はアルゴリズム1に示されており、そこでアルゴリズムの詳細についてさらに詳しく説明します。

アルゴリズム 1 REINFORCE++要件:初期ポリシーモデル

πinit、報酬モデル R、タスクプロンプトD 1:ポリシーモデルπθ ← πinit 2:ステップ= 1、...、Mを実行するDからバッチDbをサンプリングする

3:古いポリシーモデルを更新πold ← πθ質問q ∈

Dbごとに1 つの出力 o πold(· | q)をサンプリン

さ グするR を実行して、出力oiのバッチでサンプリング

4 : れた各サンプルの報酬riを計算する反復= 1、...、tに対してA q,ot を計算する

5: REINFORCE++ の目的関数 (式1)を最大化するようにポリシーモデルπθ を更新する10: 11: πθの終了保証: πθ

6:

7: 規範 プロンプトqのoのt番目のトークンについて、式8から式10まで

8:

9:

4つの実験

REINFORCE++の実証的評価は、多様なテストシナリオで実施され、そのパフォーマンスを包括的に評価しました。Bradley-Terry報酬モデルを用いた実験では、REINFORCE++をReMax、RLOO、GRPO、PPOなどの既存手法と徹底的に比較しました。計算上の制約により、長期CoT設定における実験では、REINFORCE++と、現在最先端のREINFORCEベースの手法であるGRPOとの比較に焦点を絞りました。

4.1 成果報酬モデルによるパフォーマンス

実験設定本研究では、一般領域における人間によるフィードバックを用いた強化学習（RLHF)を調査する。実験プロセスは、一般領域の言語タスクを反映した多様なデータセットを用いて微調整された指示追従型ポリシーモデルから開始された。その後、報酬モデルに基づく強化学習アルゴリズムを用いて、このモデルは改良された。具体的には、 Bradley-Terry報酬モデルを、モデル出力のペアワイズ比較によって得られた人間生成の選好データを用いて学習させた。報酬モデルは、ポリシーモデルの応答を有用性、正確性、一貫性、および人間の意図との整合性に基づいて評価し、各応答に対して結果報酬値を出力した（ Bai et al. , 2022）。

報酬モデルOuyang et al. (2022b)が提案したアプローチに従い、 SFT モデルを用いて報酬モデルを初期化する。選好学習に適応させるため、最終層をスカラー出力を生成する線形ヘッドに置き換える。[1]報酬モデルは、以下のように定義される負の対数尤度損失関数を用いて学習される。

$$\mathrm{LRM}(\theta) = -E(q,o^{+},o^{-}) \quad D \log \sigma r_{\theta}(q, o^{+}) - r_{\theta}(q, o^{-}) \tag{11}$$

データセット報酬モデルの学習には、複数の公開データセットから集約された約70万組の人間の嗜好データ[2] からなる広範なデータセットを活用しました。これらのデータセットは、コンテキストと嗜好の多様性に富み、報酬モデルが微妙な人間の判断を効果的に捉えることを可能にします。ポリシーモデルに応答生成を体系的に促すため、多様なソース[3]からサンプリングした2万件のプロンプトを慎重にバランスよく収集し、様々なシナリオとドメインを網羅的にカバーできるようにしました。この多様性により、ポリシーモデルによって生成される応答の堅牢性と一般化が促進されます。

	トークンスコアあたりのスコアの長さ		
REINFORCE++ 46.7 REINFORCE++	ベース	832	0.0561
ライン 44.2 GRPO 46.8 RLOO 44.6	ReMax	834	0.0530
45.1		860	0.0544
		866	0.0515
		805	0.0560

表1 :GRPOとREINFORCE++のスコアと長さの比較。各報酬モデルで優れた結果は太字で強調表示されています。

実験結果 :モデルの評価には、 Chat-Arena-Hard Li et al. (2024)を用いた。表1に示すように、GRPOはREINFORCE++の46.7に対して、総合スコア46.8とわずかに優れた結果となった。しかし、GRPOは平均860トークンという長いシーケンスを生成するのに対し、REINFORCE ++はわずか832トークンという短い出力しか生成しない。結果として、トークン単位で性能を評価すると、REINFORCE++はGRPOを上回り、トークン単位のスコアは0.0561、GRPOは0.0544と高い値を示した。この表は、評価モデルが長さに偏りがあるにもかかわらず、 REINFORCE++の方がより効率的な出力を提供していることを示唆している。

1https://huggingface.co/OpenRLHF/Llama-3-8b-sft-mixture 2https://huggingface.co/datasets/hendrydong/preference_700K 3https://huggingface.co/datasets/RLHFlow/prompt-collection-v0.1

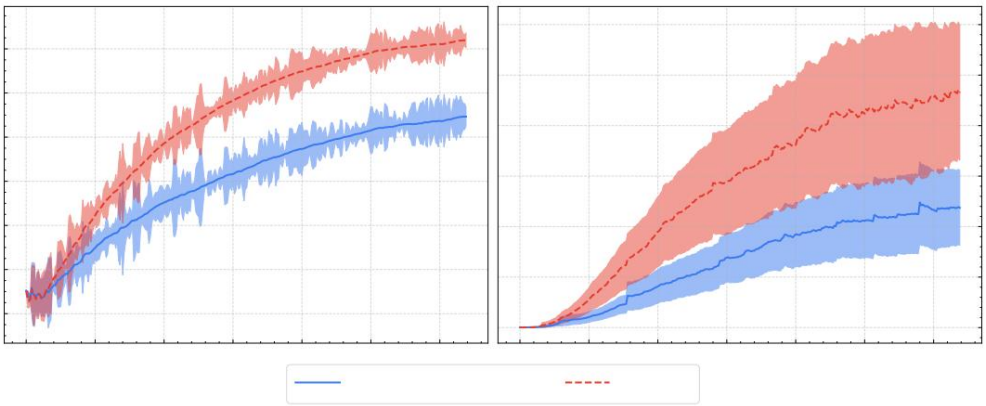

図2: 平滑化された訓練報酬とKLにおけるGRPOとREINFORCE++の比較
結果報酬モデルとの相違。

結果分析図2にテスト報酬とKLダイバージェンス曲線をプロットし、明確な
GRPOとREINFORCE++の比較挙動に関する洞察。当初、GRPOは
トレーニング全体を通して、報酬が大幅に増加し、REINFORCE++よりも優れたパフォーマンスを発揮します。しかし、
調査の結果、GRPO によって達成される優れた報酬は主に報酬ハッキングによるものであることが明らかになりました。
具体的には、GRPOのKLダイバージェンスの急激な増加は、モデルが
報酬信号は一般化性能を向上させるのではなく、報酬を膨らませる結果となる。
テストセットでは対応するゲインがない値を示す。対照的に、REINFORCE++はより
報酬は緩やかだが安定的に増加し、KLダイバージェンスも緩やかに上昇する。その結果、
好ましいトレードオフを示しており、報酬が大幅に改善される。
参照モデルからの逸脱は最小限である。KLダイバージェンスの増加は比較的小さい。
REINFORCE++では、各分岐ユニットがより効果的で
政策の改善はKLから報酬への変換効率の向上を示唆している。
我々の主張は、RLHFの設定で分布外タスクでモデルを評価した。
主に数学の問題タスク（GSM8K、MATH）とコード生成タスクが含まれます。
（HumanEval、MBPP）。評価結果を表2に示す。

	GSM8K	MATH	HumanRval	MBPP	平均	
強化++	96.21	75.20	85.98	84.39	85.45	
強化++ベースライン	95.98	72.40	78.66	85.45	83.12	
GRPO	96.21	73.80	80.49	83.33	83.46	
RLOO	96.44	72.40	79.27	82.54	82.67	
リマックス	96.59	75.40	78.66	82.54	84.05	

表2: OODベンチマークにおける異なる利点推定方法の比較。より良い
各報酬モデルの結果は太字で強調表示されます。

4.2 長文思考連鎖課題におけるパフォーマンス

4.2.1 小規模データセットの分析

実験のセットアップQwen2.5-Math-7B事前学習済みモデルを30問のみで学習した。
AIME-24の回答とAIME-25データセットにおけるパフォーマンスを評価した。
限られた訓練設定では、GRPOは小さな訓練を過剰に学習させることでその弱点を顕著に示しました。
データセット。

実験結果表3に示すように、GRPOはほぼ完璧なスコア（約
トレーニングデータセット（AIME-24)では、100点（100点)を達成しました。しかし、テストデータセット（AIME- 25)では、Pass@1
とPass@16の両方のテスト設定でほぼ0点という低いスコアに終わりました。対照的に、REINFORCE++は

パス@N	AIME-24 N = 1	AIME-24 N = 16	AIME-25
GRPO	95.0	0.0	0.4
強化++	71.0	2.5	40.0

表3: 訓練データセット（AIME -24)とテストデータセット（AIME-25)におけるGRPOとREINFORCE++の比較。各報酬モデルで優れた結果は太字で強調表示されています。

AIME-24 では 71.0 という控えめなスコアを達成していますが、 Pass@1 と Pass@16 のテスト設定ではそれぞれ 2.5 と 40.0 というスコアを達成し、より優れた一般化を示しています。

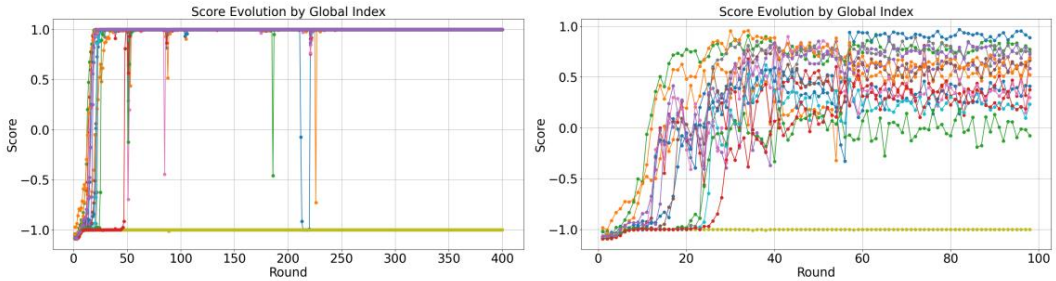

図3: 小規模なプロンプトデータセットを用いたゼロからの強化学習中にランダムに選択された15の質問に対するトレーニング曲線。左: GRPOを用いてトレーニングされたスコア。右: REINFORCE++を用いてトレーニングされたスコア

結果分析 ：さらに、ランダムに選択した15のケースのトレーニング曲線を分析しました。その結果、GRPOは数ステップで急速に100%の精度（図3では1.0のスコアで表されています)を達成することが明らかになりました。対照的に、REINFORCE++はより緩やかな改善を示し、通常10〜20ステップで同レベルの精度に達します。これらの観察結果は、GRPOがトレーニングセットに過剰適合していることを示唆しています。さらに分析すると、GRPOでトレーニングしたモデルからの応答は平均30トークンであるのに対し、REINFORCE++でトレーニングしたモデルは425トークンであり、大幅に短いことが明らかになりました。この証拠は、GRPOがREINFORCE++よりもトレーニングプロンプトの過剰適合の影響を受けやすく、テストデータセットにおける汎化性能が低下することを裏付けています。

4.2.2 教師あり微調整モデルからの強化学習

実験設定現実世界のタスクでは、ユーザーからのプロンプトは非常に多様な形式と意図を示すため、特定のモデルの長所と短所を体系的に評価することは困難です。合成データセットを使用することで、長さや難易度といった重要な要素を体系的に操作し、モデルのパフォーマンスをより直接的かつ解釈しやすい形で評価•分析することが可能になります。

データセットとハイパーパラメータLogic-RL (Xie et al., 2025) に倣い、 Knights and Knaves (K&K) パズル(Xie et al., 2024)を、論理的推論のためのアルゴリズム生成データセットとして RL トレーニングに組み込みます。これらのパズルでは、各キャラクターは常に真実を語る騎士か、常に嘘をつく悪党のいずれかです。目的は、各キャラクターの発言に基づいて、その正体を特定することです。このデータセットの重要な特徴は、その高い制御性です。プロンプトの長さは役割の数に比例し、論理演算の複雑さを変更することで難易度を調整できます。これらのパズルは、元のモデルにとって未知のデータとして機能するため、一般化能力の評価に最適です。論理的推論におけるモデルのパフォーマンスは、指示に従い、コンテキストを理解する能力に依存するため、ベースモデルを使用して実験を開始しませんでした。代わりに、より拡張されたコンテキストを処理し、制御指示に従う能力が強化されたモデルQwen2.5-7B-Instruct-1Mを選択しました。すべてのハイパーパラメータは同一のまま、優位性推定関数を変化させることで強化学習アルゴリズムを比較しました。

実験結果図4は、テストデータセットにおけるGRPOとREINFORCE++の比較分析を示しています。タスクの難易度を表す指標である人数を増やすと、

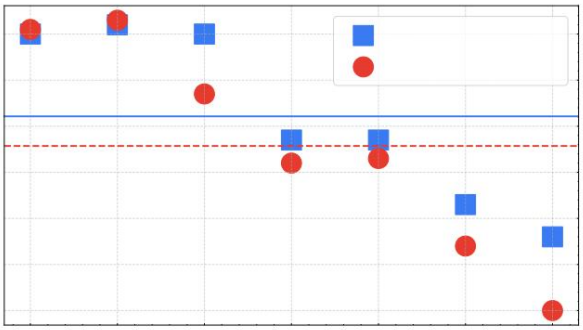

図4: 難易度の異なるロジックベンチマークにおけるGRPOとREINFORCE++の比較。線は2つの手法の平均パフォーマンスを示しています。

両手法の性能を比較すると、GRPOは2〜3人といった比較的単純なシナリオではわずかに高いスコアを記録していますが、人数が増えると性能が大幅に低下します。一方、REINFORCE++は4人以上のシナリオにおいて、より優れた性能安定性を示し、より良い結果を達成しています。この傾向は、訓練データには存在しない8人規模のOut-of-Distributionシナリオにおいて特に顕著で、 GRPOのスコアが20であるのに対し、REINFORCE++は36となっています。全体として、REINFORCE++は平均スコア62.1を記録し、平均スコア55.7を上回っています。これらの結果は、REINFORCE++が複雑でOODのシナリオにおいてより効果的に一般化することを示しており、 GRPOと比較して優れた堅牢性があることを裏付けています。

結果分析図5の訓練曲線を分析した。GRPOは数百ステップで大きな報酬値を達成することが分かった。一方、REINFORCE++は報酬の増加がより緩やかで緩やかであり、最終的にはより高い安定値に収束する。

応答の長さに関して、GRPOで学習したモデルは平均約600トークンとかなり短い応答を生成するのに対し、REINFORCE++は常に約1000トークンと長い応答を生成する。この差は、GRPOが短く、表面的な応答を生成する可能性があることを示唆しており、意味のある推論ではなく、過学習または記憶によるものである可能性を示唆している。したがって、GRPOで学習したモデルは、未知のテストケースに対しては汎化が不十分である可能性があり、過学習の脆弱性に関する以前の観察を裏付けるものとなる。

4.2.3 ゼロからのRL設定

実験設定DeepSeek-R1 (Guo et al., 2025)に着想を得て、検証報酬付き強化学習 (RLVR) を数学タスクのベースモデルに適用し、モデルの推論能力に対する様々な戦略を評価します。先行研究では、強力な推論能力を持つ強力なベースモデルから始めることが示唆されているため、ベースモデルとしてQwen2.5-Math-Base4を選択しました。

データセットこのデータセットは、主に数学トレーニングスプリットの難易度レベル3から5で構成されています。ベースモデルのコンテキスト長の制限により、モデルの出力長の変動を最大限観察できるよう、データセットから約8,000個の短いプロンプトを選択しました。

ハイパーパラメータすべてのハイパーパラメータを同一に保ち、優位性推定関数を変化させることで強化学習アルゴリズムを比較します。各探索ステップでは、計算効率と安定性のバランスを維持するために、32個の質問を選択し、質問ごとに8個の回答を生成します。アクター学習率は、ポリシー更新の収束を最適化するために5 × 10−7です。長期的な報酬の蓄積を強調するために、割引係数（γ)は1.0です。クリッピングパラメータ（)は、トレーニングを安定させ、ポリシー更新を調整するために0.2に設定されています。KLペナルティ係数（β)は、参照モデルからの逸脱を制御するために0.001に設定されています。

4https://huggingface.co/Qwen/Qwen2.5-Math-7B

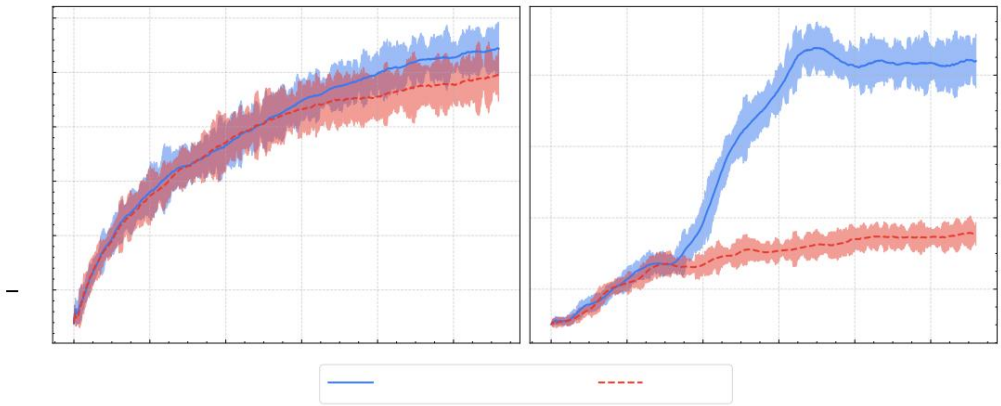

図5: 平滑化された訓練報酬と
教師あり微調整モデルからの応答の長さ。x軸はトレーニングステップを表します。

パス@N	AIME-24 AMC-23 MATH-500 N = 8 N = 8		N = 1
GRPO 18.96 強化++ 21.04		59.22	73.00
		60.47	72.00

表4: 訓練データセット（AIME-24)とテストデータセット（AIME-25)におけるGRPOとREINFORCE++の比較。各報酬モデルで優れた結果は太字で強調表示されています。

実験結果と分析表4に示すように、REINFORCE++は一貫して
GRPOは分布外設定で優れており、他の設定でも同様の結果が得られました。具体的には、
分布内テストデータセット（MATH-500)では、GRPOはわずかに高いPass@1スコアを達成しています。
73.00に対してREINFORCE++は72.00です。しかし、REINFORCE++は優れた性能を示しています。
OODスコアの一般化、AIME-24でGRPOを上回り、Pass@8スコアは21.04
18.96に対して、AIME-25では60.47に対して59.22であった。これらの結果は、
GRPOは訓練分布内では競争力を発揮するが、過剰適合の兆候を示している。
一方、REINFORCE++ は、新しい、特に難しい OOD シナリオに対してより適切に一般化します。

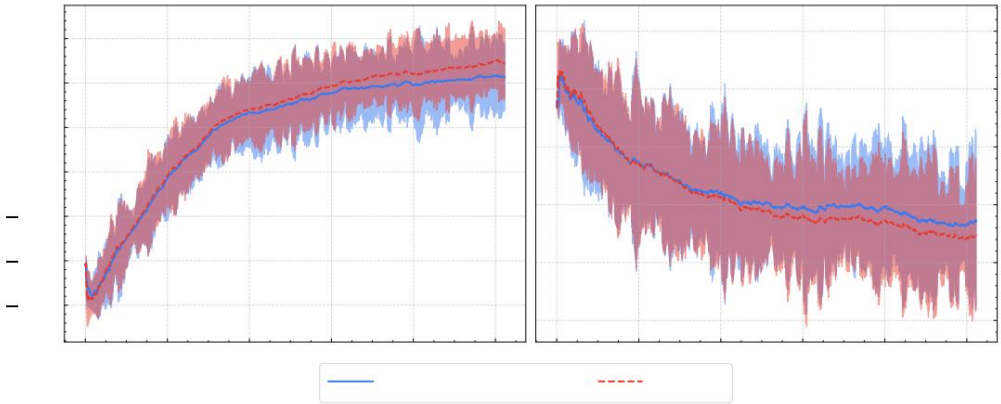

図6: 平滑化された訓練報酬とGRPOとREINFORCE++の比較
ルール報酬モデルによる応答の長さ。X軸はトレーニングステップを表します。

5 結論

本論文では、大規模言語モデル（LLM）を人間の嗜好に効率的に適合させることを目的とした、批判的思考を必要としない新しいRLHFアプローチであるREINFORCE++を紹介しました。プロンプト固有のベースラインを使用する従来の手法とは異なり、REINFORCE++は、過学習と不安定性を防ぐため、グローバルバッチ平均報酬をベースラインとして採用しています。広範な実験により、本アルゴリズムは、ブラッドリー•テリー法、ルールベース報酬モデル、長文思考連鎖（CoT）タスクなど、複数のRLHFシナリオにおいて優れたパフォーマンスと計算効率の向上を達成し、優れた汎化能力を示すことが実証されました。今後の研究課題としては、適応型正規化手法、高度な分散低減手法の検討、そしてRLHF設定を超えたREINFORCE++の拡張が挙げられます。

参考文献

ジョシュ•アチアム、スティーヴン•アドラー、サンディニ•アガルワル、ラマ•アフマド、イルゲ•アッカヤ、フロレンシア•レオニ•アレマン、ディオゴ•アルメイダ、ヤンコ•アルテンシュミット、サム•アルトマン、シャマル•アナドカット 他GPT-4技術レポート。 arXiv プレプリント arXiv:2303.08774、2023。

AI Anthropic。2023年、クロード登場。URL https://www.anthropic.com/news/ クロードの紹介。

Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、 Stanislav Fort、Deep Ganguli、Tom Henighan他。人間からのフィードバックからの強化学習を用いた、有用かつ無害なアシスタントのトレーニング。 arXivプレプリント arXiv:2204.05862、2022年。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、 Shiron Ma、Peiyi Wang、Xiao Bi 他。 Deepseek-r1:強化学習を通じて llms の推論能力を奨励します。 arXiv プレプリント arXiv:2501.12948、2025。

Tianle Li、Wei-Lin Chiang、Evan Frick、Lisa Dunlap、Tianhao Wu、Banghua Zhu、Joseph E Gonzalez、 Ion Stoica。クラウドソーシングデータから高品質ベンチマークへ :Arena-hardとベンチビルダーパイプライン。arXivプレプリント arXiv:2406.11939, 2024.

Ziniu Li、Tian Xu、Yushun Zhang、Yang Yu、Ruoyu Sun、Zhi-Quan Luo。 Remax:大規模な言語モデルを調整するためのシンプルで効果的かつ効率的な方法。 arXiv プレプリント arXiv:2310.10505、2023 。

劉一明。 RLHF における KL ダイバージェンスの再考: 単一サンプルからミニバッチ、そして期待値へ、2025 年。 URL https://www.notion.so/ Rethinking-KL-Divergence-in-RLHF-From-Single-Sample-to-Mini-Batch-to-Expectation- Notionブログ。

OpenAI. Gpt-4技術レポート. arXivプレプリント arXiv:2303.08774, 2023.

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray他「人間のフィードバックによる指示に従う言語モデルのトレーニング」arXivプレプリント arXiv:2203.02155, 2022a.

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray他.人間のフィードバックによる指示に従う言語モデルの学習. Advances in Neural Information Processing Systems, 35: 27730–27744, 2022b.

Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon、Chelsea Finn. 「直接的な選好最適化：言語モデルは実は報酬モデルである」。第37回神経情報処理システム会議、2023年。URL : https://arxiv.org/abs/2305.18290 。

ジョン•シュルマン、フィリップ•ウォルスキ、プラフラ•ダリワル、アレック•ラドフォード、オレグ•クリモフ。近接政策最適化アルゴリズム。arXivプレプリントarXiv:1707.06347、2017年。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Mingchuan Zhang、YK Li、Yu Wu、 Daya Guo。 Deepseekmath: オープン言語モデルにおける数学的推論の限界を押し広げます。 arXiv プレプリント arXiv:2402.03300、2024。

Wei ShenとChuheng Zhang. RLHFにおけるポリシーフィルタリングによるLLMのコード生成における微調整. arXiv プレプリントarXiv:2409.06957、2024年。

Geminiチーム、Rohan Anil、Sebastian Borgeaud、Yonghui Wu、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth他「Gemini :高度な機能を備えたマルチモーダルモデルファミリー」arXivプレプリント arXiv:2312.11805、2023年。

サイ•ヴェンプララ、ロジェリオ•ボナッティ、アーサー•バッカー、アシシュ•カプール。ロボット工学のためのチャット: デザイン原理とモデル能力。Microsoft Auton. Syst. Robot. Res, 2:20, 2023.

Yuan Wu et al. 基本に立ち返る: 人間からの学習のための強化スタイルの最適化の再検討
llmsにおけるフィードバック。arXivプレプリントarXiv:2402.14740, 2024。

Chulin Xie、Yangsibo Huang、Chiyuan Zhang、Da Yu、Xinyun Chen、Bill Yuchen Lin、Bo Li、Badih Ghazi、Ravi Kumar。論理的推論における大規模な言語モデルの記憶について。 arXivプレプリント arXiv:2410.23123、2024。

Tian Xie、Zitian Gao、Qingnan Ren、Haoming Luo、Yuqian Hon、Bryan Dai、Joey Zhou、Kai Qiu、 Zhirong Wu、Chong Luo。 Logic-rl: ルールベースの強化学習で llm 推論を解き放ちます。 arXiv プレプリント arXiv:2502.14768、2025。

アルゴリズムの詳細

REINFORCE++-ベースライン。使いやすさと再現性を高めるため、提案手法をOpenRLHFフレームワーク（https://github.com/OpenRLHF/OpenRLHF)内に実装しました。

上記の実験では、プロンプトごとに複数の応答を生成することでモデルを訓練することで、パフォーマンスをさらに向上できることが示されました。そこで、 REINFORCE++と複数応答生成を統合したREINFORCE++- Baselineというバリアントを導入します。具体的には、各プロンプトに対して複数の応答をサンプリングし、それらの平均報酬をベースラインとして計算することで報酬を再形成します。各プロンプトと応答のペアにおけるアドバンテージは、以下のように定義されます。

$$\text{Adv}_{q,ot} = R_{q,ot} - \text{平均グループ}(R_{q,ot})$$

$$\text{アドバンノルム}_{q,ot} = \frac{\text{Adv}_{q,ot} - \text{meansbatch}(\text{Adv}_{q,ot})}{\text{stdbatch}(\text{Adv}_{q,ot})}$$

ここで、 group は同じプロンプトに対応する生成された応答を表します。このベースライン計算はGRPOのアプローチと似ていますが、グループ正規化の標準偏差(std)を意図的にグローバルバッチ正規化に移動しています。実際には、標準偏差を含めるとトレーニングの不安定性が生じ、 RLHFプロセスにおいてより簡単なプロンプトへの過剰適合が悪化することがわかりました。KL制約については、REINFORCE++-baselineはGRPOのk3ダイバージェンス近似法ではなくk2ダイバージェンス近似法を使用します。詳細については、次の段落を参照してください。

KLペナルティ設計。REINFORCEベースのRLHF法の設計において、KLダイバージェンス推定の選択はアルゴリズムの安定性と精度に大きな影響を与えます。GRPOで広く用いられているk3推定は、損失関数によって定義されます。

$$L_{k3} = \mathbb{E}_{x \sim D,\, y \sim \pi_{\theta_{old}}(\cdot|x)} \left[\frac{\pi_{ref}(a_t|s_t)}{\pi_\theta(a_t|s_t)} - \log \frac{\pi_{ref}(a_t|s_t)}{\pi_\theta(a_t|s_t)} - 1 \right] \tag{12}$$

これらは近似的な性質のため、固有のバイアスを伴います。具体的には、 k3推定値の勾配は線形化によってk2推定値の勾配を近似します。

$$\nabla_\theta \left[\frac{1}{2} (\log x)^2 \right] \quad \text{k2勾配:} \tag{13}$$

$$\nabla_\theta \left[(x - 1) - \log x \right] \quad \text{k3勾配:} \tag{14}$$

ここで$x = \frac{\pi_{ref}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ 。k3勾配はk2勾配とほぼ一致するが、ポリシー基準値（$x \approx 1$)に近いですが、大幅な逸脱により2つの重大な欠陥が発生します。

1. バイアス:現在のポリシーが基準から大きく逸脱すると、特に線形近似誤差が非線形に拡大して、 $\pi_{ref} \gg \pi_{\theta_{old}}$に大きなバイアスが生じます。

2. 非対称性 :関数（$x - 1$)は、 $\pi_{\theta_{old}}$がπ_{ref}を上回るか下回るかによって非対称な挙動を示し、収束特性をさらに複雑にします。

したがって、計算が単純であるにもかかわらず、 k3推定はバイアスと分散の増加を招くため、理論的に不偏なk2推定よりも厳密に優れているわけではないことが示唆されます。実験的証拠は、 GRPOにおいてk3推定を用いるとk2推定よりも分散の変動が大きくなることを一貫して示しています。詳細な議論はLiu (2025)に記載されています。

PPO との関係。REINFORCE ++ は、その処方において PPO といくつかの類似点を持っています。

具体的には、PPOが$\lambda = 1$ 、割引係数$\gamma = 1$の一般化利点推定（GAE)パラメータを採用する場合、REINFORCE++は批評ネットワークを使用しないPPOに帰着し、ベースラインとしてグローバルバッチ正規化を採用します。数学的には、この関係は次のように表されます。

$$\text{GAE}(\lambda = 1,\ \gamma = 1) = r_{t+l} - V(s_t), \quad \sum_{l=0}^{\infty} \tag{15}$$

ここで、 V(st)は状態stにおける推定値関数を表します。REINFORCE ++アルゴリズムは、批判的ネットワークを除去することで、 V(st)項を効果的に除去します。同時に、グローバルバッチ正規化の導入により、分散が低減され、個々のプロンプトへの過剰適合が防止されるため、学習がさらに安定化します。

その他の実装の詳細

ミニバッチ更新トレーニング効率を高めるために、次の特性を持つミニバッチ更新を実装します。

- バッチ処理:データは、フルバッチではなく、管理しやすい小さなチャンクで処理されます。アップデート。
- 複数の更新:各ミニバッチで複数のパラメータ更新が可能になり、収束率。
- 確率的最適化:より優れた一般化のために有益なランダム性を導入します。

報酬の正規化とクリッピングトレーニングを安定させるために包括的な報酬処理を実装します。

- 正規化:外れ値を軽減するために、Z スコア正規化を使用して報酬を標準化します。
- クリッピング:不安定性を回避するために、報酬値を事前定義された範囲内に制限します。
- スケーリング:更新中の数値の安定性を確保するために、適切なスケーリング係数を適用します。

B 謝辞

- Jian Hu: REINFORCE++ と REINFORCE++-baseline の核となるアイデアを開発し、アルゴリズムを実装したほか、論文の初期ドラフトにも貢献しました。
- Jason Klein Liu: 実験コードの実装、ハイパーパラメータの微調整、論文執筆、GPUリソースの提供を行いました。
- Wei Shen: 論文執筆を主導し、主要な実験の設計と監督を行いました。