

# DAPO: オープンソースの LLM 強化学習 大規模システム

1ByteDanceシード    2 清华大学AI産業研究所 (AIR)  
3香港大学  
清華AIRとByteDance Seedの4SIA-Lab  
寄稿者リストの全著者リスト

## 抽象的な

推論スケーリングは、強化学習を複雑な推論を引き出すための中核技術として、LLMに前例のない推論能力を与えます。しかし、最先端の推論LLMの主要な技術的詳細は隠蔽されているため(OpenAI o1ブログやDeepSeek R1技術レポートなど)、コミュニティは依然としてRLトレーニング結果を再現するのに苦労しています。私たちは、Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO)アルゴリズムを提案し、Qwen2.5-32Bベースモデルを用いてAIME 2024で50ポイントを達成する最先端の大規模RLシステムを完全にオープンソース化します。トレーニングの詳細を控えているこれまでの研究とは異なり、私たちは大規模LLM RLを成功させるアルゴリズムの4つの主要な手法を紹介します。さらに、慎重にキュレーションおよび処理されたデータセットとともに、verlフレームワーク上に構築されたトレーニングコードをオープンソース化します。

私たちのオープンソース システムのこれらのコンポーネントは再現性を高め、大規模 LLM RL の将来の研究をサポートします。

日付: 2025年3月17日連絡先:  
Qiyang Yu ( yuqy22@mails.tsinghua.edu.cn )プロジェクトページ: <https://daposia.github.io/>  
<https://github.com/volcengine/verl>

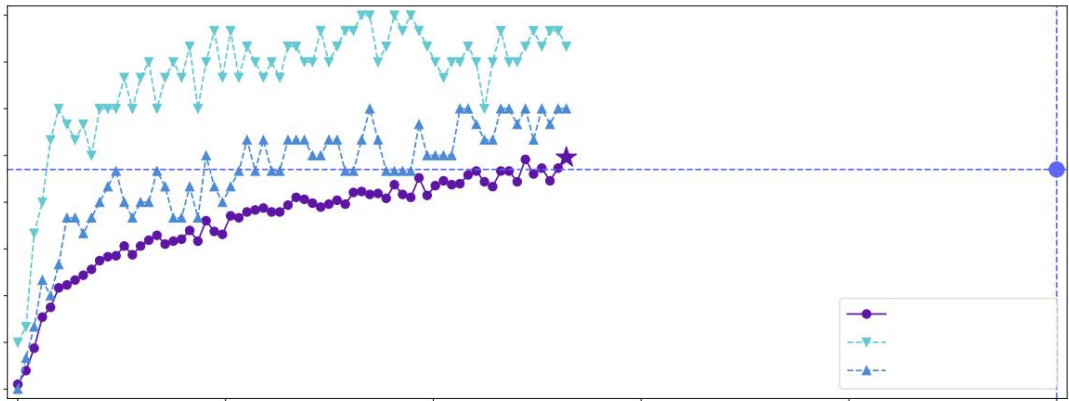


図 1 AIME 2024による Qwen2.5-32B ベース モデルのDAPOスコア。50 % のトレーニング ステップを使用した従来の SoTA DeepSeek-R1-Zero-Qwen-32B を上回っています。

1 はじめに

OpenAIのo1 [1]やDeepSeekのR1 [2]などのテスト時間スケールリングは、大規模言語モデル（LLM） [3-7]に大きなパラダイムシフトをもたらします。テスト時間スケールリングは、より長い思考連鎖（Chain-of-Thought）思考を可能にし、洗練された推論行動を誘発します。これにより、AIMEやCodeforcesのような競争的な数学およびコーディングタスクにおいて、モデルは優れた性能を発揮します。

この革命を推進する中心的な技術は、大規模な強化学習（RL）であり、自己検証や反復的改良といった複雑な推論行動を引き出します。しかしながら、スケーラブルなRL学習の実際のアルゴリズムと鍵となるレシピは依然として神話であり、既存の推論モデル[1, 2, 8-11]の技術レポートからは隠されています。本稿では、大規模なRL学習における重大な障害を明らかにし、完全にオープンソース化されたアルゴリズム、学習コード、データセットを備えたスケーラブルなRLシステムをオープンソース化します。このシステムは、業界レベルのRL結果を備えた民主的なソリューションを提供します。

強化学習（RL）の事前学習済みモデルとしてQwen2.5-32B [12]を用いた実験を行いました。最初のGRPO実行では、AIMEでわずか30ポイントしか獲得できず、DeepSeekの強化学習（47ポイント）を大幅に下回る結果となりました。徹底的な分析の結果、ナイーブなGRPOベースラインには、エントロピー崩壊、報酬ノイズ、学習の不安定性など、いくつかの重要な問題があることが明らかになりました。より広範なコミュニティは、DeepSeekの結果を再現する際に同様の課題に直面しており[13-19]、R1論文では、業界レベルの大規模かつ再現可能な強化学習システムの開発に必要な重要な学習の詳細が省略されている可能性があることを示唆しています。

このギャップを埋めるため、大規模LLM強化学習（RL）のためのオープンソースの最先端システムをリリースしました。このシステムは、Qwen2.5-32BモデルをベースにAIME 2024で50ポイントを達成し、50%の学習ステップでDeepSeek-R1-Zero-Qwen-32B [2]が達成した従来の最先端結果（47ポイント）を上回りました（図1）。本研究では、Decoupled Clip and Dynamic sAmpling Policy Optimization（DAPO）アルゴリズムを提案し、長期CoT強化学習（RL）シナリオでRLを効果的に活用するための4つの重要な手法を紹介します。詳細は第3章で説明します。

- 1. システムの多様性を促進し、エントロピーの崩壊を回避するClip-Higher.2 . トレーニングの効率と安定性を向上させる
- Dynamic Sampling。
- 3. トークンレベルのポリシー勾配損失。これは長時間の CoT RL シナリオでは重要です。
- 4. 報酬ノイズを減らし、トレーニングを安定させる、Overlong Reward Shaping。

私たちの実装はverl [20]に基づいています。学習コードとデータを含む最先端のRLシステムを完全公開することで、大規模LLM RLに関する貴重な知見を明らかにし、より大きなコミュニティに利益をもたらすことを目指しています。

2 予備

2.1 近似ポリシー最適化（PPO）

PPO [21] は、方策最適化のためにクリップされた代替目的関数を導入する。クリップを用いて方策の更新を以前の方策の近傍領域に制限することで、PPOは学習を安定化し、サンプル効率を向上させる。具体的には、PPOは以下の目的関数を最大化するように方策を更新する。

$$J_{PPO}(\theta) = E_{(q, a) \sim D, 0 \leq t} \pi_{\theta_{old}}(\cdot | q) \frac{\pi_{\theta}(ot | q, o < t)}{\pi_{\theta_{old}}(ot | q, o < t)} A_t, \quad \frac{\pi_{\theta}(ot | q, o < t)}{\pi_{\theta_{old}}(ot | q, o < t)} \cdot 1 - \epsilon, 1 + \epsilon \quad A_t, \quad (1)$$

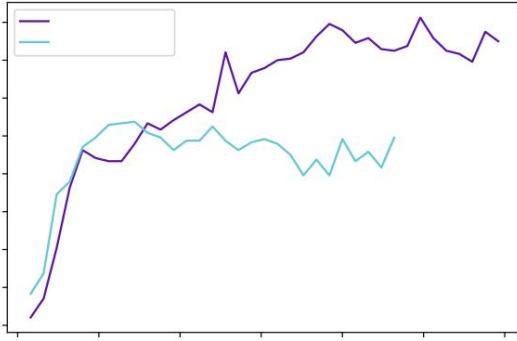
ここで、(q, a)はデータ分布Dからの質問と回答のペア、εは重要度サンプリング比のクリッピング範囲、Aは時間ステップtにおける優位性の推定値である。価値関数Vと報酬関数Rが与えられれば、A

t 一般化利点推定(GAE) [22]を使用して計算されます。

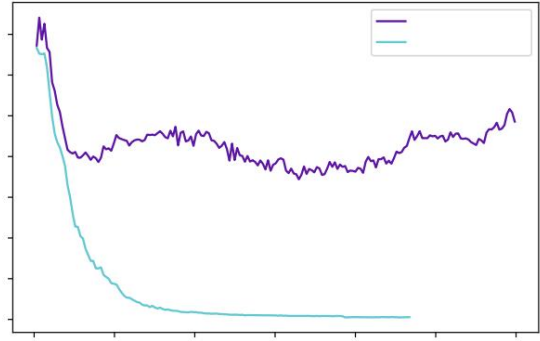
$$A_t^{GAE(\gamma, \lambda)} = \sum_{l=0}^{\infty} (\gamma \lambda)^l \delta_{t+l}, \quad (2)$$

どこ

$$\delta_l = R_l + \gamma V(s_{l+1}) - V(s_l), \quad 0 \leq \gamma, \lambda \leq 1. \quad (3)$$



(a) AIMEの精度。



(b) アクターモデルのエントロピー。

図 2 Clip-Higher 戦略を適用する前と適用した後の、AIME テスト セットの精度と RL トレーニング プロセス中のアクター モデルによって生成された確率のエントロピー。

## 2.2 グループ相対ポリシー最適化 (GRPO)

PPOと比較すると、GRPOは価値関数を排除し、グループ相対的な方法で優位性を推定する。特定の質問と回答のペア(q, a)に対して、行動方針 $\pi_{\text{old}}$ はG個の個別応答のグループ{o<sub>i</sub>}をサンプリングする。

次に、i番目の応答の利点は、グループレベルの正規化によって計算されます。

報酬 $\{R_i\}_{i=1}^G$ :

$$A_{i,t} = \frac{r_i - \frac{\sum_{i=1}^G R_i}{G}}{\text{std}(\{R_i\}_{i=1}^G)} \quad (4)$$

PPO と同様に、GRPO はクリップされた目標と、直接課される KL ペナルティ項を採用します。

$$J_{\text{GRPO}}(\theta) = E(q, a) \sum_{i=1}^G \pi_{\text{old}}(a_i | q) \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \min(r_{i,t}(\theta) A_{i,t}, \text{クリップ}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon)) - \beta \text{DKL}(\pi_{\theta} || \pi_{\text{ref}}) \right] \quad (5)$$

どこ

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_i, t | q, o_{i,<t})}{\pi_{\text{old}}(o_i, t | q, o_{i,<t})} \quad (6)$$

GRPOはサンプルレベルで目的関数を計算する点にも注目すべきです。正確には、GRPOはまず各生成シーケンス内の平均損失を計算し、その後、異なるサンプルの損失を平均化します。3.3節で説明するように、この差異はアルゴリズムの性能に影響を与える可能性があります。

## 2.3 KLダイバージェンスの除去

KLペナルティ項は、オンラインポリシーと固定参照ポリシー間の乖離を制御するために使用されます。RLHFシナリオ[23]では、強化学習の目標は、初期モデルから大きく乖離することなくモデルの挙動を整合させることです。しかし、long-CoT推論モデルの学習中は、モデルの分布が初期モデルから大きく乖離する可能性があるため、この制限は不要です。したがって、提案アルゴリズムではKL項を除外します。

## 2.4 ルールベース報酬モデリング

報酬モデルの使用は、通常、報酬ハッキング問題[24–29]に悩まされます。その代わりに、検証可能なタスクの最終的な精度を、以下の規則を用いて計算し、結果報酬として直接使用します。

$$R(y, y) = \begin{cases} \text{is\_equivalent}(y, y) & 1, \\ -1, & \text{それ以外} \end{cases} \quad (7)$$

ここで、 $y$ は真の解、 $y$ は予測解である。これは、自動定理証明[30–33]、コンピュータプログラミング[34–37]、数学コンテスト[2]など、複数の分野で示されているように、ベースモデルの推論能力を活性化する効果的なアプローチであることが証明されている。

## 3 ダポ

Decouple Clip and Dynamic sAmbling Policy Optimization (DAPO) アルゴリズムを提案します。DAPOは、各質問 $q$ と回答 $a$ の組み合わせについて、出力 $\{oi\}$ のグループをサンプリングし、以下の目的関数を用いてポリシーを最適化します。

$$J_{DAPO}(\theta) = E(q, a) \sum_{i=1}^G \pi_{\theta}(\text{old}(oi | q)) \left[ \frac{1}{G} \sum_{i=1}^G \min_{ri, t(\theta)} A_{i, t}^{\theta} - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}} \right] \quad (8)$$

st  $0 < \{oi \mid \text{is\_equivalent}(a, oi)\} < G$ ,

どこ

$$ri, t(\theta) = \frac{\pi_{\theta}(oi, t \mid q, oi, < t)}{\pi_{\theta}(\text{old}(oi, t \mid q, oi, < t))}, \quad A_{\text{それ}} = \frac{Ri - \text{平均}(\{Ri\} \text{標準偏差}_{i=1}^G)}{\text{差}(\{Ri\})} \quad (9)$$

完全なアルゴリズムはアルゴリズム 1 にあります。このセクションでは、DAPOに関連する主要な手法を紹介します。

### 3.1 上限を上げる :クリップハイアーナイーブPPO [21]ま

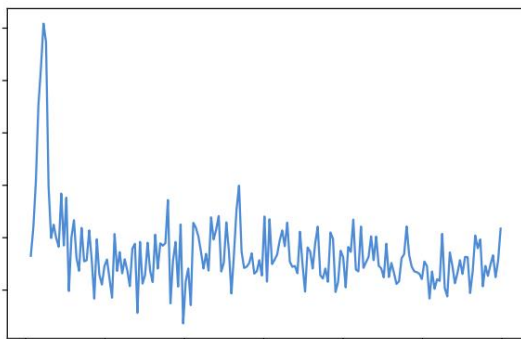
たはGRPO [38]を用いた初期実験では、エントロピー崩壊現象が観察されました。これは、学習が進むにつれて方策のエントロピーが急速に減少する現象です (図2b)。特定のグループのサンプル応答はほぼ同一になる傾向があります。これは、探索が限られており、初期の方策が決定論的であることを示しており、スケールアッププロセスを妨げる可能性があります。

この問題に対処するため、我々はClip-Higher戦略を提案する。重要度サンプリング比のクリッピングは、Clipped Proximal Policy Optimization (PPO-Clip) [21]で導入されており、信頼領域を制限して強化学習の安定性を向上させる。我々は、この上限クリッピングが方策の探索を制限する可能性があることを明らかにした。この場合、「利用トークン」の確率を高める方が、確率の低い「探索トークン」の確率を高めるよりもはるかに容易である。

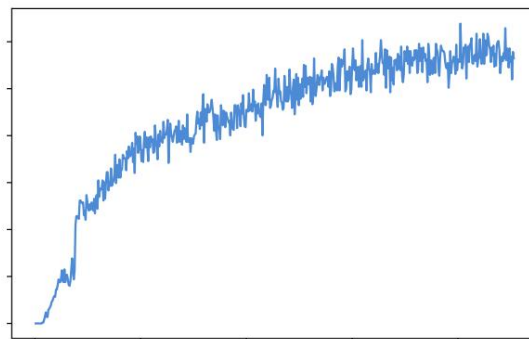
具体的には、 $\epsilon = 0.2$  (ほとんどのアルゴリズムのデフォルト値)の場合、確率 $\pi_{\theta}(\text{old}(oi \mid q)) = 0.01$ と $0.9$ の2つのアクションを考えてみましょう。更新後の確率 $\pi_{\theta}(oi \mid q)$ の最大値はそれぞれ $0.012$ と $1.08$ です。これは、より高い確率 (例えば $0.9$ )のトークンでは制約が緩いことを意味します。逆に、低い確率のトークンでは、確率を大幅に増加させることはかなり困難です。

経験的に、クリッピングされたトークンの最大確率はおよそ $\pi_{\theta}(oi \mid q) < 0.2$ であることが観察されています (図3a)。この知見は、クリッピングの上限閾値が低確率トークンの確率増加を実際に制限し、それによってシステムの多様性を潜在的に制限するという我々の分析を裏付けています。

Clip-Higher戦略に従って、下限クリッピング範囲と上限クリッピング範囲を $\epsilon_{\text{low}}$ と $\epsilon_{\text{high}}$ として分離します。



(a)最大クリップ確率。



(b)精度が 1 のサンプルの割合。

図 3アクターモデルの確率分布のエントロピーと応答の長さの変化。

式10で強調されている：

$$JDAPO(\theta) = E(q, a) \quad D_{\{oi\}} G \quad i=1 \quad \pi_{\theta old}(\cdot|q) \quad |b_i|$$

$$\frac{1}{G} \sum_{i=1}^G \min_{r_i, t(\theta)} A^{i, t, \text{クリップ}} r_i, t(\theta), 1 - \epsilon_{low}, 1 + \epsilon_{high} \quad A_{\text{それ}} \quad (10)$$

st  $0 < \{oi \mid \text{is\_equivalent}(a, oi)\} < G$  です。

低確率トークンの増加のための余地を残すため、 $\epsilon_{high}$ の値を増加させます。図2に示すように、この調整は方策のエントロピーを効果的に高め、より多様なサンプルの生成を促進します。 $\epsilon_{low}$ は比較的小さく保つことを選択しました。これは、 $\epsilon_{low}$ の値を増加させると、これらのトークンの確率が0に抑制され、サンプリング空間が崩壊してしまうためです。

### 3.2 多ければ多いほど良い :ダイナミックサンプリング

既存の強化学習アルゴリズムは、一部のプロンプトの精度が1である場合に勾配減少問題に悩まされます。例えばGRPOの場合、特定のプロンプトのすべての出力 $\{oi\}$ が正しく、同じ報酬1を受け取る場合、このグループのアドバンテージは0になります。アドバンテージが0になると、ポリシー更新の勾配がなくなり、サンプル効率が低下します。経験的には、図3bに示すように、精度が1であるサンプルの数は増加し続けます。これは、各バッチ内の有効なプロンプト数が減少し続けることを意味し、勾配の分散が大きくなり、モデルトレーニングにおける勾配シグナルが弱まる可能性があります。

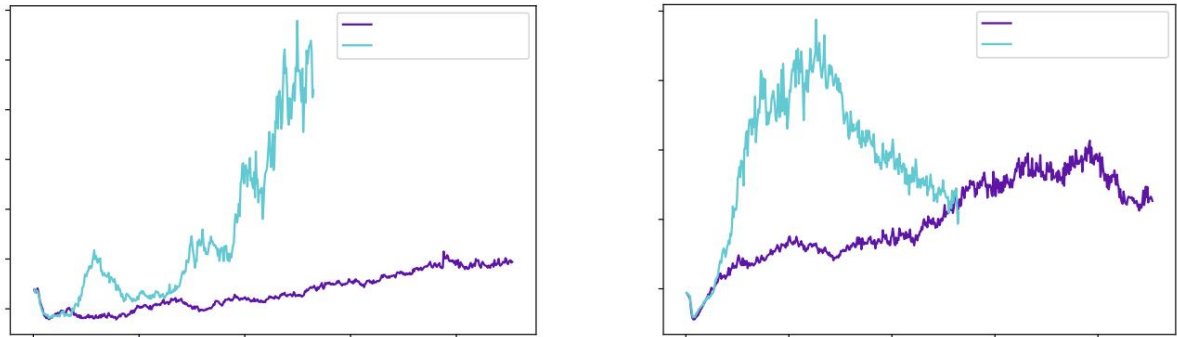
この目的のため、式11に示すような精度が1と0のプロンプトをオーバーサンプリングして除外し、バッチ内のすべてのプロンプトに有効な勾配を残しつつ、プロンプトの数を一定に保つことを提案する。学習前に、バッチが精度が0でも1でもないサンプルで完全に満たされるまでサンプリングを続ける。

$$JDAPO(\theta) = E(q, a) \quad D_{\{oi\}} G \quad i=1 \quad \pi_{\theta old}(\cdot|q) \quad |b_i|$$

$$\frac{1}{G} \sum_{i=1}^G \min_{r_i, t(\theta)} A^{i, t, \text{クリップ}} r_i, t(\theta), 1 - \epsilon_{low}, 1 + \epsilon_{high} \quad A_{\text{それ}} \quad (11)$$

st  $0 < \{oi \mid \text{is\_equivalent}(a, oi)\} < G$  です。

この戦略は必ずしも訓練効率を阻害するわけではないことに注意すべきである。なぜなら、RLシステムが同期され、生成段階が



(a)アクターモデルの生成確率のエントロピー。

(b)アクターモデル生成応答の平均長さ

図 4アクターモデルの確率分布のエントロピーと応答の長さの変化。

パイプライン化されていません。さらに、図6に示すように、動的サンプリングを使用すると、実験では同じパフォーマンスがより速く達成されることがわかりました。

### 3.3 リバランス行為：トークンレベルのポリシー勾配損失

オリジナルのGRPOアルゴリズムは、サンプルレベルの損失計算を採用しています。これは、まず各サンプル内のトークンごとに損失を平均化し、次にサンプル全体の損失を集計するというものです。このアプローチでは、最終的な損失計算において各サンプルに等しい重みが割り当てられます。しかし、この損失削減方法は、長時間CoT強化学習シナリオにおいていくつかの課題をもたらすことが分かっています。

損失計算ではすべてのサンプルに同じ重みが割り当てられるため、長いレスポンス（トークン数が多い）内のトークンは、全体の損失への寄与が釣り合いに低くなる可能性があります。これは2つの悪影響につながる可能性があります。まず、高品質の長いサンプルの場合、この影響により、モデルがサンプル内の推論関連パターンを学習する能力が阻害される可能性があります。次に、過度に長いサンプルは、意味不明な単語や繰り返し単語などの低品質なパターンを示すことがしばしばあります。したがって、サンプルレベルの損失計算では、長いサンプル内のこれらの望ましくないパターンを効果的にペナルティとして適用できないため、図4aと図4bに示すように、エントロピーとレスポンス長の不健全な増加につながります。

上記の制限に対処するために、long-CoT RL シナリオにトークンレベルのポリシー勾配損失を導入します。

$$J_{DAPO}(\theta) = E(q, a) \sum_{i=1}^G \frac{1}{G} \pi_{\theta}^{\text{old}}(a_i | q) \log \pi_{\theta}(a_i | q) \quad (12)$$

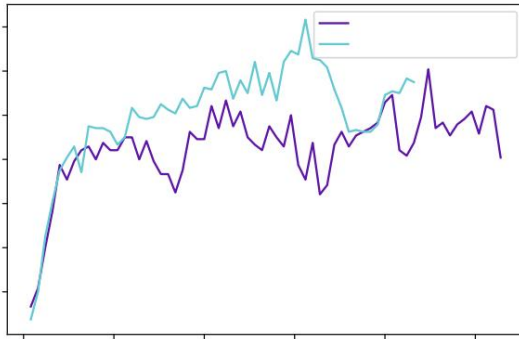
st  $0 < \{oi \mid is\_equivalent(a, oi)\} < G$  です。

この設定では、長いシーケンスは短いシーケンスよりも全体的な勾配更新に大きな影響を与える可能性があります。さらに、個々のトークンの観点から見ると、特定の生成パターンが報酬の増加または減少につながる場合、それが現れる応答の長さに関わらず、同様に促進または抑制されます。

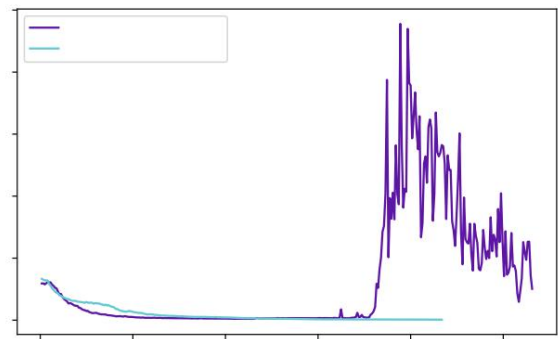
### 3.4 かくれんぼ：長すぎる報酬形成

RL トレーニングでは通常、生成の最大長を設定し、長すぎるサンプルはそれに応じて切り捨てられます。切り捨てられたサンプルに対する報酬の形成が不適切だと、報酬ノイズが発生し、トレーニング プロセスが大幅に中断される可能性があることがわかりました。





(a) AIMEでのパフォーマンス。



(b) アクターモデルのエントロピー。

図 5 Overlong Reward Shaping戦略を適用する前と適用した後の AIME 上のアクター モデルの精度とその生成確率のエントロピー。

デフォルトでは、切り捨てられたサンプルには懲罰的な報酬を割り当てます。このアプローチは、健全な推論プロセスであっても、その長さが長すぎるという理由だけでペナルティが課せられる可能性があるため、学習プロセスにノイズをもたらす可能性があります。このようなペナルティは、モデルが推論プロセスの妥当性について混乱を引き起こす可能性があります。

この報酬ノイズの影響を調査するために、まず、切り捨てられたサンプルの損失をマスクするOverlong Filtering戦略を適用しました。図5に示すように、このアプローチによりトレーニングが大幅に安定化し、パフォーマンスが向上することがわかりました。

---

#### アルゴリズム1 DAPO:分離クリップと動的サンプリングポリシーの最適化

---

入力初期ポリシーモデル $\pi_{\theta}$ ;報酬モデル $R$ ;タスクプロンプト $D$ ;ハイパーパラメータ $\epsilon_{\text{low}}$ ,  $\epsilon_{\text{high}}$  1:ステップ=1, ...,  $M$ を実行する2:  $D$ からバッチ $D_b$ をサン

プリングする3: 古いポリシーモデル $\pi_{\theta_{\text{old}}}$

←  $\pi_{\theta}$ を更新する4:  $G$ の出力 $\{o_i\}$ をサンプリングする

5: 報酬を計算する $\{r_i\}$  6:  $\pi_{\theta_{\text{old}}}(\cdot|q)$  の各質問  $q \in D_b$  の各サンプル  
出力 $o_i$ について  $R$  を実行して

$o_i$ をフィルタリングし、残りを動的サンプリングバッファに追加します (動的サンプリング方程式 (11)) バッファサイズ $n_b < N$ の場合 続行

7:

8:

9: バッファ内の各 $o_i$ について、 $A_{10}$  を計算する: 反復= 1, ...,  $\mu$ について それ  $o_i$ の $t$ 番目のトークンについて (式 (9))

て 11: を実行する:

DAPO目的関数を最大化するように政策モデル $\pi_{\theta}$ を更新する (式 (8))

出力 $\pi_{\theta}$

---

さらに、我々はSoft Overlong Punishment (式13)を提案する。これは、長さを考慮したペナルティメカニズムであり、切り捨てられたサンプルに対する報酬を形成するために設計されている。具体的には、応答の長さが事前に定義された最大値を超えた場合、ペナルティ区間を定義する。この区間内では、応答が長いほど、より大きなペナルティが与えられる。このペナルティは、元のルールベースの正解報酬に加算され、モデルに過度に長い応答を回避するよう指示する。

$$R_{\text{長さ}}(y) = \begin{cases} 0, & |y| \leq L_{\text{max}} - L_{\text{cache}} \\ \frac{(L_{\text{max}} - L_{\text{cache}}) - |y|}{L_{\text{キャッシュ}}}, & L_{\text{max}} - L_{\text{cache}} < |y| \leq L_{\text{max}} \\ -1, & L_{\text{max}} < |y| \end{cases} \quad (13)$$

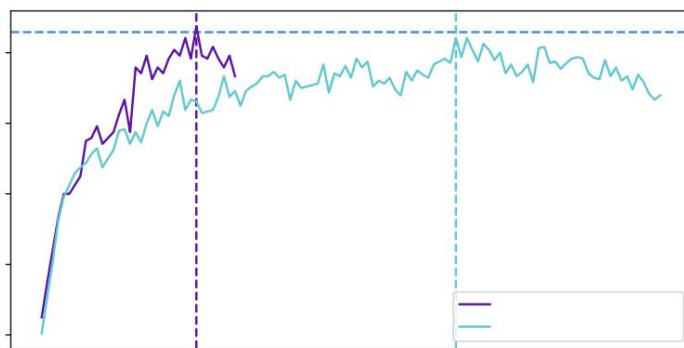


図 6 ベースライン設定で動的サンプリングを適用する前と後のトレーニングの進行状況。

### 3.5 データセットの変換

私たちのデータセットは、AoPS1ウェブサイトと公式コンテストホームページからウェブスクレイピングと手動アノテーションの組み合わせによって取得されました。数学データセットの回答は通常、式、公式、数値などさまざまな形式で提供されるため、それらを解析するための包括的なルールを設計することが困難です。ルールを使用して正確な報酬信号を提供し、数式パーサーによって導入されるエラーを最小限に抑えるために、AIME にヒントを得て、回答を選択して解析しやすい整数に変換します。たとえば、元の回答が  $a + \sqrt{b}$  の形式で表現されている場合、LLM に質問を変更するように指示して、期待される回答が  $a + b + c$  になるようにします。選択と変換の後、17K のプロンプトとそれぞれが回答として整数とペアになっている DAPO-Math-17K データセットを取得しました。

## 4つの実験

### 4.1 トレーニングの詳細

本研究では、他のタスクにも容易に応用可能なアルゴリズムを評価するために、特に数学的タスクに焦点を当てる。訓練には Ver1 フレームワーク [20] を採用する。ベースラインアルゴリズムとして ナイブ GRPO [38] を使用し、グループ報酬正規化を用いて優位性を推定する。

ハイパーパラメータについては、20 回のロールアウトステップにわたる線形ウォームアップを組み込んだ、一定の学習率  $1 \times 10^{-6}$  の AdamW、[39] オプティマイザーを使用します。ロールアウトでは、プロンプトのバッチサイズは 512 で、プロンプトごとに 16 の応答をサンプリングします。トレーニングでは、ミニバッチサイズは 512 に設定され、つまり、ロールアウトステップごとに 16 回の勾配更新が行われます。Overlong Reward Shaping では、予想される最大長を 16,384 トークンに設定し、追加の 4,096 トークンをソフト パニッシュ キャッシュとして割り当てます。したがって、生成用のトークンの最大数は 20,480 トークンに設定されます。Clip-Higher メカニズムについては、クリッピングパラメータ  $\epsilon_{\text{low}}$  を 0.2、 $\epsilon_{\text{high}}$  を 0.28 に設定し、探索と活用のトレードオフを効果的にバランスさせます。AIME での評価では、評価セットを 32 回繰り返し、結果の安定性について  $\text{avg@32}$  を報告します。評価の推論ハイパーパラメータは、温度 1.0、 $\text{topp}$  0.7 に設定されています。

### 4.2 主な結果

AIME 2024 を用いた実験では、DAPO が Qwen-32B ベースモデルを強力な推論モデルへと学習し、DeepSeek による R1 アプローチを用いた Qwen2.5-32B の実験を上回る性能を達成したことが実証されました。図 1 では、AIME 2024 において性能が大幅に向上し、精度がほぼ 0% から 50% に向上していることがわかります。注目すべきは、この性能向上が DeepSeek-R1-Zero-Qwen-32B で必要な学習ステップのわずか 50% で達成されていることです。

<sup>1</sup><https://artofproblemsolve.com/>



表1 DAPOに適用された進歩的な技術の主な結果

モデル	AIME24平均@32
ディープシーク-R1-ゼロ-Qwen-32B	47
ナীবGRPO	30
+ フィルタリングが長すぎる	36
+ クリップハイ	38
+ ソフトな長すぎる罰	41
+ トークンレベルの損失	42
+ ダイナミックサンプリング (DAPO)	50

表1に詳述されているように、私たちの方法論における各トレーニング手法の貢献を分析します。観測された改善は、これらの手法が RL トレーニングに有効であることを示しています。それぞれが AIME 2024 で数ポイントの精度向上に貢献しています。特に、バニラ GRPO 設定では、Qwen2.5-32B ベースモデルからのトレーニングでは30%の精度しか達成できません。

トークンレベルの損失の場合、パフォーマンスの向上は少なくなりますが、トレーニングの安定性が向上し、長さがより健全に増加することがわかります。

ダイナミックサンプリングを適用すると、ゼロ勾配データのフィルタリングによりより多くのデータをサンプリングする必要がありますが、全体的な学習時間には大きな影響はありません。図6に示すように、サンプリングインスタンス数が増加しても、必要な学習ステップ数が少なくなるため、モデルの収束時間はさらに短縮されます。

#### 4.3 トレーニングダイナミクス

大規模言語モデルにおける強化学習は、最先端の研究分野であるだけでなく、様々なサブシステムの相互依存性を特徴とする、本質的に複雑なシステムエンジニアリングの課題でもあります。単一のサブシステムへの変更はシステム全体に波及し、これらのコンポーネント間の複雑な相互作用により予期せぬ結果をもたらす可能性があります。データやハイパーパラメータの変動など、初期条件における一見些細な変化でさえ、反復的な強化学習プロセスを通じて増幅され、結果に大きな偏差をもたらす可能性があります。この複雑さは、研究者をしばしばジレンマに陥れます。綿密な分析を行い、変更によって学習プロセスの特定の側面が改善されるという十分な根拠のある期待を抱いたとしても、実際の結果は期待された軌跡からしばしば乖離するのです。

したがって、実験中の主要な中間結果を監視することは、矛盾の原因を迅速に特定し、最終的にシステムを改良するために不可欠です。

- 図7aに示すように、生成された応答の長さは、トレーニングの安定性とパフォーマンスに密接に関連する指標です。長さが長くなると、モデルの探索空間が広がり、より複雑な推論行動をサンプリングし、トレーニングを通じて徐々に強化することが可能になります。しかし、長さは訓練中に常に継続的に上昇するわけではないことに注意することが重要です。かなりの期間、停滞傾向や低下傾向を示すことがあり、これは[2]でも実証されています。実験の劣化を評価する指標として、長さと検証精度を併用することが一般的です。
- 図7bに示すように、訓練中の報酬のダイナミクスは、強化学習において常に重要なモニタリング指標の一つとなっています。私たちの実験の大部分において、報酬の増加傾向は比較的安定しており、実験設定の調整によって大きく変動したり減少したりすることはありません。これは、信頼性の高い報酬信号が与えられれば、言語モデルが訓練セットの分布にロバストに適合できることを示しています。しかしながら、訓練セットの最終的な報酬は検証セットの精度とほとんど相関を示さないことが多く、これは訓練セットへの過剰適合を示唆しています。
- アクターモデルのエントロピーと生成確率は、モデルの探索能力と関連しており、実験において注意深く監視する重要な指標です。直感的に言えば、モデルの

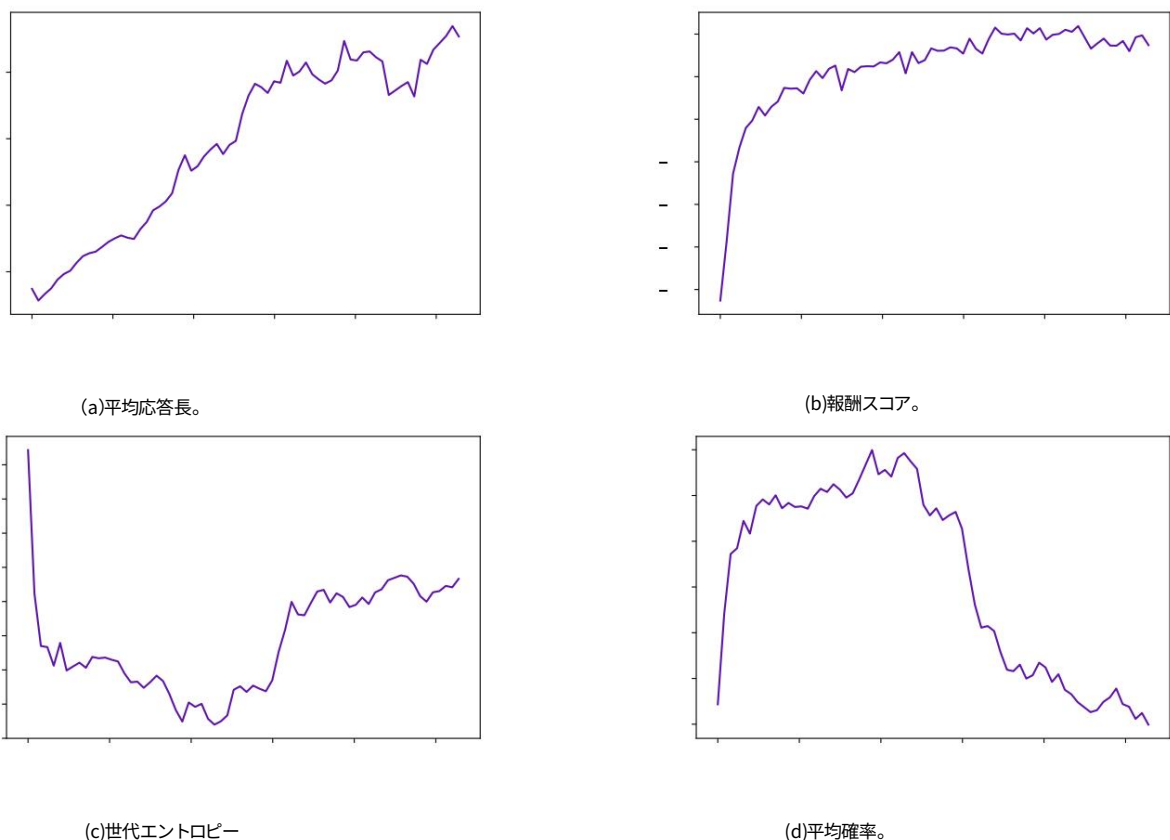


図 7 応答の長さ、報酬スコア、生成エントロピー、および DAPO の平均確率のメトリック曲線。これらは RL トレーニングのダイナミクスを示し、潜在的な問題を特定するための重要な監視指標として機能します。

エントロピーは適切な範囲内に維持される必要がある。エントロピーが過度に低い場合、確率分布が急峻になり、探索能力の低下につながる。逆に、エントロピーが過度に高い場合、意味不明な情報や繰り返し生成といった過剰探索の問題がしばしば引き起こされる。生成確率に関しては、状況は全く逆である。第3.1節で示したように、Clip-Higher戦略を適用することで、エントロピー崩壊の問題に効果的に対処できた。

その後の実験では、エントロピーの緩やかな上昇傾向を維持することがモデル パフォーマンスの向上につながることがわかりました(図 7cと図 7d に示されています)。

#### 4.4 ケーススタディ

強化学習の学習プロセスにおいて、興味深い現象が観察されました。それは、アクターモデルの推論パターンが時間の経過とともに動的に進化するというものです。具体的には、このアルゴリズムは、正しい問題解決を促進する既存の推論パターンを強化するだけでなく、当初は存在しなかった全く新しい推論モードを徐々に生み出します。この発見は、強化学習アルゴリズムの適応性と探索能力を明らかにし、モデルの学習メカニズムに関する新たな知見をもたらします。

例えば、モデルの学習初期段階では、以前の推論ステップの確認や振り返りはほとんど見られませんでした。しかし、学習が進むにつれて、モデルは表2に示すように、振り返りとバックトラックという明確な行動を示すようになります。この観察結果は、強化学習における推論能力の発現を解釈するための更なる探求に光を当てており、今後の研究課題とします。

問題 :正四面

体S — ABCの底辺ABCは正三角形であり、点Aの面SBCへの投影Hは△SBCの垂心であり、二面角H — AB — Cは30 °、 SA = 2であるとき、正四面体の体積を求めよ。答えはk + mの形である。

$\frac{k}{m}$  の値を与える

回答:与えら

れた問題を考えてみましょう: 次のことを決定する必要があります ...

...

次に、 Sの座標を(x0, y0, z0)とします。SA = 2なので、距離の公式は = 4となります。Hは = 2なので、xは次のように書きます。

$$\sqrt{(x_0)^2 + (y_0)^2 + (z_0)^2} = 2$$

...

しかし、少し待ってください。平面に関する二面角について、より思慮深い幾何学的な観点から考えてみましょう。平面α1 = ABC はxy 座標平面 ( z = 0 として)上に完全に存在します。平面α2 = SBC とします。平面α2に垂直に投影された点A はH上に着地します。

線分 l = AB ...

...

表2強化学習における反射行動の出現

5 結論

本稿では、大規模LLM強化学習 (LLM RL)のためのアルゴリズム、コードインフラストラクチャ、データセットを含む完全なオープンソースシステムを公開します。このシステムは、最先端の大規模LLM強化学習パフォーマンス ( Qwen-32B事前学習済みモデルを用いたAIME 50)を達成しています。Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO)アルゴリズムを提案し、long-CoT強化学習シナリオにおいて強化学習を強力かつ効率的にするための4つの主要手法を紹介します。さらに、学習コードとデータセットをオープンソース化することで、より広範な研究コミュニティと社会にスケーラブルな強化学習ソリューションへの実用的なアクセスを提供し、誰もがこれらの進歩の恩恵を受けることができますようになります。

## 貢献

プロジェクトリーダー

チーイン・ユー<sup>1,2,4</sup>

アルゴリズム

チーイン・ユー<sup>1,2,4</sup>、鄭張<sup>1</sup>、朱若菲<sup>1</sup>、ユフェン・ユアン<sup>1</sup>、左曉晨<sup>1</sup>、ユウユエ<sup>1</sup>

インフラストラクチャ

天天ファン<sup>1</sup>、劉高紅<sup>1</sup>、劉玲俊<sup>1</sup>、シン・リウ<sup>1</sup>、林海斌<sup>1</sup>、林志奇<sup>1</sup>、ボレ・マ<sup>1</sup>、光明盛<sup>1,3</sup>、Yuxuan Tong<sup>1,2,4</sup>、Qiyang Yu<sup>1,2,4</sup>、チ・チャン<sup>1</sup>、Mofan Zhang<sup>1</sup>、Wang Zhang<sup>1</sup>、Hang Zhu<sup>1</sup>、金華朱<sup>1</sup>

姓（アルファベット順）

データセット

陳佳澤<sup>1</sup>、陳江傑<sup>1,4</sup>、チェンイー・ワン<sup>1</sup>、ホンリー・ユ<sup>1,2,4</sup>、渭南大<sup>1,2,4</sup>、宋玉軒<sup>1,2,4</sup>、翔鵬ウェイ<sup>1</sup>、Qiyang Yu<sup>1,2,4</sup>

監督

周浩<sup>2,4</sup>、劉静静<sup>2,4</sup>、ウェイイン・マー<sup>2,4</sup>、張雅琴<sup>2,4</sup>、リン・ヤン<sup>1,4</sup>、穆喬<sup>1,4</sup>、吳永輝<sup>1</sup>、ミンシュアン・ワン<sup>1,4</sup>

所属

1ByteDanceシード

<sup>2</sup> 清華大学AI産業研究所（AIR）

3香港大学

清華AIRとByteDance Seedの4SIA-Lab

## 謝辞

DAPOプロジェクトを支援してくれたZhengyin Du、Kai Shen、Tianyang Zhan、Zhen Xiao、Renjie Zheng、Li Han、Kaihua Jiang、およびByteDance の他の同僚に感謝します。

## 参考文献

- [1] OpenAI.LLMSを使った推論の学習、2024年。
- [2] 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。

---
- [3] OpenAI.GPT4技術レポート.arXivプレプリントarXiv:2303.08774,2023。

---
- [4] アントロピック、クロード3.5ソネット、2024年。
- [5] Tom Brown,Benjamin Mann,Nick Ryder,Melanie Subbiah,Jared D Kaplan,Prafulla Dhariwal,Arvind Neelakantan,Pranav Shyam,Girish Sastry,Amanda Askell他。言語モデルは少数ショット学習器である。  
神経情報処理システムの進歩,33:1877–1901,2020年。

---
- [6] Aakanksha Chowdhery,Sharan Narang,Jacob Devlin,Maarten Bosma,Gaurav Mishra,Adam Roberts,Paul Barham,Hyung Won Chung,Charles Sutton,Sebastian Gehrmann,et al. Palm :パスウェイによる言語モデルのスケーリング。機械学習研究ジャーナル,24(240):1–113,2023年。

---
- [7] 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。

---
- [8] XAI. Grok 3ベータ版 - 推論エージェントの時代、2024年。
- [9] Google DeepMind.Gemini 2.0 フラッシュシンキング、2024年。
- [10] Qwen.Qwq-32b:強化学習の力を受け入れる、2024年。
- [11] キミチーム、アンガン・ドウ、ボフェイ・ガオ、ボウェイ・シン、チャンジウ・ジャン、チェン・チェン、チェン・リー、チェンジュン・シャオ、チェンチュアン・ドウ、チョンファ・リャオ、他。キミk1。 5: llms による強化学習のスケーリング。 arXiv プレプリント arXiv:2501.12599,2025。

---
- [12] An Yang,Baosong Yang,Beichen Zhang,Binyuan Hui,Bo Zheng,Bowen Yu,Chengyuan Li,Dayiheng Liu,Fei Huang,Haoran Wei,他。クウェン2。 5 テクニカルレポート。 arXiv プレプリント arXiv:2412.15115,2024。

---
- [13] Zhipeng Chen,Yingqian Min,Beichen Zhang,Jie Chen,Jinhao Jiang,Daixuan Cheng,Wayne Xin Zhao,Zheng Liu,Xu Miao,Yang Lu 他。 r1 に似た推論モデルの導出と改善に関する実証研究。 arXiv プレプリント arXiv:2503.04548,2025。

---
- [14] ジンチェン・フー、インミン・チャン、チー・ハン、ダシン・ジャン、フンヨン・シュム・シャンユー・チャン。 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 アプローチ。 <https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero>、 2025年。
- [15] Jian Hu. Reinforce++:大規模言語モデルをアラインメントするためのシンプルで効率的なアプローチ.arXivプレプリントarXiv:2501.03262, 2025。

---
- [16] 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。

---
- [17] Jung Hyun Lee,June Yong Yang,Byeongho Heo,Dongyoon Han,Kang Min Yoo. 大規模言語モデルの数学的推論能力を強化するためのトークン教師あり価値モデル.arXivプレプリントarXiv:2407.12863, 2024。

---
- [18] アミールホセイン・カゼムネジャド、ミラド・アガジョハリ、エヴァ・ポータランス、アレックス・ソルドニ、シヴァ・レディ、アーロン・クールヴィル、ニコラス・ル・ヴィネッポ :洗練された単位割り当てによる法文学的推論のRLポテンシャルの解放.arXivプレプリントarXiv :2410.01679,2024。

---
- [19] Yufeng Yuan,Yu Yue,Ruofei Zhu,Tiantian Fan,Lin Yan. long-cotにおけるPPOの崩壊の背後には何があるのか?その秘密は値最適化にある.arXivプレプリントarXiv:2503.01491, 2025。

---
- [20] Guangming Sheng,Chi Zhang,Zilingfeng Ye,Xibin Wu,Wang Zhang,Ru Zhang,Yanghua Peng,Haibin Lin,および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。

---
- [21] ジョン・シュルマン、フィリップ・ウォルスキ、プラフル・ダリワル、アレック・ラドフォード、オレグ・クリモフ。近似政策最適化アルゴリズム.arXivプレプリント arXiv:1707.06347, 2017。

---

- [22] ジョン・シュルマン、フィリップ・モーリッツ、セルゲイ・レヴィン、マイケル・ジョーダン、ピーター・アビール、一般化利点推定を用いた高次元連続制御、2018年。
- [23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, Ryan Lowe. 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。S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh編、『Advances in Neural Information Processing Systems』第35巻、27730～27744ページ、Curran Associates, Inc., 2022年。
- 
- [24] ダリオ・アモデイ、クリス・オラー、ジェイコブ・スタインハート、ポール・クリスティアーノ、ジョン・シュルマン、ダン・マネ。具体的な問題 AI安全性、2016年。
- [25] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, Shane Legg. 破損した報酬チャネルを用いた強化学習、2017年。
- [26] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, Shane Legg. 仕様ゲーミング : AIの創意工夫の裏側、2020年。
- [27] トム・エヴェリット、マーカス・ハッター、ラマナ・クマール、ピクトリア・クラコフナ。強化学習における報酬改ざんの問題と解決策 : 因果影響図の観点、2021年。
- [28] レオ・ガオ、ジョン・シュルマン、ジェイコブ・ヒルトン。報酬モデルの過剰最適化に関するスケーリング則、2022年。
- [29] リリアン・ウエン。強化学習における報酬ハッキング。lilianweng.github.io、2024年11月。
- [30] スタニスラス・ポールとイリヤ・スツケヴァー「自動定理証明のための生成言語モデリング」2020年。
- [31] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, Thang Luong. 人間のデモンストレーションなしでオリンピック幾何学を解く。Nature, 625(7995):476–482, 2024。
- [32] Trieu TrinhとThang Luong. Alphageometry: 幾何学のためのオリンピックレベルのAIシステム、2024年。
- [33] AlphaProofとAlphaGeometryチーム。AIが国際数学オリンピックの問題を解き、銀メダルレベルを達成、2024年。
- [34] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, Steven Chu Hong Hoi. Coder1: 事前学習済みモデルと深層強化学習によるコード生成の習得。ニューラル情報処理システムの進歩, 35:21314–21328, 2022。
- 
- [35] ノア・シン、フェデリコ・カッサーノ、エドワード・パーマン、アシュウィン・ゴピナス、カルティク・ナラシンハン、ヤオ・シュンユウ。Reflection: 言語強化学習を備えた言語エージェント、2023年。
- [36] Xinyun Chen, Maxwell Lin, Nathanael Schärli, Denny Zhou. 大規模言語モデルに自己デバッグを教える、2023年。
- [37] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, Gabriel Synnaeve, RLEF : 強化学習による実行フィードバックにおけるコードモデルのグラウンディング、2025年。
- [38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, Dada 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。
- 
- [39] イリヤ・ロシュテロフとフランク・ハッター「分離重み減衰正則化」国際学習会議表現、2019年。
-



# 付録

## データセットの変換

以下はデータ変換の例です。元の問題、変換後の問題、変換プロセス中のモデル出力を示しています (元の問題はマークダウン テキストで、読みやすいようにレンダリングされています)。

モデルがこのタスクを高精度で完了できるようにするために、私たちの目標は、幻覚を回避しながら包括的な推論を実行させることです。これは、徹底的な推論を促す明確な思考フレームワークを提供することで実現します。具体的には、思考連鎖 (CoT) 推論と、明確に定義された4つのステップ (解答形式の抽出、問題文の書き換え、修正された問題の解決、そして最終的な解答として整数値の提供) を採用しています。各ステップでは、モデルの推論を導くための、少数の試行による解や詳細なガイドラインを提供します。

このアプローチを通じて、ほとんどの場合、LLM は形式と品質の両方が満足できる再定式化を生成できることがわかりました。

元の問題xとyを、xの値がxとな  
るような実数とします。答え:  $11 - 2\sqrt{6}$   $x^2 + y^2 - 22x - 16y + 113 = 0$ 。可能な限り最小のものを求めよ。

変形問題xとyを  $x - 22x - 16y + 113 =$   
0となる実数とします。xの最小値を求めなさい。元の解は  $k - \frac{22+y}{m}\sqrt{n}$  の形です。ここで、k、 m、 nは整数です。k + m + n の値を求めなさい。答え： 19

## モデルの出力 (マークダウンテキスト)

変換の思考プロセスこの問題の元の答えは代  
数式  $11 - 2\sqrt{6}$  です。これを整数形式に変換する必要があります。まず、答えは整数部分と平方根項で構成されていることに  
注意してください。  
与えられた規則に従って、整数部分と平方根部分を未定係数に置き換えます。元の問題ではxとyが変数として使われているので、他の文字を未定係数とし  
て選びます。答えは  $k - m\sqrt{n}$  の形で表し、k + m + nを求めます。

解答形式  $k - m\sqrt{n}$  変換問  
題xとyを  $x -$   
 $22x - 16y + 113 = 0$  となる実数とします。  
xの最小値を求めなさい。元の解答は  $k - m\sqrt{n}$  の形式で、 k、 m、 nは整数です。k + m + n の値を求めなさい。

変換された問題の計算の答え  
•元の答えは、 $11 - 2\sqrt{6}$  です。•必要な形式は、 $k - m\sqrt{n}$  です。•係数を一致させると、k = 11、m = 2、  
n = 6 になります。•最終結果は、k + m + n = 11 + 2 + 6 = 19 です。

最終回答  
19

## B 補足事例

---

問題 :エメヴィ

ルの住民900人のうち、ダイヤモンドの指輪を持っている人は195人、ゴルフクラブセットを持っている人は367人、園芸用のスコップを持っている人は562人います。さらに、900人の住民はそれぞれハート型のキャンディーが入った袋を1つ持っています。これらのものをちょうど2つ持っている住民は437人、3つ持っている住民は234人です。エメヴィルの住民のうち、これら4つすべてを所有している人の数を求めなさい。

---

応答:ステッ

ブバーステップの解決方

法: 1. 次の変数を定義しましょう。A をダイヤモンドの指輪を所有している住民の集合とします。

B をゴルフクラブのセットを所有する住民の集合とします。

C を庭用のスコップを所有する住民の集合とします。

ハート型キャンディーの袋を所有する住民の集合をDとします。900人の住民全員がハート型キャンディーの袋を所有しているので、 $|D| = 900$ となります。

4つのアイテムのうち  $i$  個だけを所有している居住者の数を  $a_i$  で表します。

包含排除の原理を、A、B、C、D の 4 つの集合の和集合に適用してみましょう。

...

$$a_4 = \frac{-219}{-4} = \frac{219}{4} = 54.75。$$

4つのアイテムすべてを所有している人の数である  $a_4$  は整数でなければならないため、アイテムの異なる組み合わせを所有している人の組み合わせカウントという観点から現在の **アプローチを再検討する必要があります。**

包含排除原理を直接使用する代わりに、アイテム A、B、C、D の組み合わせを所有している人の数を考慮する別のアプローチを使用できます。

A、B、C のうち  $i$  番目のアイテムを所有している人の数を  $x_i$  とします(全員が所有しているアイテム D は除く)。

すると、A、B、C、D のうち、ちょうど  $k$  個のアイテムを所有している人の総数は次のように分類されます。

...

---

表3反射行動の出現事例