

R1-Zero のようなトレーニングを理解する: 批判的な視点

Zichen Liu* † 1,2 Changyu Chen* 1,3 Wenjun Li* 3 Penhui Qi* 1,2

パン・ティエン1 チャオ・ドウ1 ウィー・サン・リー2ミン・リン1

1Sea AI Lab 2シンガポール国立大学3シンガポール経営大学

2025年3月21日

抽象的な

DeepSeek-R1-Zero は、大規模な強化学習 (RL) が教師ありの微調整なしで LLM の推論機能を直接強化できることを示しました。この研究では、R1-Zero のようなトレーニングを、その 2 つのコア コンポーネントである基本モデルと RL を分析することにより、批判的に調べます。DeepSeek-V3-Base を含む幅広い基本モデルを調査し、事前トレーニングの特性が RL パフォーマンスにどのように影響するかを理解します。分析により、DeepSeek-V3-Base がすでに「アハ モーメント」を示しているのに対し、Qwen2.5 基本モデルはプロンプト テンプレートがなくても強力な推論機能を示し、潜在的な事前トレーニング バイアスを示唆していることが明らかになりました。さらに、Group Relative Policy Optimization (GRPO) の最適化バイアスを特定しました。これは、トレーニング中に応答の長さ (特に誤った出力の場合) を人為的に増加させます。これを解決するために、推論パフォーマンスを維持しながらトークン効率を向上させる偏りのない最適化手法である Dr. GRPO を紹介します。これらの洞察を活用して、7B ベース モデルで AIME 2024 で 43.3% の精度を達成し、新たな最先端を確立する最小限の R1-Zero レシピを提示します。

R1-Zero のようなトレーニングを理解する: 批判的な視点 <https://github.com/sail-sg/理解-r1-zero1>

R1-Zero のようなトレーニングを理解する: 批判的な視点

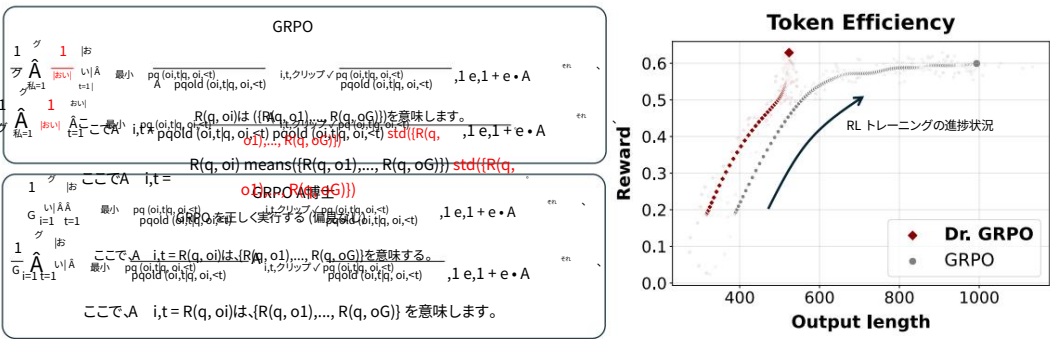


図 1: 左: GRPO 博士は、長さ標準偏差の正規化項を削除することで、GRPO (Shao et al., 2024) のバイアスに対処するためのシンプルでありながら重要な変更を導入しています。右: 当社の偏りのないオプティマイザーは、モデルが次第に長い誤った応答を生成するのを効果的に防止し、トークンの効率を高めます。

コア貢献者。

† プロジェクトリーダー。

1 LLM RL フレームワーク Oat を使用して開発されました: <https://github.com/sail-sg/oat>。

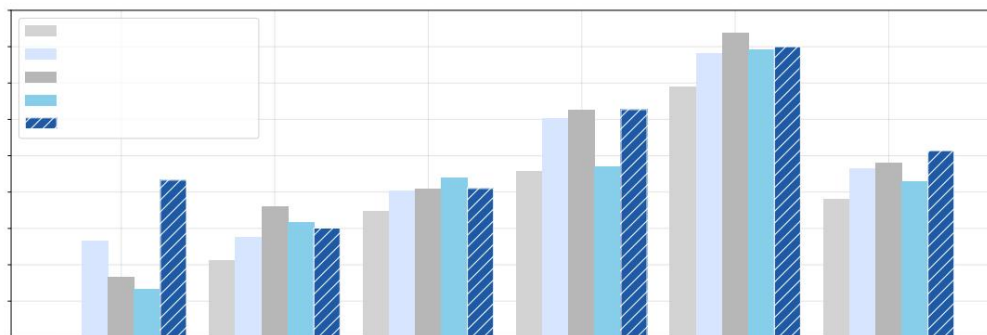


図 2: モデルのパフォーマンス比較。Oat -Zero-7Bは、セクション1 (3 番目の段落) で説明した最小限のレシピを使用して RL 調整されています。詳細な結果については、付録Bを参照してください。

1 はじめに

DeepSeek-R1-Zero (Guo et al., 2025) は、R1-Zero のようなトレーニング パラダイムを導入することで、大規模言語モデル (LLM) のトレーニング後のパイプラインに革命をもたらします。つまり、予備ステップとして教師あり微調整 (SFT) に依存せずに、RL をベース LLM に直接適用します。この新しいパラダイムは、そのシンプルさと実証されたRL スケーリング現象により魅力的です。モデルの推論機能は、モデルの応答長の継続的な増加とともに向上します。この現象には、モデルが自己反省などの新たなスキルを学習する「アハ モーメント」も伴います。

本稿では、基本モデルと強化学習という2つの重要な要素を研究することで、R1-Zeroのようなトレーニングを理解することを目指します。第1部では、基本モデルのさまざまな属性を調査します。特に、R1-Zeroを再現する最近の試みで使用されているQwen2.5モデルファミリー (Yang et al., 2024a;b) (Pan et al., 2025; Zeng et al., 2025; Liu et al., 2025b; Hu et al., 2025)と、実際のR1-Zeroモデルが強化学習で調整されたDeepSeek-V3-Base (Liu et al., 2024)に焦点を当てます。第2部では、GRPO (Shao et al., 2024)の最適化におけるバイアスを特定します。このバイアスは、次第に長い誤った応答につながる可能性があります。この目的のために、我々はバイアスを排除するための簡単な変更、すなわち、GRPO Done Right (Dr. GRPO) を提案します。これにより、トークン効率が向上します(図1 で強調表示)。

ベースモデルとRLに関する私たちの分析は、R1-Zeroのようなトレーニングのための最小限のレシピを示唆しています。Qwen-Mathテンプレートを使用して、MATH (Hendrycks et al., 2021)レベル3~5の質問に対して (偏りのない)Dr. GRPOアルゴリズムを使用してQwen2.5-Math-7BをRLチューニングし、8× A100 GPUでわずか27時間の計算で最先端のパフォーマンス (図2)を達成しました。この論文で提示した調査結果、公開されたモデル、オープンソース化されたコードベースが、この分野の将来の研究に役立つことを願っています。

概要として、この論文の要点を以下にまとめます。

- (セクション2.1)テンプレートは、ベース モデルが文を完成させるのではなく質問に答えるために重要です。さらに、すべてのベース モデルは、RL の前にすでに数学を解く能力を備えています。
- (Sec. 2.2)興味深いことに、Qwen-2.5 ベースモデルはテンプレートを使用しないことで即座に約60%の改善が得られ、モデルをクックする際に連結された質問と回答のテキストで事前トレーニングできるのではないかと推測されます。
- (Sec. 2.3) DeepSeek-V3-Baseを含め、ほぼすべてのベースモデルですでに「Aha moment」が見られます。
- (Sec. 3.1、Sec. 3.2) Dr. GRPO は最適化におけるGRPO のバイアスを効果的に修正し、トークン効率を向上させました。
- (Sec. 3.3)モデルとテンプレートの不一致により、RLの前に推論機能が破壊される可能性があります。それを再構築します。
- (Sec. 3.4) Llama-3.2-3B での数学の事前トレーニングにより、RL 上限が向上します。

2 基本モデルの分析このセクションでは、

Qwen-2.5ファミリー (Yang et al., 2024a;b)、Llama-3.1 (Grattafiori et al., 2024)、DeepSeekシリーズ (Liu et al., 2024; Shao et al., 2024; Guo et al., 2025)を含む幅広い基本モデルを精査し、MATH (Hendrycks et al., 2021)トレーニングセットからサンプリングした500の質問をして、その応答を分析します。

2.1 R1-Zero の訓練可能性: テンプレートは探索的基本ポリシーを構築します。基本モデルか

らの訓練は R1-Zero のようなパラダイムの基本的な設定であるため、まず、一般的に文の完成 (つまり、 $p_\theta(x)$) のために訓練される、広く使用されているオープンソースの基本モデルが、適切なテンプレートを通じて質問応答機能を効果的に引き出し、質問応答の基本ポリシー $\pi_\theta(\cdot|q)$ として機能するかどうかを調査します。Guo ら(2025) の R1 テンプレート (テンプレート 1)に加えて、Zeng ら(2025)が使用したQwen-Math テンプレート (テンプレート2)と、テンプレートなし (テンプレート3) を検討します。

テンプレート 1 (R1 テンプレート)。ユーザーとアシスタント間の会話。ユーザーが質問をし、アシスタントがそれを解決します。アシスタントはまず頭の中で推論プロセスを考え、次にユーザーに答えを提供します。推論プロセスは<think> </think>タグで囲み、答えは<answer> </answer>タグで囲みます。つまり、<think>推論プロセスはここに</think> <answer>答えはここに</answer> です。\\nユーザー: {question}\\nアシスタント: <think>

テンプレート 2 (Qwen-Math テンプレート)。<\\im start>system\\nステップごとに推論し、最終的な答えを \\boxed{} 内に入力してください。<\\im end>\\n<\\im start>user\\n{question}<\\im end>\\n<\\im start>assistant\\n

テンプレート 3 (テンプレートなし)。{question}

実験設定。実験には、Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.5-7B、Llama-3.1-8B、DeepSeek-Math-7B、DeepSeek-V3-Base-685Bを含めます。各モデルについて、最初にテンプレートなしを適用してモデル応答を取得し、次にGPT-4o-miniにモデル応答が回答形式 (品質に関係なく)であるか、文完成パターンであるかを判断させます。質問に答える傾向がある応答の割合をメトリックとして記録します。次に、R1テンプレートとQwen-Mathテンプレートの両方を適用してモデル応答を取得し、メトリックに基づいて各モデルに最適なテンプレートを決定します。最後に、対応するテンプレートを使用して各モデルのpass@8精度を評価し、基本ポリシーがRL改善のための報酬軌道を探索できるかどうかを評価します。

結果。図3の左のプロットは、基本モデル (テンプレートあり、なし) が、提供された質問にどの程度適切に回答するかを示しています。Llama モデルと DeepSeek モデルはすべて、適切なテンプレート (R1 テンプレート) を使用することで回答能力が向上していることがわかります。ただし、Qwen2.5モデルは、テンプレートを使用しない場合に最もよく機能します (回答率 100%)。この興味深い特性は、セクション2.2 で説明するように、さらに調査を進める動機となります。一方、テンプレートなしでの回答率が最も低いことは、DeepSeek-V3-Base がほぼ純粋な基本モデルであることを示唆しています。この観察結果から、DeepSeek-V3-Base のような純粋な基本モデルがAha モーメントを示すかどうかを調べる動機となります (セクション2.3)。図3の中央のプロットは、異なるサンプリング温度での異なる基本モデル (テンプレートあり) のpass@8精度を示しています。このメトリックは、基本ポリシーの探索能力の指標として使用できます。たとえば、ベースポリシーが正しい最終回答につながる単一の軌跡をサンプリングすることさえできない場合、報酬信号がないため、RL でポリシーを改善することは不可能です。私たちの結果は、テストされたすべてのモデルが探索的 (したがって RL の準備ができています) であり、Qwen2.5 モデルが最高のパフォーマンスを発揮する (DeepSeek-V3-Base を上回る) ことを示しています。これは、ほとんどの R1-Zero プロジェクト (Zeng ら、2025 年、Hu ら、2025 年)が Qwen2.5 モデルに基づいていることを部分的に説明している可能性があります。

2.2 Qwen-2.5モデルはテンプレートを破棄するときに最高のパフォーマンスを発揮します

次に、Qwen2.5 ベースのモデルはすべて、テンプレートがなくてもチャット モデルとして使用できるという興味深い観察結果 (図 3(左) を参照) を詳しく調べます。さらに一歩進んで、

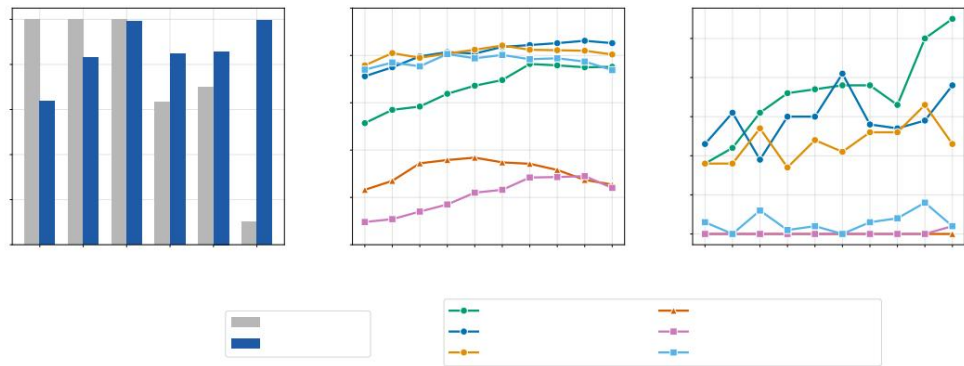


図3: 3つの側面にわたるモデル属性。質問応答能力: 事前学習された言語モデルが質問に直接答えるのではなく、継続または拡張;探索能力: pass@8は、ベースがどれだけ優れているかを測定します モデル探索;自己反省:カウントは、付録Cに詳述されているように、キーワードベースの検出と LLM ベースの検出があります。

基本モデル + テンプレートAIME24 AMC MATH500 Minerva OlympiadBench平均						
Qwen2.5-数学-1.5B						
(4ショットプロンプト)	0.0	20.0	50.4	12.1	15.9	19.7
R1テンプレート	0.0	9.6	21.2	6.6	2.2	7.9
Qwenテンプレート テ	20.0	32.5	33.0	12.5	22.8	24.2
ンプレートなし	16.7	43.4	61.8	15.1	28.4	33.1
Qwen2.5-数学-7B						
(4ショットプロンプト)	3.3	22.5	61.6	10.7	20.9	23.8
R1 テンプレート	0.0	0.0	0.0	0.0	0.1	0.0
Qwen テンプレート	16.7	38.6	50.6	9.9	16.6	26.5
テンプレートなし	0.2	45.8	69.0	21.3	34.7	38.2

表1: Qwen2.5-Mathモデルは、連結された質問と回答のテキストで事前学習される可能性がある。テンプレートが適用されていない場合でも、最高のパフォーマンスが得られます。

Qwen2.5-Mathモデルの推論能力を5つの標準ベンチマークで評価する :AIME 2024 (Li et al., 2024)、AMC (Li et al., 2024)、MATH500 (Hendrycks et al., 2021)、Minerva 数学 (Lewkowycz et al., 2022) 、およびOlympiadBench (He et al., 2024) 。 実際には、貪欲なデコードを使用し、サンプリング バジレットを 3000 トークンに制限します。

表1に示すように、テンプレートを使用しないと平均パフォーマンスが大幅に向上します。従来の 4 ショットプロンプトと比較して約 60% の改善が実現しました。 Qwen2.5-Math (Yang et al., 2024b)はチャットモデルのデータ (質問と回答のペア)を使用するため 事前訓練の段階では、連結された テキストはlog pθ (q; o)を直接最大化します。仮説が正しいと判明した場合、我々はより Qwen2.5モデルを使用してDeepSeek-R1-Zeroを再現する際には注意が必要です。ベースモデルは テンプレートなしでもすでに SFT に似ています。

2.3 Aha Moment は、DeepSeek-V3-Base を含むベース モデルにすでに登場しています

DeepSeek-R1-Zeroの最も感動的な成果の一つは、自己反省の出現である。 行動、つまり「アハ体験」を、純粋なRLトレーニングを通じて学ぶ。いくつかの先行研究 (Liu et al., 2025b; Yeo et al., 2025)は、オープンソースには「アハ」の瞬間がないかもしれないと示唆している。 R1 レプリケーションは、使用する基本モデルにすでに自己反映キーワードが含まれているためです。 しかし、彼らは実際のR1-Zeroモデルが使用されたDeepSeek-V3-Baseをテストしていない。 RL調整済み。DeepSeek-V3-Base-685Bを自社でホストすることで、この欠けている部分を補うことができます。

要約し,R1 テ

ンプレートを使用して 500 の MATH 質問に対する応答を調査します。図 3 の右側のプロットから、DeepSeek-V3-Base もかなりの量の自己反映を生成し、Liu ら(2025b) の主張をさらに検証していることがわかります。また、図4には、DeepSeek-V3-Base が「なるほど」、「待つ」、「COLM 2025 問題を確認する」などのキーワードを生成する例も示しています。

質問(a): $1/\sin 45 \sin 46 + 1/\sin 47 \sin 48 + \cdots + 1/\sin 133 \sin 134 = 1/\sin n$ となる最小の正の整数nを求めます。	
回答まず、方 程式の左側が $\sum_{k=1}^n \frac{1}{\sin a_k \sin a_{k+1}}$ の形式であり、 $a_1 = 45$ 、 $a_2 = 47$ 、...、 $a_{90} = 133$ であることに気が付きました。また、 $\sin a_{k+1} = \sin(a_k + 2)$ であることに気が付きました。 ... 単純化すると、 $\sin(3a) = 3 \sin a \cos^2 a - \sin^3 a$ となります。 なるほど! これを使うと $\sin(a + 2b) = \sin((a + b) + b) = 3 \sin(a + b) \cos^2 b - \sin^3(a + b)$ が得られます... 見た目は乱雑ですが、再び $\sin a \cos b + \cos a \sin b$ という形式になっていることに気が付きました。 これによって、あるアイデアが浮かびました。おそらく、この恒等式を繰り返し使用して、方程式の左側を簡略化できるでしょう。 ...	
質問(b):最初の15行の中から要素がランダムに選択されます...	質問(c):デナリとネイトは犬の散歩ビジネスで働いており、給料は...
回答パスカルの三角形では、すべての行は1で始まり、1で終わります。 ... これは次のように計算できます: *気まずい沈黙* ちょっと待ってください、考えすぎです。もう一度やり直しましょう。 パスカルの三角形の最初の n 行の要素の数は $(n+1)!/[2! (n-1)!]$ です。 ...	回答この問題を解決するには、まず変数を定義することから始めましょう... ... 繰り返しますが、 $x^6 = 0$ なので、この方程式も有効な解を与えません。 間違いがないか確認してみましょう。問題の条件を再度確認する必要があります。正しいアプローチは... ...

図 4: DeepSeek-V3-Base がRL チューニング前の5 回の参照でもすでに「アハ体験」を示していることを示す事例。

6 付録

さらに重要な疑問は、自己反省行動がRL トレーニング後のモデル パフォーマンスの向上と関連しているかどうかです。これを調査するために、DeepSeek- 7をホストしています。ここに他の追加セクションを含めることができます。
R1-Zero と MATH データセットからの同じ質問に対する回答を分析します。R1-Zero では自己反省行動がより頻繁に発生しますが、これらの行動が必ずしも高い精度を意味するわけではないことがわかります。詳細な分析は付録Dにあります。

3 強化学習の分析

言語モデルの生成は、トークンレベルのマルコフ決定過程(MDP) $M = (S, A, r, p_Q)$ として定式化できます。各生成ステップ t で、状態 $st \in S$ は、入力質問とこれまでに生成された出力応答の連結です: $st = q; o \leq t = [q_1, \dots, q_M, o_1, \dots, o_{t-1}]$ 。ポリシー $\pi_\theta(\cdot | st)$ は語彙 A から次のトークン ot を選択し、次の状態 $st+1 = st; [ot]$ への決定論的遷移をもたらします。生成プロセスは、一連の質問から初期状態 $s_1 = q$ から p_Q をサンプリングすることから始まり、自己帰帰ポリシーが $[eos]$ トークンを生成するか、予算を使い果たすと停止します。

通常、エントロピー正規化目的関数を最大化します(Schulman et al., 2017a)。

$$J(\pi_\theta) = E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot | q)} [R(q, o)] - \beta D_{KL}[\pi_\theta(\cdot | q)] \| \pi_{ref}(\cdot | q)]、$$

(1)

ただし $R(q, o) = \sum_{t=1}^{|o|} r(st, ot)$ は軌道 q ; o のリターン (Sutton & Barto, 2018)であり、 π_{ref} は参照ポリシーです。KL 正則化項は通常、人間のフィードバックからの強化学習に採用されます (Christiano et al., 2017)。ここで r は報酬です。

R1-Zero のようなトレーニングを理解する: 批判的な視点

π_{ref} によって収集されたデータから学習されたモデル。この場合、正則化は、 π_{θ} が報酬モデルが正確な分布から大きく逸脱するのを防ぐのに役立ちます(Jaques et al., 2019; Stiennon et al., 2020)。ただし、RL チューニング推論モデルでは通常、ルールベースの検証器が r として使用される(Lambert et al., 2024) ため、分布シフトの懸念は排除されます。

これにより、KL 項を削除できます。これにより、トレーニング中に π_{ref} に必要なメモリと計算が節約されるだけでなく、R1-Zero のようなトレーニングのパフォーマンスが向上する可能性もあります(Hu et al., 2025)。この論文では、 $\beta = 0$ と仮定します。

ポリシー最適化アルゴリズム。上記の目的関数 ($\beta = 0$ の式 (1))で π_{θ} を最適化するために、近似ポリシー最適化 (PPO) (Schulman et al., 2017b) は次の代理目的関数を最大化します。PPO (π_{θ}) = Eq $p_{Q, \theta} \pi_{\theta}^{\text{old}}(\cdot | q) \pi_{\theta}(\text{ot} | q, o < t) \pi_{\theta}(\text{ot} | q, o < t)$

$$\sum_{t=1}^{|a|} \text{分} \frac{A_t}{\text{clip}(\pi_{\theta}^{\text{old}}(\text{ot} | q, o < t) \pi_{\theta}^{\text{old}}(\text{ot} | q, o < t) \pi_{\theta}(\text{ot} | q, o < t) \pi_{\theta}(\text{ot} | q, o < t))} \cdot 1 - , 1 +) A_t \quad (2)$$

で、 $\pi_{\theta}^{\text{old}}$ は更新前のポリシー、 θ はクリッピングハイパーパラメータ、 A_t は t 番目のトークンのアドバンテージ関数の推定量です。 A_t を推定する標準的な方法は、学習した価値モデル V を使用して一般化アドバンテージ推定 (GAE) (Schulman et al., 2015) を計算することです。ただし、LLM RL チューニングのコンテキストでは、価値モデルの学習は計算コストが高いため、 V を使用せずに A を推定する方法が実際には好まれます。たとえば、Shao et al. (2024) は、まず質問ごとに回答のグループ $\{o_1, \dots, o_G\}$ をサンプリングし、その戻り値 $R = \{R_1, \dots, R_G\}$ を計算し、次に o_i からのすべてのトークンを A として利用する t

$$t = \frac{R_i - \text{平均}(R)}{\text{標準偏差}(R)}$$

3.1 GRPO は偏った最適化につながるDeepseek-R1-Zero (Guo

et al., 2025) では、トレーニング プロセス全体を通じて応答の長さが一貫して増加するという注目すべき傾向が見られます。これは、自己反省などの高度な推論能力の発達の兆候として解釈されることがよくあります。最近の研究(Pan et al., 2025; Zeng et al., 2025; Hu et al., 2025)では、さまざまなアルゴリズムと実装を使用してこの現象を再現しています。ただし、観察された応答の長さの増加は、GRPO (Shao et al., 2024)の目的関数に固有のバイアスにも起因している可能性があると言います。GRPO (π_{θ}) = Eq $p_{Q, \theta} \{o_i\} G$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \text{分} \frac{\pi_{\theta}(o_i, t | q, o_i, < t)}{\text{(} o_i, t | q, o_i, < t \text{) をクリップする}} A_{i, t} \cdot \frac{\pi_{\theta}^{\text{old}}(o_i, t | q, o_i, < t)}{\pi_{\theta}^{\text{old}}(o_i, t | q, o_i, < t)} \cdot 1 - , 1 + A_{i, t} \quad (3)$$

どこ

$$A_{i, t} = \frac{R(q, o_i) - \text{means}(\{R(q, o_1), \dots, R(q, o_G)\})}{\text{std}(\{R(q, o_1), \dots, R(q, o_G)\})}$$

戻り値 $R(q, o_i)$ には通常、LLM推論における結果検証可能な報酬のみが含まれます (この分析はプロセス報酬の場合にも適用されます)。

式(2)の目的関数と比較して、GRPOは2つのバイアスを導入します (図5も参照)。

- 応答レベルの長さバイアス : これは $|o_i|$ で割ることから生じます。正の利点 ($A_{i, t} > 0$, 正しい応答を示す)の場合、このバイアスにより、より短い応答の勾配更新が大きくなり、ポリシーは正しい回答の簡潔さを優先するようになります。逆に、負の利点 ($A_{i, t} < 0$, 不正解の応答を示す)の場合、長い応答は $|o_i|$ が大きいいためペナルティが小さくなり、ポリシーは不正解の中でもより長い応答を優先するようになります。

• 質問レベルの難易度バイアス : これは、中心化された結果報酬を $\text{std}(\{R(q, o_1), \dots, R(q, o_G)\})$ で割ることによって発生します。標準偏差が低い質問 (例 : 簡単すぎるか難しすぎる質問で、結果報酬がほぼすべて1または0である質問)は、ポリシーの更新中に高い重みが与えられます。アドバンテージ正規化はRLで一般的なトリックですが (Andrychowicz et al., 2021) 、通常はバッチ全体で計算されます。

対照的に、質問レベルの正規化では、質問ごとに目的の重みが異なり、最適化の難易度の偏りが生じます。

R1-Zero のようなトレーニングを理解する: 批判的な視点

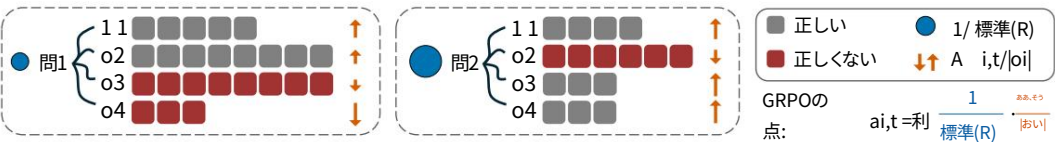


図5: GRPOにおけるバイアスの図解。GRPOの有効な利点は、ai,t

は、不偏優位性Aの再加重バージョンに相当する。
項std(R)と|oi|は、異なる重みを割り当てることで最適化にバイアスをかける可能性がある。
異なる質問と回答は青い円の大きさと長さで示される
オレンジ色の矢印の。

オープンソースのPPO実装にも長さの偏りは存在する。我々はまた、いくつかのLLMのポストトレーニング用の一般的なPPOアルゴリズムのオープンソース実装。驚いたことに、これらの実装はすべて、応答の長さによって損失を正規化しています（リスト1を参照）。および表2は、式（2）で定義されたPPOの目的と一致していません。この定式化と実装の不一致は、GRPOの出版前から存在していました。この不一致は、事前トレーニングの段階（Shoeybi et al., 2019）に起因する可能性がありますと推測しています。
すべてのトークンが固定長のコンテキストにパックされ、損失が正規化される。
コンテキストの長さ（つまり、loss.mean(-1)を計算する）は数値安定性を向上させる。しかし、RLチューニング段階では、典型的な実装（von Werra et al., 2020）では損失を次のように正規化する。
応答の長さは一定ではなく、意図しない長さのバイアスが生じます。

リスト1: 典型的なオープンソースPPO損失実装との比較
バイアス（赤）と実装（緑）。MAX_TOKENSは、
トレーニング全体（予算カリキュラムが有効でない場合）の最大数を指定します
生成トークンの。他の定数も勾配ノルムの差に応じて機能します。

```
1 定義マスク平均（テンソル、マスク、薄暗い）:  
2 戻り値（テンソル * マスク）.sum（axis = dim）/ マスク .sum（axis = dim）  
3 + 戻り値（テンソル * マスク）.sum（軸 = -1）/ MAX_TOKENS  
4  
5 ppo_loss = # トークンごとのppo損失を計算  
6 response_mask = # トークンごとのレスポンスマスク  
7 # レスポンスの長さの正規化（例 :OpenRLHF）  
8 loss_variant1 = masked_mean（ppo_loss response_mask dim = -1）.mean()  
9 # OR バッチごとの長さの正規化（例 :trl verl）  
10 loss_variant2 = masked_mean（ppo_loss response_mask dim = None）.mean()
```

リポジトリ trl	コードリンクは公平ですか？
(von Werra et al., 2020)	PPO損失 ✗
OpenRLHF (Hu et al., 2024) verl	PPO損失 ✗
(Sheng et al., 2024)	PPO損失 ✗
SimpleRL-Zero (Zeng et al., 2025)	PPO損失 ✗
オープンリーゾナーゼロ (Hu et al., 2025)	PPO損失 ✗

表 2: 多くのオープンソース PPO 実装には長さのバイアスがあります。

3.2 Dr. GRPO: グループ相対ポリシー最適化を正しく行う

GRPOにおける前述の最適化バイアスを回避するために、我々は単純に
およびstd((R(q, o1), ..., R(q, oG)))正規化項。一方、
偏りのない最適化目標を実装するには、mask.sum(axis=dim)を置き換えることができます。
リスト1のマスク平均関数に定数（例えば、発電予算）を入れると、
緑色の線で強調表示されています。注目すべきは、これらの単純な変更によりPPOが回復されることです。
式（2）の目的関数は、モンテカルロリターンによって推定された利点を、偏りのない
ベースライン（Sutton & Barto, 2018年）の詳細な導出は付録Aで示しています。
GRPO博士と同様の最適化アルゴリズムを採用し、その有効性を実験的に検証します。

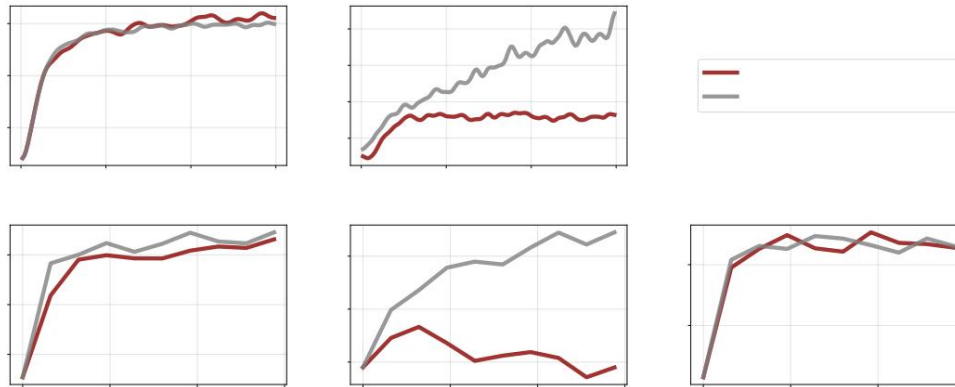


図 6: トレーニング ダイナミクス (上) と評価結果 (下)の観点から見た Dr. GRPO と GRPO の比較。

実験設定。私たちは、モジュール式で研究に適した効率的な LLM RL フレームワークである Oat (Liu et al., 2025a)を使用してアルゴリズムを実装します。オンライン RL チューニングには、Qwen2.5-1.5B ベースモデルと R1 テンプレート (テンプレート1) を採用しています。検証ベースの報酬関数は、次の最小限のルールを使用して、Math-Verify2 を使用して実装します。

$$R(q, o) = \begin{cases} 1 & \text{if } o \text{ is the correct final answer to } q \\ 0 & \text{otherwise} \end{cases}$$

MATH (Hendrycks et al., 2021) トレーニング データセットからサンプリングした質問に対して RL を実行し、バニラ GRPO と提案された Dr. GRPO を比較します。オンライン モデルを、AIME2024、AMC、MATH500、Minerva Math、OlympiadBench の 5 つのベンチマークで評価します。

ハイパーパラメータを含むより詳細な実験については、[オープンソースのコードベース](#)をご覧ください。

結果。図6にさまざまな指標を報告し、Dr. GRPO が最適化バイアスを効果的に緩和し、トークン効率を向上させることができることを示しています。特に、まず、GRPO と Dr. GRPO はどちらも DeepSeek-R1-Zero (Guo et al., 2025) と同様の傾向を示し、応答の長さがトレーニング報酬とともに増加することがわかります (プロット 1 および 2)。ただし、GRPO は、報酬の改善が遅くとも、継続的に長い応答を生成する傾向があることがわかります (プロット 2)。このような現象は、RL による long-CoT の「出現」と呼ばれることがよくありますが (Zeng et al., 2025; Hu et al., 2025)、最適化中の応答レベルの長さのバイアス (セクション3.1) によっても混乱すると主張します³。

対

照的に、GRPO 博士は、偏りのないポリシー勾配を計算することで、トレーニング中に応答の長さが急激に長くなるのを防ぎます (プロット 2)。さらに、評価ベンチマークでは、GRPO 博士はベースラインと比較して誤った応答の長さを大幅に短縮しており (プロット 4)、偏りのないオプティマイザーが考えすぎを軽減することも示唆しています (Chenら、2024)。

3.3 RLダイナミクスにおけるテンプレートと質問セットのカバレッジのデュエット

Qwen2.5-Math ベースモデルは、プロンプトテンプレートなしでも高い精度で質問に簡単に答えることができることを思い出してください (セクション2.2)。この興味深い観察に基づいて、さまざまなテンプレートが RL トレーニングにどのように影響するかに興味があります。さらに、質問セットの範囲が広いほどパフォーマンスが向上するという一般的な考え (Luo ら、2025 年、Hu ら、2025 年) を考慮して、さまざまなテンプレートとさまざまなレベルの質問範囲の間の相互作用も研究します。

²<https://github.com/huggingface/Math-Verify>.

³Zeng et al. (2025) と Hu et al. (2025) はどちらも PPO を採用しており、これは定式化。ただし、損失の実装では依然として長さのバイアスが発生します (リスト1 を参照)。

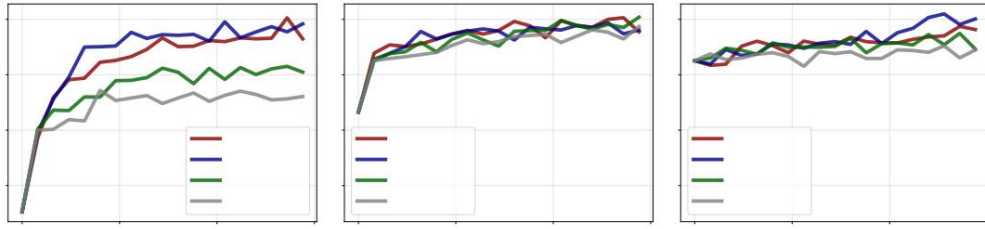


図7: RLトレーニング中のさまざまな{テンプレート、質問セット}の組み合わせの平均ベンチマーク精度。

実験設定。Qwen2.5-Math-1.5B ベースモデルから始めて、R1テンプレート、Qwen-Math テンプレート、テンプレートなしをそれぞれ適用し、Dr. GRPO を使用して RL を実行します。すべての実験は、表 3 に詳述されているさまざまな質問セットに対して繰り返されます。

質問セット	説明
オーズ	AIME、Numina-Math、Tulu3 MATH を組み合わせた多様で大規模な (57k)
数学	高校数学コンテスト問題 (12k)
グローバル	より簡単な小学校の算数の質問 (8k)
ASDiv	基本的な代数 (+ − × ÷) の問題 (2k)

表 3: 難易度と範囲が異なるさまざまな質問セット。

結果。図7 は、さまざまな実行の RL 曲線を示しています。この曲線から、いくつかの興味深い観察を行うことができます。1) テンプレートは初期ポリシーのパフォーマンスを決定しますが、RL はすべてのポリシーを 40% 程度の同等のパフォーマンスに向上させることができます (適切な質問セットが与えられている場合)。2) R1 テンプレートを使用する場合、質問セットは RL のダイナミクスに大きな影響を与え、範囲が狭すぎるとプラトー パフォーマンスが低下します。ただし、Qwen-Math テンプレートを使用する場合、GSM-8K の RL によって最高の最終パフォーマンスが達成され、はるかに単純な (そして良い) 質問でトレーニングすると、難しい質問のテスト精度が大幅に向上 (ほぼ 2 倍) することがわかります。これらの観察から、次の洞察が得られます。

- Qwen2.5-Math-1.5B ベースモデルは、すでに強力な数学計算能力を備えています (図7 の右側のプロットの開始点を参照)。テンプレートを適用すると、実際には RL が再構築する前に能力が破壊されます。これは、純粋な RL によってもたらされる大きな利益を主張する際には、より慎重になる必要があることを意味します。
- ベースモデルとテンプレートの間に大きな不一致がある場合 (例: R1 テンプレートが Qwen2.5-Math-1.5B と不一致)、ポリシーの改善は主に RL チューニングによってもたらされるため、質問セットが良好なカバレッジを持つことが必要になります (図7 の左のプロット)。それ以外の場合は、小さくて完全に適切な質問セットでも、新しい知識を注入するのではなく、正しい推論動作を強化することで、推論能力を同様に誘導できます。

3.4 ドメイン固有の事前トレーニングにより RL の上限が改善

最近の成功した R1-Zero のような数学推論システムの複製では、ほとんどの場合、初期ポリシーとして Qwen2.5 ベースモデルが採用されています (Zeng ら、2025 年、Cui ら、2025 年、Hu ら、2025 年)。これらはすでに強力な数学ソルバーであり、自己反映パターンを示しています (セクション 2.2 および 2.3)。このセクションでは、反対側を検討したいと思います。つまり、R1-Zero のようなトレーニングは、もともと弱い (数学推論の観点から) ベースモデルで成功できるのでしょうか? この質問には、数学の事前トレーニングによって RL の上限が向上するという観察に基づいて肯定的に答えます。

実験設定。Llama-3.2-3B ベースモデルを出発点として採用し、R1 テンプレートを使用した RL チューニングには偏りのない Dr. GRPO アルゴリズムを使用します。

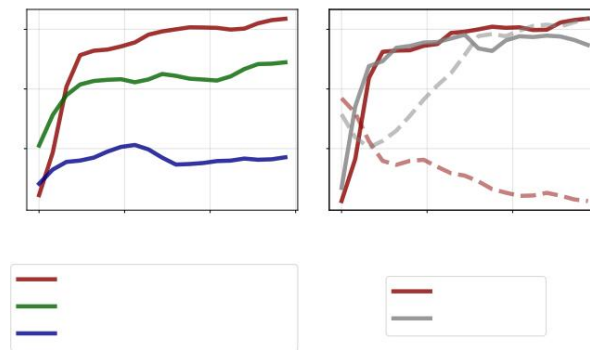


図 8: 左: さまざまなベース モデルの平均ベンチマーク パフォーマンス曲線。右: GRPO博士とGRPOの推論精度の比較 (実線)

実線) とモデル応答の長さ (破線)。

ドメイン固有の事前トレーニングはRLに役立つため、Llama-3.2-3B-FineMath4を採用する。これはFineMathデータセット (Allal et al., 2025)で継続的に事前学習されています。さらに、Qwen2.5モデルは連結された質問応答テキストで事前トレーニングされている可能性が高いという仮説を立て (2.2節)、同様にNuminaMath- 1.5 (Li et al., 2024)から連結されたデータセットを準備し、学習率 $1e-5$ で2エポックにわたってLlama-3.2-3B-FineMathを継続的に事前トレーニングします。連結された継続的な事前トレーニング済みモデルをLlama-3.2-3B-NuminaQAと呼びます。

結果。図8の左側のプロットに、異なるベースモデルのRL曲線を示します。RL はバニラの Llama ベース モデルを改善することさえできますが、得られるメリットは最小限であることがわかります。数学領域を埋め込むための継続的な事前トレーニング (および連結された継続的な事前トレーニング)の後知識を蓄積することで、Llamaモデルはより強力なRLパフォーマンスを発揮し、私たちの仮説を検証することができます。また、GRPOの最適化バイアスをLlamaベースモデルで再検討します。図8のプロットは、GRPOで訓練されたモデルのパフォーマンスと応答の長さを比較したもので、GRPO博士。GRPOが「二重増加」現象を生み出すことができることは明らかです。潜在的に、長いCoTはLlamaモデルでも出現する可能性があるという誤解につながる可能性がある。数学の事前トレーニング。残念ながら、長さの増加は最適化によるものかもしれない。バイアス (第3.1節)は、提案されたDr.GRPO (第3.2節および

図8の右側のグラフ)。

4 閉会の辞

R1-Zeroのようなトレーニングに使用されるベースモデルを批判的な視点から検討しました。強化学習に使われるアルゴリズムも分析しました。分析を通じて、事前学習がどのようにバイアスがRLの結果にどのように影響するか、そしてGRPOのような最適化の選択が意図せず形状モデルの挙動。提案されたDr.GRPOでは、改善する簡単な修正を提供します。推論性能を維持しながらトークン効率を向上させる。私たちの結果は、RLのスケールアップが

効果的かつ効率的であり、時には少ないほうがより効果的であることもあります。

参考文献

Arash Ahmadian, Chris Cremer, Matthias Galle, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Ustun, Sara Hooker. 基本に戻る: LLMsにおける人間のフィードバックから学習するための強化スタイルの最適化を再検討。arXiv プレプリント arXiv :2402.14740、2024年。

ルブナ・ベン・アラル、アントン・ロシュコフ、エリー・バクーシュ、ガブリエル・マルティン・プラスケス、ギリエルメ・ペネド、ルイス・タンストール、アンドレス・マラフィオティ、ハイネック・キドル・チェック、アグスト・ピケレス・ラハルン、

4<https://huggingface.co/HuggingFaceTB/FineMath-Llama-3B>.

Vaibhav Srivastav, 他 「SmolLM2: Smol がビッグデータ中心の小規模言語モデルのトレーニングに移行するとき」 arXiv プレプリント arXiv:2502.02737, 2025。

マルシン・アンドリコピッチ、アントン・ライチュク、ピョートル・スタンチュク、マヌ・オルシーニ、セルタン・ジルギン、ラファ・エル・マリニエ、レナード・フスノー、マシュー・ガイスト、オリヴィエ・ピエカン、マルシン・ミハルスキー 他。
オンポリシーディープアクタークリティック法にとって何が重要か? 大規模研究。 2021年国際学習表現会議にて。

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhang, Zhuosheng Zhang 他 2+3= はそこまで考えないでしょうか? o1 のような llms の考えすぎについて。 arXiv プレプリント arXiv:2412.21187, 2024。

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, Dario Amodei。
人間の好みからの深層強化学習。ニューラル情報処理システムの進歩, 30, 2017年。

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, 他。暗黙的な報酬によるプロセスの強化。 arXiv プレプリント arXiv:2502.01456, 2025。

アーロン・グラッタフィオーリ、アビマニュ・ダベイ、アビナフ・ジャウリ、アビナフ・パンデイ、アビシエク・カディアン、アフマド・アルダール、アイシャ・レットマン、アキル・マトゥール、アラン・シェルテン、アレックス・ヴォーン 他 ラマ3 の群れのモデル。
arXiv プレプリント arXiv:2407.21783, 2024。

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。

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, 他。OlympiadBench: オリンピックレベルのバイリンガルで多角的な科学的問題を伴う AG を促進するための挑戦的なベンチマーク。 arXiv プレプリント arXiv:2402.14008, 2024。

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, Jacob Steinhardt。
数学データセットによる数学の問題解決の測定。arXiv プレプリント arXiv:2103.03874, 2021 年。

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, Yu Cao。Open-rlhf: 使いやすく、スケーラブルで高性能な rlhf フレームワーク。 arXiv プレプリント arXiv:2405.11143, 2024。

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Heung-Yeung Shum, Xiangyu Zhang です。
Open-reasoner-zero: 基本モデルで強化学習をスケーリングするためのオープンソースのアプローチ。https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, 2025年。

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, および Rosalind Picard。対話における暗黙の人間の好み、ポリシーから大きく外れたバッチ深層強化学習。arXiv プレプリント arXiv:1907.00456, 2019 年。

ワウター・クール、ヘルケ・ファン・ホーフ、マックス・ウェリング。4 つの強化サンプルを購入し、ベースラインを取得します
無料!、2019年。

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, 他。Tulu 3: オープン言語モデルのポストトレーニングにおけるフロンティアの推進。arXiv プレプリント arXiv:2411.15124, 2024。

アイトール・レウコウィッチ、アンダース・アンドレアッセン、デヴィッド・ドーハン、イーサン・ダイアー、ヘンリク・ミハレフスキー、ヴィナイ・ラマシュ、アンブローズ・スローン、ジェム・アニル、イマノール・シュラク、テオ・ガットマン=ソロ 他
言語モデルによる定量的推論問題の解決。ニューラル情報処理システムの進歩, 35:3843–3857, 2022年。

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen 他. NuminaMath: 86 万組の競技数学問題と解答を含む、ai4maths 最大の公開データセット。
Hugging Face リポジトリ, 13/9/2024。

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, Stanislas Polu. Nu-minamath。 (<https://huggingface.co/AI-MO/NuminaMath-1.5>) (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)、2024年。

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan 他 Deepseek-v3 の技術レポート。 arXiv プレプリント arXiv:2412.19437, 2024。

Zichen Liu, Changyu Chen, Chao Du, Wee Sun Lee, および Min Lin. Oat: LLM オンライン アライメントのための研究に適したフレームワーク。 <https://github.com/sail-sg/oat>、2025年頃

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, および Min Lin. R1-zero のようなトレーニングでは、AI 体験は得られない可能性があります - パイロット スタディ。 <https://oatllm.notion.site/oat-zero>、2025b. Notion ブログ。

Michael Luo, Sijun Tan, Justin Wong, Xiaxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, Ion Stoica. Deepscaler: RL をスケールして 1.5b モデルで o1-preview を上回ります。 <https://github.com/agentica-project/deepscaler>、2025年。

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, Alane Suhr. タイニーゼロ。
<https://github.com/Jiayi-Pan/TinyZero>、2025年。アクセス日: 2025年1月24日。

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, Pieter Abbeel. 一般化利点推定を使用した高次元連続制御。 arXiv プレプリント arXiv:1506.02438, 2015 年。

ジョン・シュルマン、シー・チェン、ピーター・アビール。政策勾配とソフトの等価性
q学習。 arXiv プレプリント arXiv:1704.06440, 2017a。

ジョン・シュルマン、フィリップ・ウォルスキ、ブラフ・ダリワル、アレック・ラドフォード、オレグ・クリモフ。近位
ポリシー最適化アルゴリズム。 arXiv プレプリント arXiv:1707.06347, 2017b。

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu 他 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。

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, Chuan Wu. Hybridflow: 柔軟で効率的な rlhf フレームワーク。 arXiv プレプリント arXiv:2409.19256, 2024。

モハメド・シューイビ、モストファ・パトワリー、ラウル・ブリ、パトリック・ルグレスリー、ジャレッド・キャスパー、ブライアン・カタン
ザーロ。 Megatron-lm: モデル並列処理を使用して、数十億のパラメーター言語モデルをトレーニングします。 arXiv プレ
プリント arXiv:1909.08053, 2019。

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul F Christiano. 人間のフィードバックによる要約の学習。神経情報処理システムの進歩, 33:3008–3021, 2020年。

Richard S. Sutton および Andrew G. Barto. 強化学習: 入門。MIT出版, 第2版, 2018年。

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, Quentin Gallouedec. Trl: トランスフォーマー強化学習。 <https://github.com/huggingface/trl>、2020年。

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei 他。Qwen2.5 テクニカルレポート。arXiv プレプリント arXiv:2412.15115, 2024a。

An Yang, Beichen Zhang, Binyuan Hui, Bohei Gao, Bowen Yu, Chengpeng Liu, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin 他。Qwen2.5-math テクニカルレポート: 自己改善による数学的エキスパートモデルへ。arXiv プレプリント arXiv:2409.12122, 2024b。

エドワード・ヨー、ユシュアン・トン、モリー・ニウ、グラハム・ニュービッツ、シャン・ユエ。謎を解き明かす長い LMS における思考連鎖推論。arXiv プレプリント arXiv:2502.03373, 2025。

Weihaio Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, Junxian He。7b モデルと 8k の例: 強化学習による新たな推論は効果的かつ効率的です。<https://hkust-nlp.notion.site/simplerl-reason.2025> 年。Notion ブログ。

ポリシー勾配導出

LLM後の訓練のためのRLの文脈では、通常、

$$J(\pi_\theta) = E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot|q)} [R(q, o)], \quad (4)$$

ここで、 $R(q, o) = \sum_{t=1}^{|o|} r(q, o \leq t)$ は軌道 q ; o のリターン (Sutton & Barto, 2018) であり、 $o \leq t$ は、応答 o における t 番目のトークンのトークンレベルの報酬を表します。

式(4)のモンテカルロ政策勾配(Sutton & Barto, 2018)は、

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &= E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot|q)} [\nabla_\theta \log \pi_\theta(o|q) R(q, o)] \\ &= E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot|q)} [\nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) R(q, o)] \\ &= E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot|q)} [\nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \sum_{t'=t}^{|o|} r(q, o \leq t')] \\ &= E_{q \sim p_Q} E_{o \sim \pi_\theta(\cdot|q)} [\nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) \sum_{t'=t}^{|o|} r(q, o \leq t') - B(q, o_{<t})] \end{aligned} \quad (5)$$

ここで、 $B(q, o_{<t})$ は分散減少項であり、 o_t に関して不変であるため、 $[\nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})] = E_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} [\nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})]$ となる。

$$= [\sum_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})] = [\sum_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} \nabla_\theta \log \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})]$$

$$\nabla_\theta \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t}) = [\nabla_\theta \sum_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})]$$

$$\begin{aligned} &= [\nabla_\theta \sum_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})] \\ &= [\nabla_\theta \sum_{o_t \sim \pi_\theta(\cdot|q, o_{<t})} \pi_\theta(o_t|q, o_{<t}) B(q, o_{<t})] = 0 \text{ です。} \end{aligned}$$

通常、 $B(q, o_{<t}) = [\sum_{o \geq t} E_{o \sim \pi_\theta(\cdot|q, o_{<t})} r(q, o \leq t)]$ は、期待される累積

将来の相対報酬 (現在の状態の値とも呼ばれる) を表し、 $|o|$ と表記します。

$A(o_t|q, o_{<t}) = \sum_{t'=t}^{|o|} r(q, o \leq t') - B(q, o_{<t})$ を利点として考える。結果報酬の場合、 $\sum_{t'=t}^{|o|} r(q, o \leq t') = \sum_{t=1}^{|o|} r(q, o \leq t) = R(q, o)$ です。

$B(q, o_{<t}) = \text{mean}(\{R(q, o_1), \dots, R(q, o_G)\})$ と設定すると、式(5)のポリシー勾配は、

$$\nabla_\theta J(\pi_\theta) = E_{q \sim p_Q} E_{\{o_i\}_{i=1}^G \sim \pi_\theta(\cdot|q)} \left[\frac{1}{G} \sum_{t=1}^{|o_i|} \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}) A(o_{i,t}|q, o_{i,<t}) \right], \quad (6)$$

どこ

$$= \frac{R(q, o_i) - \text{平均}(\{R(q, o_1), \dots, R(q, o_G)\})}{\text{std}(\{R(q, o_1), \dots, R(q, o_G)\})}$$

式(6)を計算するためにPPO (Schulman et al., 2017b) 目的関数を採用する。

$$\begin{aligned} J(\pi_\theta) &= E[q \sim p_Q, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)] \\ &= \frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})} A(o_{i,t}|q, o_{i,<t}) \frac{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} \cdot 1 - \dots, 1 + A \dots \end{aligned}$$

そこから、stdと $|o|$ の両方がRL目的関数に現れるべきではないという結論が導き出されます。

Aの公平性 i, t Aに注目する。 上記で計算された i, t はREINFORCEの i, t と同等である。
Leave-One-Out (RLOO) (Ahmadian et al., 2024; Kool et al., 2019) をスケール係数まで適用すると、
RLダイナミクスに影響を与えずに学習率に組み込むことができる。具体的には、

$$\begin{aligned} \frac{G}{G-1} \cdot \text{あ}_{\text{それ}} &= \frac{G}{G-1} R(q, oi) - \frac{G}{G-1} \frac{1}{G} \sum_{j=1}^G R(q, oj) \\ &= \frac{G}{G-1} R(q, oi) - \frac{1}{G-1} \sum_{j=1, j \neq i}^G R(q, oj) - \frac{1}{G-1} R(q, oi) \\ &= A_{\text{RLOO}_{\text{それ}}} \end{aligned}$$

B 詳細なベンチマーク結果

3つのスケール (1.5B、3B、7B)の詳細なベンチマーク結果を表4に示す。
比較のために、同じスケールの instruct モデルと R1-Distill モデルも含めます。
コンテキスト長が
4kの場合、比較するすべてのベースラインで生成予算を3kに制限します。
より長いコンテキスト (OpenReasoner-Zero end R1-Distill-Qwen)向けに訓練された結果、
8K 世代予算でのパフォーマンス。

ベースモデル + メソッド	AIME24 AMC MATH500 Minerva OlympiadBench平均						
Qwen2.5-Math-1.5B	20.0	32.5	16.7	33.0	12.5	22.8	24.2
Qwen2.5-Math-1.5B*オートゼ	43.4	20.0	53.0	61.8	15.1	28.4	33.1
ロ-1.5B				74.2	25.7	37.6	42.1
R1-蒸留-Qwen-1.5B @ 3k R1-蒸留-	2.5	21.7	20.0	52.2	16.3	17.3	22.0
Qwen-1.5B @ 8k Qwen2.5-数	49.4	10.0	48.2	77.4	25.0	35.8	41.5
学-1.5B-命令				74.2	26.5	40.2	39.8
Llama-3.2-3B + RL	0.0	2.4	6.4	6.3	1.3	3.3	
w. Dr. GRPO Llama-3.2-3B-	3.3	7.2	10.0	11.0	2.2	6.8	
FineMath + RL w. Dr. GRPO	0.0	3.6	18.4	5.9	2.2	6.0	
Llama-3.2-3B-NuminaQA	3.3	10.8	38.0	12.9	9.0	14.8	
+ RL w. Dr. GRPO (Oat-Zero-3B) 6.7	0.0	0.0	0.6	0.0	0.1	0.14	
Llama-3.2-3B-Instruct 6.7		18.1	50.0	14.3	14.7	20.7	
		15.7	38.8	11.8	12.6	17.1	
Qwen2.5-Math-7B	16.7	38.6	0.2	50.6	9.9	16.6	26.5
Qwen2.5-Math-7B*	45.8	26.7	60.2	69.0	21.3	34.7	38.2
SimpleRL-Zero-7B PRIME-	16.7	62.7	13.3	78.2	27.6	40.3	46.6
Zero-7B	47.0	13.3	54.2	83.8	36.0	40.9	48.0
OpenReasoner-Zero-7B @ 3k	43.3	62.7		79.2	31.6	44.0	43.0
OpenReasoner-Zero-7B @ 8k				82.4	31.6	47.9	45.9
オートゼロ7B				80.0	30.1	41.0	51.4
R1-蒸留-Qwen-7B @ 3k R1-蒸留-	10.0	26.2	60.1	23.0	23.1	28.5	
Qwen-7B @ 8k Qwen2.5-	33.3	68.4	88.1	35.9	47.7	54.7	
Math-7B-Instruct	16.7	53.0	83.6	29.8	42.7	45.1	

表4: ベンチマークスコアの比較。Oat-Zeroモデルは、
ミニマリストレシビ (第1節)。*は、最適なテンプレート (テンプレートなし)を使用して生成することを意味します。
回答は、テストのスコアが最も高くなり、受験者の能力を忠実に反映できるように、
ベースモデル。

C キーワードベースの検出とLLMベースの識別
自己反省行動

私たちは、自己反省を示す厳選されたキーワードとフレーズのプールを構築します
LLMの応答には行動が含まれている。しかし、LLMが生成した応答には、

幻覚やトピックから外れた内容が多く含まれているため、必ずしも本物の内省を示すとは限らない単純であいまいなキーワードが存在することになります。たとえば、「待つ」や「もう一度試す」などの用語は、誤検知につながることがよくあります。誤検知を減らすために、私たちは、内省を強く示唆する用語で構成された、小規模で高度に選択的なキーワード プールを維持しています。私たちの実験では、キーワード プールは、recheck、rethink、reassess、reevaluate、re-evaluate、reevaluation、re-examine、reexamine、reconsider、reanalyze、double-check、check again、think again、verify again、go over the steps に限定されています。

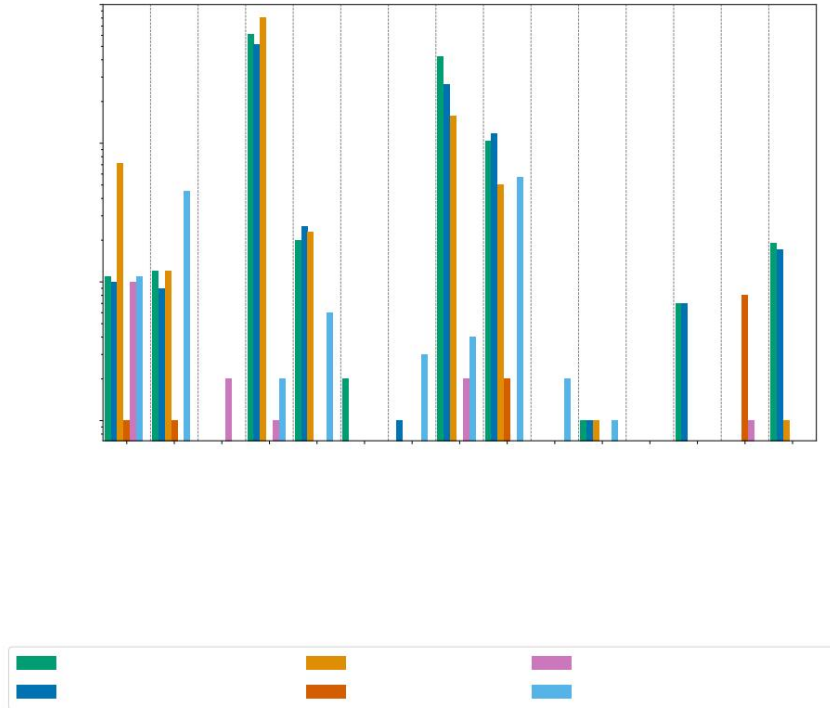


図 9: 40,000 件の回答のうちのキーワードの出現回数 (500 件の質問×質問あたり8 件の回答× 10 の温度)。y は対数スケールです。

図 9 では、さまざまなモデルによって生成された応答におけるさまざまなキーワードの出現を示します。興味深いことに、異なるモデル ファミリーでは強調されるキーワードが異なります。たとえば、「再度確認する」、「再確認する」、「再評価する」、「再調査する」、「再確認する」、「再検討する」、「再度検証する」などのフレーズは、Qwen2.5 ファミリーで最も頻繁に登場します。対照的に、「再評価する」、「再調査する」、「再度検証する」は DeepSeek ファミリーの応答には現れませんが、Llama モデルでは「もう一度考える」というフレーズが頻繁に使用されます。私たちは、この現象は、特に推論と数学に関して、事前トレーニング データの違いによって生じると仮定しています。

キーワードプールは慎重に選択していますが、特定のキーワードを含まない暗黙的な内省行動を識別するには不十分な場合があります。さらに、図10のケース (a) に示すように、誤検知につながる可能性があります。これらの制限に対処し、ベースモデルの内省機能をより正確に評価するために、より強力な LLM (実験では GPT-4o-mini) を活用して応答を分析し、明示的な内省 (「再確認」や「再評価」などのキーワードなど) を示しているのか、暗黙的な内省 (キーワードマッチングでは簡単に捉えられない、より洗練されたパターンなど) を示しているのかを判断します。このアプローチは、真の内省行動と、関連用語の表面的または偶発的な使用を区別するのに役立ちます。

LLMベースの検出は、キーワードベースの検出から誤検出を効果的に排除し、暗黙の自己反省行動を識別しますが、応答が長くて複雑な場合は特に、誤分類される可能性があります。たとえば、図10のケース (b) は、LLMベースの検出で誤検出を示しており、応答は自己反省として分類されています。

R1-Zero のようなトレーニングを理解する: 批判的な視点

Case(a): What is the smallest positive multiple of 23 that is 4 more than a multiple of 89?

Response

To find the smallest positive multiple of 23 that, ...
Start with the smallest possible value of k by checking if they satisfy the equation. If not, increase k and **check again**.

...
Now, trying $k = 21$: $23 \cdot 21 = 483$...

...

Case(b): What is the greatest multiple of 3 that is less than 500 and is also a multiple of 4?

Response

To find the greatest multiple of 3 that is less than 500 and is also a multiple of 4, I need to first consider the greatest multiple of 4 less than 500, which is 496. However, 496 is not a multiple of 3. I then need to consider the next multiple of 4, which is 492. 492 is a multiple of 3, since $492 / 3 = 164$ with no remainder. </think>

<answer> 492 <answer>

図 10: ケース (a): キーワードベースの検出における誤検知。 ケース (b): LLM ベースの検出における誤検知。

LLM によって検出されるが、実際には自己反映を示さないエラー。このタイプのエラーは、キーワードベースの検出によって除去できます。堅牢性を高めるために、キーワードベースと LLM ベースの検出をクロス検証によって統合します。組み合わせた検出結果と、キーワードベースおよび LLM ベースの方法からの個々の結果を図11に示します。

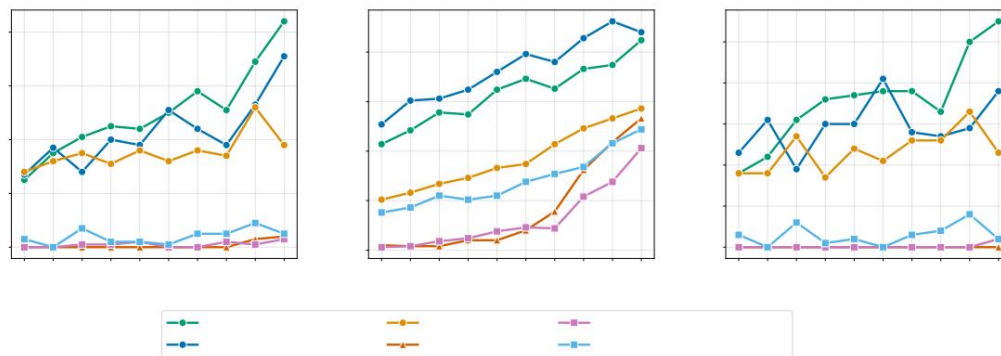


図 11: キーワードベースの検出、LLM ベースの検出、およびクロス検出の比較。自己反省は 500 の質問にわたって質問レベルでカウントされ、8 つの回答のうち少なくとも 1 つに自己反省が見られる場合、その質問は自己反省があるとマークされます。

D DeepSeek-V3-BaseとDeepSeek-R1-Zeroの比較

DeepSeek-V3-Base と DeepSeek-R1-Zero を分析して、R1-Zero トレーニング中のモデルの動作の変化を理解します。図12 は、両方のモデルで評価した 500 の数学問題に対する難易度別の応答カテゴリの内訳を示しています。結果は、ほとんどの誤った応答が RL トレーニング後に修正され、R1-Zero トレーニングによる大幅なパフォーマンスの向上を示していることを示しています。一方、フォーマットされていない応答が増加していることがわかりました。これは、Liu ら(2025b)の観察と一致しています。

表5 では、カテゴリ全体の平均回答長を報告しています。切り捨てられた回答は、より大きなコンテキスト サイズを使用した場合、他の 3 つのカテゴリのいずれかに分類されることに注意してください。そのため、表から除外しています。結果は、正解を含むすべてのカテゴリで回答長が大幅に増加していることを示しています。これは、Guo ら(2025)の図 3 の結果と一致しています。ただし、不正解の回答の平均長は、正解の回答よりも著しく長くなっています。これは、より難しい質問では、推論の複雑さが増すため、一般的に回答が長くなるためであると仮定しています。

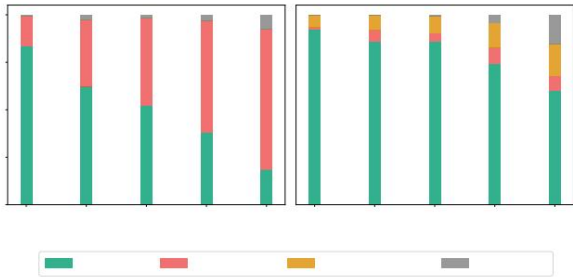


図12: 回答カテゴリーの内訳
DeepSeek-V3-Base および DeepSeek-R1-Zeroの MATH データセットの難易度レベル。

カテゴリ	ベースR1-ゼロ
フォーマットされていない	880.7
正しい	7870.3
間違っている	4965.4
1038.9	8206.1

表5: 平均応答文字列
カテゴリー間の長さ
DeepSeek-V3-Base (ベース)と
DeepSeek-R1-Zero (R1-Zero)。

難しい質問では間違った回答が多くなり、結果的に平均的な長さ。

自己反省は必ずしも高い精度を意味するわけではない。自己反省行動がモデルのパフォーマンスと関連しているかどうかを調べるために、推論（自己反省が訓練中の探索を改善する（このセクションの範囲外の潜在的なプラス効果）ため、我々は分析する8回の試行を通じて DeepSeek-R1-Zero から少なくとも1つの自己反省を伴う応答を引き出す質問。各質問に対して100件の回答を抽出し、彼らを2つのグループに分けます。自己反省のあるグループと次に、各質問について、これら2つのグループ間の正解率の差を計算します。

図13に示されているように、結果はほぼ半数自己反省を伴う回答は、より高い成果を達成しない自己反省のない人よりも正確性が高く、自己反省は必ずしも高い DeepSeek-R1-Zero の推論段階の精度。

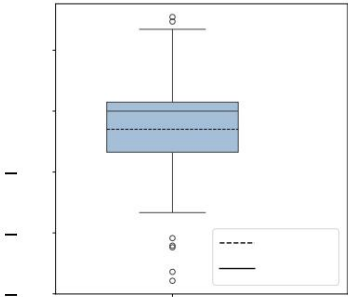


図13:回答の有無による精度の違い
各質問に対する自己反省（ DeepSeek-R1-Zeroからサンプリングされた回答）。

GPT-As-A-Judge で使用される E プロンプト

応答に自己反省行動が含まれているかどうかを判断するために、LLM ベースの検出を要求します。

LLM ベースの自己反省検出

数学の問題と詳細な回答をお送りします。あなたの仕事は、回答が質問に答えようとしているかどうかを判断することです。回答がトピックから外れていたり、幻覚的であったり、ランダムな話であったり、その他無関係である場合は、0 とマークしてください。

それ以外の場合は、応答が自己反省を示しているかどうかを評価します。

分類ルール: 1. カテゴリー0: 回

答がトピックから外れている、意味不明、一貫性がない、過度に繰り返しが多かったり、論理的思考に欠けている。

- 例: -回答が質問と関係がない。
-無意味または幻覚的な内容が含まれています。
-一貫性のない過度な繰り返しが構成されています。

2. カテゴリー1: 回答は質問に答えようとしていますが、自己反省は示されていません。

- 例: -応答により、手順を再実行せずに問題が直接解決されます。
-回答の正確さを検証したり、代替の解決策を検討したりする試みは行われません。

3. カテゴリー2: 回答はあらゆるレベルで自己反省を示しています。

- これには、次のものが含まれます。-明示的な自己反省のキーワード、たとえば、*再確認、再考、再評価、再評価、再評価、再評価、再評価、再評価、再確認、再度確認、再度考える、再度検証、手順を確認する* など。

-解決策を再検討したり、仮定に疑問を呈したり、明確なキーワードなしで代替アプローチを検討したりするなど、暗黙の自己反省行動。 •何らかの形の自己反省がある場合は、常に2 に分類し、

正確性や回答の質が低下します。

4. カテゴリー3: 応答は、自己反省を示さずに計算用の Python コードのみで構成されています。 •事例: -応答では、検証、再評価、または代替の検討を行わずに、ソリューションを計算する Python スクリプトのみが提供されています。

出力形式: 回答では、ま

ず分析の簡単な説明を記述し、最後に1つのカテゴリ番号 (0、1、2、または3) を記述する必要があります。回答の最後にはカテゴリ番号を必ず含めてください。

出力例:

- 「回答はトピックから外れており、質問に答えようとしていません。0。」
- 「回答は自己反省なしに直接的な解決策を提供しています。1。」
- 「回答は自己反省を示しています。2。」
- 「回答は自己反省のない Python コードのみで構成されています。3。」

問: {question}

応答: {応答}

モデルの質問応答能力を確認するためのプロンプト。

質問応答能力をチェックするためのプロンプト

私はあなたに質問と、LLM によって生成された長い応答を送信します。あなたの仕事は、出力が質問に答えようとしているかどうかを判断することです。出力には、無関係なコンテンツ、幻覚、またはランダムでトピックから外れた応答が含まれる場合があります。

出力を次のいずれかのカテゴリに分類してください: 出力形式: 応答は1つの整数(0 または 1) で始まり、その後に簡単な説明が続く必要があります。

- 0 を返す: →出力は質問に答えようとしていません (例: 無関係な内容、ランダムな会話、幻覚)。出力例: '0: 応答はトピックから外れており、質問に答えていません。'
- 1 を返します: →回答の完全性や正確性に関係なく、出力は質問に答えようとしています。出力例: 「1:回答が不完全または不正確であっても、応答は質問に対応します。」

質問: {質問}

応答: {応答}