

## DeepSeek-R1: LLMにおける推論能力の奨励 強化学習

ディープシークAI

お問い合わせ

抽象的な

第一世代の推論モデルである DeepSeek-R1-Zero と DeepSeek-R1 を紹介します。  
DeepSeek-R1-Zero は、予備段階として教師あり微調整 (SFT)を行わずに大規模強化学習 (RL) でトレーニングされたモデルであり、優れた推論能力を発揮します。  
RL を通じて、DeepSeek-R1-Zero は、数多くの強力に興味深い推論動作を自然に実現します。ただし、読みにくさや言語の混在などの課題に直面します。これらの問題に対処し、推論パフォーマンスをさらに向上させるために、RL の前に多段階のトレーニングとコールド スタート データを組み込んだ DeepSeek-R1 を導入します。  
DeepSeek-R1 は、推論タスクで OpenAI-o1-1217 に匹敵するパフォーマンスを実現します。研究コミュニティをサポートするために、DeepSeek-R1-Zero、DeepSeek-R1、および Qwen と Llama に基づく DeepSeek-R1 から抽出された 6 つの高密度モデル(1.5B、7B、8B、14B、32B、70B) をオープンソース化します。

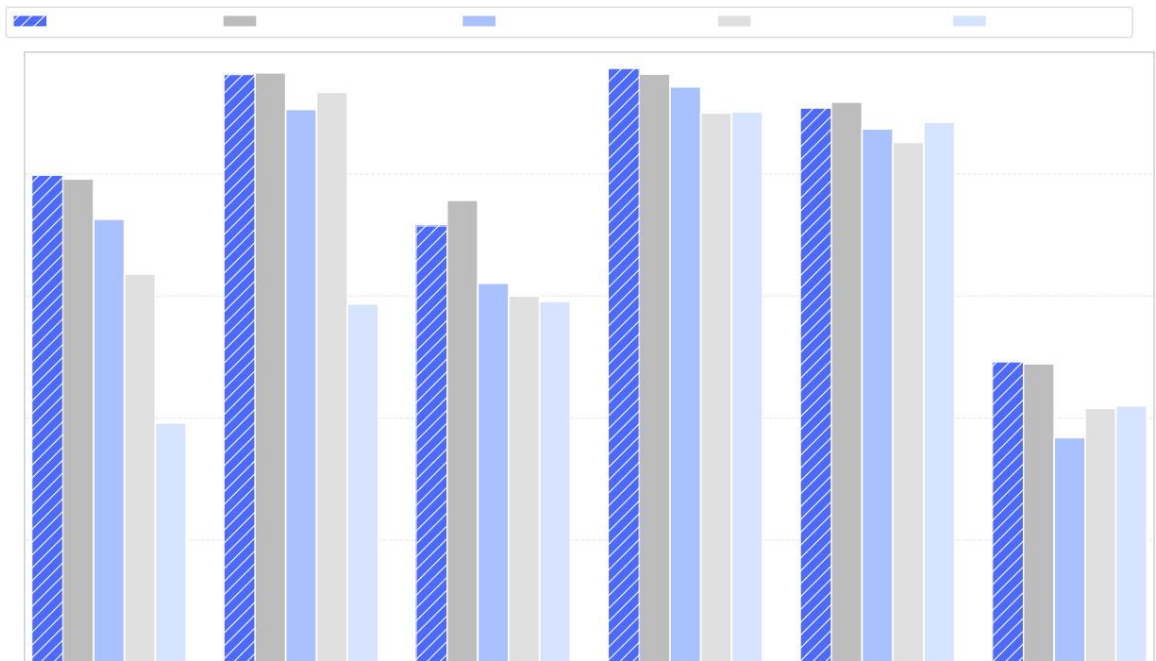


図 1 | DeepSeek-R1 のベンチマーク パフォーマンス。

## コンテンツ

|   |    |
|---|----|
| 1 はじめに  | 3  |
| 1.1 貢献 ..   | 4  |
| 1.2 評価結果の要約 ...                                   | 4  |
| 2 アプローチ   | 5  |
| 2.1 概要 .  | 5  |
| 2.2 DeepSeek-R1-Zero: ベースモデルでの強化学習。               | 5  |
| 2.2.1 強化学習アルゴリズム                                  | 5  |
| 2.2.2 報酬モデリング .                                   | 6  |
| 2.2.3 トレーニング テンプレート .                             | 6  |
| 2.2.4 DeepSeek-R1-Zero 6のパフォーマンス、自己進化プロセス、そしてアハ体験 |    |
| 2.3 DeepSeek-R1: コールドスタートによる強化学習。                 | 9  |
| 2.3.1 コールドスタート。                                   | 9  |
| 2.3.2 推論指向強化学習 ..                                 | 10 |
| 2.3.3 拒否サンプリングと教師あり微調整。                           | 10 |
| 2.3.4 あらゆるシナリオに対する強化学習。                           | 11 |
| 2.4 蒸留: 小さなモデルに推論機能を追加する。                         | 11 |
| 3 実験3.1 DeepSeek-R1の                              | 11 |
| 評価 ..   | 13 |
| 3.2 蒸留モデルの評価。                                     | 14 |
| 4 議論  | 14 |
| 4.1 蒸留学習と強化学習 ....                                | 14 |
| 4.2 失敗した試み...                                     | 15 |
| 5 結論、限界、今後の課題                                     | 16 |
| 貢献と謝辞   | 20 |

## 1. はじめに

近年、大規模言語モデル (LLM) は急速な反復と進化を遂げており (Anthropic, 2024 年、Google, 2024 年、OpenAI, 2024a)、汎用人工知能 (AGI) とのギャップは徐々に縮小しています。

最近、トレーニング後のトレーニングは、完全なトレーニング パイプラインの重要な要素として浮上しています。この手法は、事前トレーニングに比べて比較的最小限の計算リソースで、推論タスクの精度を高め、社会的価値観に合わせ、ユーザーの好みに適応することが示されています。推論機能の観点では、OpenAI の o1 (OpenAI, 2024b) シリーズ モデルが、Chain-of-Thought 推論プロセスの長さを増やすことで推論時間のスケーリングを導入した最初のモデルです。このアプローチにより、数学、コーディング、科学的推論など、さまざまな推論タスクで大幅な改善が達成されています。ただし、効果的なテスト時間のスケーリングの課題は、研究コミュニティにとって未解決の問題のまです。これまでの研究では、プロセスベースの報酬モデル (Lightman et al., 2023; Uesato et al., 2022; Wang et al., 2023)、強化学習 (Kumar et al., 2024)、モンテカルロツリー探索やビーム探索などの探索アルゴリズム (Feng et al., 2024; Trinh et al., 2024; Xin et al., 2024) など、さまざまなアプローチが検討されてきました。ただし、これらの方法のいずれも、OpenAI の o1 シリーズモデルに匹敵する一般的な推論パフォーマンスを達成していません。

本稿では、純粋な強化学習 (RL) を使用して言語モデルの推論機能を向上させるための第一歩を踏み出します。私たちの目標は、純粋な RL プロセスによる自己進化に焦点を当て、教師ありデータなしで推論機能を開発する LLM の可能性を探ることです。具体的には、DeepSeek-V3-Base を基本モデルとして使用し、GRPO (Shao et al., 2024) を RL フレームワークとして採用して、推論におけるモデルのパフォーマンスを向上させます。

トレーニング中、DeepSeek-R1-Zero は、数多くの強力な興味深い推論動作を自然に実現しました。数千の RL ステップを経て、DeepSeek-R1-Zero は推論ベンチマークで優れたパフォーマンスを発揮します。たとえば、AIME 2024 の pass@1 スコアは 15.6% から 71.0% に増加し、多数決によりスコアはさらに 86.7% に向上し、OpenAI-o1-0912 のパフォーマンスと一致します。

しかし、DeepSeek-R1-Zero では、読みにくさや言語の混在などの課題があります。これらの問題に対処し、推論パフォーマンスをさらに向上させるために、少量のコールド スタート データと多段階のトレーニング パイプラインを組み込んだ DeepSeek-R1 を導入しました。具体的には、まず何千ものコールド スタート データを収集して、DeepSeek-V3-Base モデルを微調整します。その後、DeepSeek-R1-Zero のような推論指向の RL を実行します。RL プロセスが収束に近づく、RL チェックポイントでの拒否サンプリングを通じて新しい SFT データを作成し、ライティング、事実に基づく QA、自己認識などのドメインにおける DeepSeek-V3 の教師ありデータと組み合わせ、DeepSeek-V3-Base モデルを再トレーニングします。

新しいデータで微調整した後、チェックポイントは、すべてのシナリオからのプロンプトを考慮しながら追加の RL プロセスを経ます。これらの手順の後、OpenAI-o1-1217 と同等のパフォーマンスを実現する DeepSeek-R1 と呼ばれるチェックポイントを取得しました。

さらに、DeepSeek-R1 からより小さな密なモデルへの蒸留についても調査しました。Qwen2.5-32B (Qwen, 2024b) をベースモデルとして使用し、DeepSeek-R1 からの直接蒸留は、これに RL を適用するよりも優れたパフォーマンスを発揮しました。これは、より大きなベースモデルによって発見された推論パターンが推論機能の向上に不可欠であることを示しています。蒸留された Qwen および Llama (Dubey et al., 2024) シリーズをオープンソース化しました。特に、蒸留された 14B モデルは最先端のオープンソース QwQ-32B-Preview (Qwen, 2024a) を大幅に上回り、蒸留された 32B および 70B モデルは密なモデル間の推論ベンチマークで新記録を樹立しました。

### 1.1. 貢献

#### トレーニング後: ベースモデルでの大規模強化学習

- 予備ステップとして教師あり微調整 (SFT) に頼らずに、ベースモデルに RL を直接適用します。このアプローチにより、モデルは複雑な問題を解決するための思考の連鎖 (CoT) を探索できるようになり、DeepSeek-R1-Zero が開発されました。DeepSeek-R1-Zero は、自己検証、リフレクション、長い CoT の生成などの機能を実証しており、研究コミュニティにとって重要なマイルストーンとなっています。特に、これは **SFT を必要とせずに、LLM の推論機能を RL のみでインセンティブ化できることを検証した最初のオープンリサーチです**。このブレークスルーは、この分野での将来の進歩への道を開きます。
- DeepSeek-R1 を開発するためのパイプラインを紹介します。このパイプラインには、改善された推論パターンを発見し、**人間の好みに合わせることを目的とした2つの RL ステージと、モデルの推論機能と非推論機能の種となる2つの SFT ステージ**が組み込まれています。このパイプラインは、より優れたモデルを作成することで業界に利益をもたらすと考えています。

#### 蒸留: 小さなモデルでも強力になる

- **大規模モデルの推論パターンを小規模モデルに抽出**することで、小規模モデルでの強化学習で発見された推論パターンに比べてパフォーマンスが向上することを実証しました。オープンソースの DeepSeek-R1 とその API は、将来的に研究コミュニティがより優れた小規模モデルを抽出するのに役立つでしょう。
- DeepSeek-R1 によって生成された推論データを使用して、研究コミュニティで広く使用されているいくつかの高密度モデルを微調整しました。評価結果では、蒸留された小さな高密度モデルがベンチマークで非常に優れたパフォーマンスを発揮することが示されています。DeepSeek-R1-Distill-Qwen-7B は AIME 2024 で 55.5% を達成し、QwQ-32B-Preview を上回りました。さらに、DeepSeek-R1-Distill-Qwen-32B は AIME 2024 で 72.6%、MATH-500 で 94.3%、LiveCodeBench で 57.2% のスコアを獲得しました。これらの結果は、以前のオープンソースモデルを大幅に上回り、o1-mini に匹敵します。私たちは、Qwen2.5 および Llama3 シリーズに基づいて、15 億、70 億、80 億、14 億、32 億、および 70 億のチェックポイントを抽出し、コミュニティにオープンソース化しました。

### 1.2. 評価結果の要約

- 推論タスク: (1) DeepSeek-R1は**AIME 2024**で79.8% Pass@1のスコアを達成し、OpenAI-o1-1217をわずかに上回りました。**MATH-500**では97.3%という素晴らしいスコアを達成し、OpenAI-o1-1217と同等のパフォーマンスを発揮し、他のモデルを大幅に上回りました。(2) コーディング関連のタスクでは、DeepSeek-R1 はコード競技タスクでエキスパート レベルを示し、Codeforces で 2,029 Elo レーティングを達成し、競技に参加した人間の 96.3% を上回りました。エンジニアリング関連のタスクでは、DeepSeek-R1 はDeepSeek-V3 よりもわずかに優れたパフォーマンスを発揮し、実際のタスクで開発者の役に立つ可能性があります。
- 知識: MMLU、MMLU-Pro、GPQA Diamond などのベンチマークでは、DeepSeek-R1 は優れた結果を達成し、MMLU で 90.8%、MMLU-Pro で 84.0%、GPQA Diamond で 71.5% のスコアで DeepSeek-V3 を大幅に上回りました。これらのベンチマークでは、DeepSeek-R1 のパフォーマンスは OpenAI-o1-1217 よりわずかに劣りますが、他のクローズドソースモデルを上回り、教育タスクにおける競争力を示しています。事実のベンチマークである SimpleQA では、DeepSeek-R1 は DeepSeek-V3 を上回り、事実に基づくクエリを処理する能力を示しています。同様の傾向が見られ、OpenAI-o1 はこのベンチマークで 4o を上回りました。

- その他: DeepSeek-R1は、クリエイティブライティング、一般的な質問への回答、編集、要約など。印象的な AlpacaEval 2.0 では長さ制御の勝率が 87.6%、ArenaHard では勝率が 92.3% となり、試験以外のクエリをインテリジェントに処理する優れた能力が示されました。さらに、DeepSeek-R1は、次のようなタスクで優れたパフォーマンスを発揮します。ロングコンテキスト理解はDeepSeek-V3を大幅に上回るベンチマーク。

## 2. アプローチ

### 2.1. 概要

これまでの研究では、モデルを強化するために大量の教師ありデータに大きく依存してきた。パフォーマンス。この研究では、推論能力が教師あり学習を使わなくても、大規模な強化学習（RL）によって改善された。コールドスタートとして微調整（SFT）を行います。さらに、パフォーマンスをさらに向上させるには、少量のコールドスタートデータを含める。以下のセクションでは、(1) DeepSeek-R1-Zeroは、SFTデータなしでベースモデルに直接RLを適用し、(2) DeepSeek-R1は、数千のチェックポイントからRLを適用し、長い思考連鎖（CoT）の例。3) DeepSeek-R1の推論機能を抽出して小型で高密度なモデル。

### 2.2. DeepSeek-R1-Zero: ベースモデルでの強化学習

強化学習は、これまでの研究（Shao et al., 2024; Wang et al., 2023）で証明されているように、推論タスクにおいて大きな有効性を示しています。しかし、これらの研究は教師ありデータに大きく依存しており、その収集には時間がかかります。このセクションでは、教師なしデータなしで推論能力を開発するLLMの可能性を探る。純粋な強化学習プロセスを通じて自己進化に焦点を当てています。RLアルゴリズムの簡単な概要と、いくつかの興味深い結果の発表、そしてこれがコミュニティに貴重な洞察を提供することを願っています。

#### 2.2.1. 強化学習アルゴリズム

グループ相対ポリシー最適化強化学習のトレーニングコストを節約するために、グループ相対ポリシー最適化を採用する。相対的政策最適化（GRPO）（Shao et al., 2024）は、批評家モデルを放棄し、通常はポリシーモデルと同じサイズであり、代わりにグループスコアからベースラインを推定します。具体的には、各質問に対して、GRPOは古いものから出力のグループ $\{1, 2, \dots\}$ をサンプリングします。ポリシーを作成し、次の目標を最大化することでポリシーモデルを最適化します。

$$J(\pi) = E \left[ \sum_{i=1}^n \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{r_g(\pi)}{r_g(\pi_{\text{clip}})} \right] \quad (1)$$

$$r_g(\pi) = \frac{r_g(\pi)}{r_g(\pi_{\text{clip}})} - \log \frac{r_g(\pi)}{1, (|\mathcal{G}|)} \quad (2)$$

ここで、 $\pi$  および  $\pi_{\text{clip}}$  はハイパーパラメータであり、 $\pi$  および  $\pi_{\text{clip}}$  は、各グループ内の出力に対応する報酬 $\{1, 2, \dots, |\mathcal{G}|\}$ 。

$$r_g(\pi) = \frac{-m(\{1, 2, \dots, |\mathcal{G}|\})}{(|\mathcal{G}|, \{1, 2, \dots, |\mathcal{G}|\})} \quad (3)$$

---

ユーザーとアシスタント間の会話。ユーザーが質問し、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、それからユーザーに答えを提供します。推論プロセスと答えはそれぞれ `<think>` `</think>` タグと `<answer>` `</answer>` タグで囲まれます。つまり、`<think>` 推論プロセスはここに `</think>` `<answer>` 答えはここに `</answer>` となります。ユーザー: **プロンプト**。アシスタント:

---

表 1 | DeepSeek-R1-Zero のテンプレート。**プロンプト**は、トレーニング中に特定の推論質問に置き換えられます。

### 2.2.2. 報酬モデリング

報酬はトレーニング信号のソースであり、RL の最適化方向を決定します。

DeepSeek-R1-Zero をトレーニングするために、主に 2種類の報酬で構成されるルールベースの報酬システムを採用しています。

- **精度報酬**: 精度報酬モデルは、応答が正しいかどうかを評価します。

たとえば、決定論的な結果を伴う数学の問題の場合、モデルは最終解答を指定された形式 (ボックス内など) で提供する必要があります。これにより、信頼性の高いルールベースの正しさの検証が可能になります。同様に、LeetCode の問題の場合、コンパイラを使用して、定義済みのテスト ケースに基づいてフィードバックを生成できます。形式報酬: 精度報酬モデルに加えて、モデルが思考プロセスを `'<think>'` タグと `'</think>'` タグの間に配置するように強制する形式報酬モデルを採用しています。

DeepSeek-R1-Zero の開発では、結果またはプロセスのニューラル報酬モデルは適用しません。これは、ニューラル報酬モデルが大規模な強化学習プロセスで報酬ハッキングの影響を受ける可能性があり、報酬モデルの再トレーニングには追加のトレーニングリソースが必要になり、トレーニング パイプライン全体が複雑になることがわかったためです。

### 2.2.3. トレーニングテンプレート

DeepSeek-R1-Zero をトレーニングするには、まず、ベース モデルが指定した指示に従うようにガイドする簡単なテンプレートを設計します。表 1 に示すように、このテンプレートでは、DeepSeek-R1-Zero が最初に推論プロセスを生成し、その後に最終的な答えを生成する必要があります。

私たちは意図的にこの構造形式に制約を限定し、コンテンツ固有のバイアス (反省的推論の義務付けや特定の問題解決戦略の促進など) を回避して、RL プロセス中にモデルの自然な進行を正確に観察できるようにします。

### 2.2.4. DeepSeek-R1-Zero のパフォーマンス、自己進化プロセス、そしてアハ体験

DeepSeek-R1-Zero のパフォーマンス図 2 は、RL トレーニング プロセス全体にわたる AIME 2024 ベンチマークでの DeepSeek-R1-Zero のパフォーマンスの軌跡を示しています。図に示すように、DeepSeek-R1-Zero は、RL トレーニングが進むにつれて、着実かつ一貫してパフォーマンスが向上しています。特に、AIME 2024 の平均 pass@1 スコアは大幅に増加し、最初の 15.6% から驚異的な 71.0% に跳ね上がり、OpenAI-o1-0912 に匹敵するパフォーマンス レベルに達しています。この大幅な改善は、時間の経過とともにモデルのパフォーマンスを最適化する RL アルゴリズムの有効性を強調しています。

表2は、DeepSeek-R1-ZeroとOpenAIのo1-0912モデルを、さまざまな推論関連のベンチマークで比較分析したものです。その結果、RLは



| モデル                   | エイム2024 |       | 数学500 | GPQA ライブコード<br>ダイヤモンドベンチ |      | コードフォース |
|-----------------------|---------|-------|-------|--------------------------|------|---------|
|                       | パス@1    | 欠点@64 | パス@1  | パス@1                     | パス@1 | 評価      |
| OpenAI-o1-mini        | 63.6    | 80.0  | 90.0  | 60.0                     | 53.8 | 1820    |
| OpenAI-o1-0912        | 74.4    | 83.3  | 94.8  | 77.3                     | 63.4 | 1843    |
| DeepSeek-R1-Zero 71.0 |         | 86.7  | 95.9  | 73.3                     | 50.0 | 1444    |

表2 | DeepSeek-R1-ZeroとOpenAI o1モデルの推論関連の比較  
ベンチマーク。

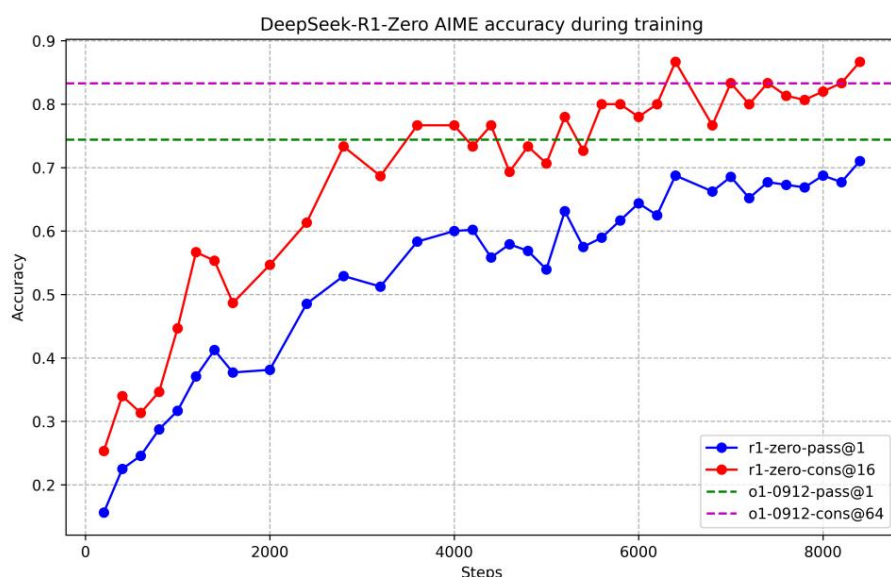


図2 | トレーニング中のDeepSeek-R1-ZeroのAIME精度。各質問について、サンプルを抽出します。  
16 個の回答から全体の平均精度を計算し、安定した評価を確保します。

DeepSeek-R1-Zeroは、教師なしでも堅牢な推論能力を実現します。  
データの微調整。これは注目すべき成果であり、モデルの能力を強調している。  
RLのみで効果的に学習し、一般化することができます。さらに、DeepSeek-R1-Zeroのパフォーマンスは、多数決を適用することでさらに強化されます。たとえば、  
AIMEベンチマークで多数決を採用した場合、DeepSeek-R1-Zeroのパフォーマンスは  
71.0%から86.7%に上昇し、OpenAI-o1-0912の性能を上回りました。  
DeepSeek-R1-Zeroは、  
多数決投票は、その強力な基盤能力とさらなる発展の可能性を強調している。  
推論タスクの進歩。

DeepSeek-R1-Zeroの自己進化プロセス  
RLがモデルを駆動して推論能力を向上させる方法を示す魅力的なデモンストレーションです。  
自律的に。ベースモデルから直接RLを開始することで、モデルの  
監督された微調整段階の影響を受けずに進行する。このアプローチは  
モデルが時間の経過とともにどのように進化するかを明確に把握する。特に、モデルが処理する能力の面で  
複雑な推論タスク。

図3に示すように、DeepSeek-R1-Zeroの思考時間は一貫して改善している。

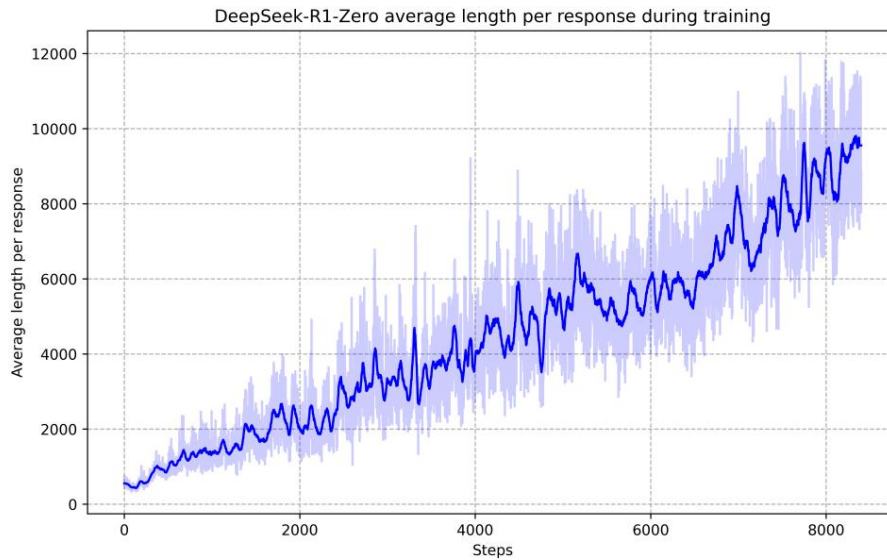


図 3 | RL プロセス中のトレーニング セットにおける DeepSeek-R1-Zero の平均応答長。DeepSeek-R1-Zero は、より長い思考時間で推論タスクを解決することを自然に学習します。

トレーニング プロセス全体を通じて、この改善は外部調整の結果ではなく、モデル内部の本質的な発展によるものです。DeepSeek-R1-Zero は、テスト時間の延長された計算を活用することで、ますます複雑化する推論タスクを解決する能力を自然に獲得します。この計算は、数百から数千の推論トークンを生成することから始まり、モデルが思考プロセスをより深く探求し、改良することを可能にします。

この自己進化の最も注目すべき側面の 1 つは、テスト時間の計算が増加するにつれて、洗練された動作が出現することです。モデルが以前の手順を再訪して再評価するリフレクションや、問題解決への代替アプローチの探索などの動作は、自発的に発生します。これらの動作は明示的にプログラムされているわけではなく、モデルと強化学習環境との相互作用の結果として発生します。この自発的な開発により、DeepSeek-R1-Zero の推論機能が大幅に強化され、より効率的かつ正確に、より困難なタスクに取り組むことができるようになります。

DeepSeek-R1-Zero の「なるほど」という瞬間 DeepSeek-R1-Zero のトレーニング中に観察される特に興味深い現象は、「なるほど」という瞬間の発生です。表 3 に示すように、この瞬間はモデルの中間バージョンで発生します。このフェーズでは、DeepSeek-R1-Zero は、最初のアプローチを再評価することで、問題により多くの思考時間を割り当てることを学習します。この動作は、モデルの推論能力が向上していることの証であるだけでなく、強化学習が予想しない高度な結果につながる可能性があることを示す魅力的な例でもあります。

この瞬間は、モデルにとってだけでなく、その動作を観察している研究者にとっても「ひらめきの瞬間」です。これは強化学習の力と美しさを強調しています。モデルに問題の解決方法を明示的に教えるのではなく、適切なインセンティブを与えるだけで、モデルは高度な問題解決戦略を自律的に開発します。この「ひらめきの瞬間」は、強化学習が人工システムの新しいレベルの知能を解放する可能性を強力に思い出させ、将来的にさらに自律的で適応性の高いモデルへの道を開きます。



---

質問:  $> 1$  の場合、 $\sqrt{x} - \sqrt{x+1}$  の実数解の合計は

---

回答: <考えてください>

方程式  $\sqrt{x} - \sqrt{x+1} = \sqrt{x} - \sqrt{x+1}$  を解くには、まず、両方を二乗することから始めましょう・・・  

$$\frac{\sqrt{x} - \sqrt{x+1}}{2} = 2 \Rightarrow -\sqrt{x+1} = 2$$

内側の平方根項を分離するために並べ替えます:  $(-\frac{2}{2})^2 = 4 \Rightarrow 4 - 2x - 1 = 0$   
 ...

待って、待って。それは私がここでフラグを立てることができる「なるほど」という瞬間です。

これを段階的に再評価して、正しい合計が次のようにできるか確認してみましょう。・・・次の

方程式から始めました:  $\sqrt{x} - \sqrt{x+1} =$

まず、両辺を二乗

します:  $-\sqrt{x+1} = 2 \Rightarrow \sqrt{x+1} = -2$  次に、  
 次の方程式を次のように扱い、両辺をも

う一度二乗します: ...

---

表 3 | DeepSeek-R1-Zero の中間バージョンの興味深い「ひらめき」。モデルは擬人化された口調を使用して考え直すことを学習します。これは私たちにとってもひらめきの瞬間であり、強化学習の威力と美しさを目の当たりにすることができます。

DeepSeek-R1-Zero の欠点 DeepSeek-R1-Zero は強力な推論機能を発揮し、予期しない強力な推論動作を自律的に開発しますが、いくつかの問題に直面しています。たとえば、DeepSeek-R1-Zero は、読みにくさや言語の混在などの課題に悩まされています。推論プロセスをより読みやすくし、オープンコミュニティと共有するために、人間に優しいコールドスタートデータを使用して RL を利用する方法である DeepSeek-R1 を検討します。

### 2.3. DeepSeek-R1: コールドスタートによる強化学習

DeepSeek-R1-Zero の有望な結果に触発されて、2 つの自然な疑問が生じます。1) 少量の高品質データをコールドスタートとして組み込むことで、推論パフォーマンスをさらに向上させたり、収束を加速したりできるでしょうか。2) 明確で首尾一貫した思考の連鎖 (CoT) を生成するだけでなく、強力な一般機能も示す、ユーザーフレンドリなモデルをどのようにトレーニングできるでしょうか。これらの疑問に対処するために、DeepSeek-R1 をトレーニングするためのパイプラインを設計します。パイプラインは、次のように概説される 4 つのステージで構成されています。

#### 2.3.1. コールドスタート

DeepSeek-R1-Zero とは異なり、ベースモデルからの RL トレーニングの初期の不安定なコールドスタートフェーズを防ぐために、DeepSeek-R1 では少量の長い CoT データを構築して収集し、モデルを初期の RL アクターとして微調整します。このようなデータを収集するために、いくつかのアプローチを検討しました。長い CoT を例として、数回のプロンプトを使用する、モデルに直接プロンプトして詳細な回答を生成させ、反映と検証を行う、DeepSeek-R1-Zero の出力を読み取り可能な形式で収集する、人間による後処理で結果を洗練する

注釈者。

この研究では、DeepSeek-V3-Base を RL の出発点として微調整するために、何千ものコールドスタートデータを収集しました。DeepSeek-R1-Zero と比較して、コールドスタートデータの利点は次のとおりです。

含む：

- 読みやすさ: DeepSeek-R1-Zero の主な制限は、そのコンテンツが読みにくいことが多いことです。応答には複数の言語が混在していたり、ユーザー向けに回答を強調するためのマークダウン形式が欠けていたりすることがあります。対照的に、DeepSeek-R1 のコールド スタート データを作成するときは、各応答の最後に要約が含まれ、読みにくい応答を除外する読みやすいパターンを設計します。ここでは、出力形式を `[special_token|<reasoning_process>|special_token|summary]` として定義します。ここで、推論プロセスはクエリの CoT であり、要約は推論結果を要約するために使用されます。
- 可能性: 人間の事前情報を使用してコールド スタート データのパターンを慎重に設計することで、DeepSeek-R1-Zero よりも優れたパフォーマンスが観察されます。反復トレーニングは推論モデルのより良い方法であると考えています。

### 2.3.2. 推論指向強化学習

コールド スタート データで DeepSeek-V3-Base を微調整した後、DeepSeek-R1-Zero で採用されているものと同じ大規模強化学習トレーニング プロセスを適用します。このフェーズでは、特にコーディング、数学、科学、論理推論などの推論集約型のタスク（明確な解決策を持つ明確に定義された問題を含む）におけるモデルの推論機能の強化に重点が置かれます。トレーニング プロセス中に、特に RL プロンプトに複数の言語が含まれる場合、CoT で言語の混合が頻繁に発生することが観察されます。言語の混合の問題を軽減するために、RL トレーニング中に言語の一貫性の報酬を導入します。これは、CoT 内のターゲット言語の単語の割合として計算されます。アブレーション実験では、このような調整によってモデルのパフォーマンスがわずかに低下することが示されていますが、この報酬は人間の好みと一致し、読みやすくなります。最後に、推論タスクの精度と言語の一貫性の報酬を直接合計して組み合わせ、最終的な報酬を形成します。次に、微調整されたモデルに RL トレーニングを適用し、推論タスクの収束を達成します。

### 2.3.3. 棄却サンプリングと教師あり微調整

推論指向の強化学習が収束すると、結果のチェックポイントを利用して、次のラウンドの SFT (教師あり微調整) データを収集します。主に推論に焦点を当てた初期のコールド スタート データとは異なり、この段階では他のドメインからのデータが組み込まれ、ライティング、ロール プレーイング、およびその他の汎用タスクにおけるモデルの機能が向上します。具体的には、以下のようにデータを生成し、モデルを微調整します。

推論データ上記の RL トレーニングのチェックポイントから拒否サンプリングを実行して、推論プロンプトをキュレートし、推論軌道を生成します。前の段階では、ルールベースの報酬を使用して評価できるデータのみを含めました。ただし、この段階では、追加のデータを組み込むことでデータセットを拡張します。その一部は、判断のためにグラウンドトゥールズとモデル予測を DeepSeek-V3 に入力することで生成報酬モデルを使用します。

さらに、モデルの出力は混乱していて読みにくい場合があるため、混合言語、長い言い換え、コード ブロックを含む思考の連鎖を除外しました。プロンプトごとに複数の応答をサンプリングし、正しい応答のみを保持します。合計で、推論関連のトレーニングサンプルを約 60 万件収集します。

非推論データライティング、事実に基づく QA、自己認識、翻訳などの非推論データについては、DeepSeek-V3 パイプラインを採用し、DeepSeek-V3 の SFT データセットの一部を再利用します。特定の非推論タスクについては、プロンプトで質問に答える前に、DeepSeek-V3 を呼び出して潜在的な思考の連鎖を生成します。ただし、「こんにちは」などのより単純なクエリについては、応答として CoT を提供しません。最終的に、推論とは関係のないトレーニング サンプルを合計で約 20 万個収集しました。

上記の約 80 万サンプルのキュレーションされたデータセットを使用して、DeepSeek-V3-Base を 2 エポックにわたって微調整します。

### 2.3.4. あらゆるシナリオに対応する強化学習

モデルを人間の好みにさらに合わせるために、モデルの有用性と無害性を向上させると同時にその推論能力を洗練させることを目的とした二次強化学習段階を実装します。具体的には、報酬信号と多様なプロンプト分布の組み合わせを使用してモデルをトレーニングします。推論データについては、DeepSeek-R1-Zero で概説されている方法論に従います。この方法では、ルールベースの報酬を使用して、数学、コード、および論理推論ドメインでの学習プロセスをガイドします。一般的なデータについては、複雑で微妙なシナリオで人間の好みを捉えるために報酬モデルを使用します。DeepSeek-V3 パイプラインを基盤とし、同様の好みのペアとトレーニングプロンプトの分布を採用します。有用性については、最終的な要約のみに焦点を当て、基礎となる推論プロセスへの干渉を最小限に抑えながら、ユーザーへの応答の有用性と関連性を強調する評価を確実に行います。無害性については、推論プロセスと要約の両方を含むモデルの応答全体を評価して、生成プロセス中に発生する可能性のある潜在的なリスク、バイアス、または有害なコンテンツを特定して軽減します。最終的には、報酬信号と多様なデータ分布を統合することで、有用性と無害性を優先しながら推論に優れたモデルをトレーニングできます。

## 2.4. 蒸留: 小さなモデルに推論機能を追加する

より効率的な小規模モデルに DeepSeek-R1 のような推論機能を搭載するために、§ 2.3.3 で詳述されているように、DeepSeek-R1 でキュレーションされた 80 万個のサンプルを使用して、Qwen (Qwen, 2024b) や Llama (AI@Meta, 2024) などのオープンソース モデルを直接微調整しました。調査結果によると、**この単純な蒸留方法により、小規模モデルの推論機能が大幅に向上することが示されています**。ここで使用する基本モデルは、Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.5-14B、Qwen2.5-32B、Llama-3.1-8B、および Llama-3.3-70B-Instruct です。Llama-3.3 を選択したのは、その推論機能が Llama-3.1 よりもわずかに優れているためです。

蒸留モデルの場合、RL を組み込むとモデルのパフォーマンスが大幅に向上する可能性があるにもかかわらず、SFT のみを適用し、RL ステージは含めません。ここでの主な目標は、蒸留手法の有効性を実証することであり、RL ステージの調査はより広範な研究コミュニティに委ねます。

## 3. 実験

ベンチマークMMLU (Hendrycks et al., 2020)、MMLU-Redux (Gema et al., 2024)、MMLU-Pro (Wang et al., 2024)、C-Eval (Huang et al., 2023)、CMMLU (Li et al., 2023)、IFEval (Zhou et al., 2023)、FRAMES (Krishna et al., 2024)、GPQA Diamond (Rein et al., 2023)、SimpleQA (OpenAI, 2024c)、C-SimpleQA (He et al., 2024)、SWE-Bench Verified (OpenAI,

2024d).Aider<sup>1</sup>、LiveCodeBench (Jain et al., 2024) (2024-08 – 2025-01)、Codeforces<sup>2</sup>、中国語 National High School Mathematics Olympiad (CNMO 2024)<sup>3</sup>、American Invitational Mathematics Examination 2024 (AIME 2024) (MAA、2024) などの数学オリンピックでモデルを評価しています。標準ベンチマークに加えて、LLM を審査員として使用して、オープンエンド生成タスクでモデルを評価します。具体的には、GPT-4-Turbo-1106 をペアワイズ比較の審査員として活用する AlpacaEval 2.0 (Dubois et al., 2024) と Arena-Hard (Li et al., 2024) の元の構成に準拠しています。ここでは、長さのバイアスを回避するために、最終サマリーのみを評価にフィードします。抽出されたモデルについては、AIME 2024、MATH-500、GPQA Diamond、Codeforces、およびLiveCodeBench での代表的な結果を報告します。

評価プロンプトDeepSeek-V3 のセットアップに従って、MMLU、DROP、GPQA Diamond、SimpleQA などの標準ベンチマークが、simple-evals フレームワークのプロンプトを使用して評価されます。MMLU-Redux の場合、ゼロショット設定で Zero-Eval プロンプト形式 (Lin、2024) を採用します。MMLU-Pro、C-Eval、CLUE-WSC に関しては、元のプロンプトが少数ショットであるため、プロンプトをゼロショット設定に少し変更します。少数ショットの CoT は DeepSeek-R1 のパフォーマンスに悪影響を及ぼす可能性があります。他のデータセットは、作成者によって提供されるデフォルトのプロンプトを使用して、元の評価プロトコルに従います。コードと数学のベンチマークの場合、HumanEval-Mul データセットは 8 つの主流のプログラミング言語 (Python、Java、C++、C#、JavaScript、TypeScript、PHP、Bash) をカバーしています。LiveCodeBench のモデル パフォーマンスは、2024 年 8 月から 2025 年 1 月の間に収集されたデータを使用して CoT 形式で評価されます。Codeforcesデータセットは、10 の Div.2 コンテストの問題と専門家が作成したテスト ケースを使用して評価され、その後、予想される評価と競合者の割合が計算されます。SWE-Bench 検証結果は、エージェントレス フレームワーク (Xia 他、2024) を介して取得されます。AIDER 関連のベンチマークは、「diff」形式を使用して測定されます。DeepSeek-R1 出力は、ベンチマークごとに最大 32,768 トークンに制限されます。

ベースラインDeepSeek-V3、Claude-Sonnet-3.5-1022、GPT-4o-0513、OpenAI-o1-mini、OpenAI-o1-1217 など、いくつかの強力なベースラインに対して包括的な評価を実施します。

中国本土ではOpenAI-o1-1217 APIへのアクセスが難しいため、公式レポートに基づいてそのパフォーマンスを報告します。抽出モデルについては、オープンソースモデルQwQ-32B-Preview (Qwen, 2024a)とも比較します。

評価設定モデルの最大生成長を 32,768 トークンに設定しました。

貪欲なデコードを使用して長い出力の推論モデルを評価すると、繰り返し率が高くなり、さまざまなチェックポイント間で大きな変動が生じることがわかりました。そのため、デフォルトで pass@ 評価 (Chen et al., 2021) を使用し、ゼロ以外の温度を使用して pass@1 を報告します。

具体的には、サンプリング温度0.6、最高値0.95を使用して、各質問に対する回答 (テストセットのサイズに応じて、通常は4~64)を生成します。Pass@1は次のように計算されます。

$$1 \text{ Pass}@1 = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\text{correct}}(i)$$

ここで、 $\mathbb{1}_{\text{correct}}(i)$  は  $i$  番目の応答の正確さを表します。この方法は、より信頼性の高いパフォーマンス推定を提供します。AIME 2024 では、cons@64 と表記される 64 個のサンプルを使用したコンセンサス (多数決) 結果 (Wang et al., 2022) も報告します。

<sup>1</sup><https://aider.chat>

<sup>2</sup><https://codeforces.com>

<sup>3</sup><https://www.cms.org.cn/Home/comp/comp/cid/12.html>

## 3.1. DeepSeek-R1の評価

| ベンチマーク（メトリック） |                           | Claude-3.5- GPT-4o DeepSeek OpenAI OpenAI DeepSeek<br>ソネット-1022 0513 V3 o1-mini o1-1217 |      |       | R1   |      |       |
|---------------|---------------------------|---|------|-------|------|------|-------|
| 建築            |                           | -   | -    | 文部科学省 | -    | -    | 文部科学省 |
|               | # アクティブ化されたパラメータ          | -   | -    | 37B   | -    | -    | 37B   |
|               | # 合計パラメータ                 | -   | -    | 671B  | -    | -    | 671B  |
| 英語            | MMLU（パス@1）                | 88.3  | 87.2 | 88.5  | 85.2 | 91.8 | 90.8  |
|               | MMLU-Redux（EM）            | 88.9  | 88.0 | 89.1  | 86.7 | -    | 92.9  |
|               | MMLU-プロ（EM）               | 78.0  | 72.6 | 75.9  | 80.3 | -    | 84.0  |
|               | DROP（3ショットF1）             | 88.3  | 83.7 | 91.6  | 83.9 | 90.2 | 92.2  |
|               | IF-Eval（プロンプト厳密）          | 86.5  | 84.3 | 86.1  | 84.8 | -    | 83.3  |
|               | GPQA ダイヤモンド(合格@1)         | 65.0  | 49.9 | 59.1  | 60.0 | 75.7 | 71.5  |
|               | SimpleQA（正解）              | 28.4  | 38.2 | 24.9  | 7.0  | 47.0 | 30.1  |
|               | フレーム（付属品）                 | 72.5  | 80.5 | 73.3  | 76.9 | -    | 82.5  |
|               | AlpacaEval2.0（LC 勝率）      | 52.0  | 51.1 | 70.0  | 57.8 | -    | 87.6  |
|               | アリーナハード(GPT-4-1106)       | 85.2  | 80.4 | 85.5  | 92.0 | -    | 92.3  |
| コード           | LiveCodeBench（Pass@1-COT） | 38.9  | 32.9 | 36.2  | 53.8 | 63.4 | 65.9  |
|               | コードフォース（パーセンタイル）          | 20.3  | 23.6 | 58.7  | 93.4 | 96.6 | 96.3  |
|               | コードフォース(評価)               | 717   | 759  | 1134  | 1820 | 2061 | 2029  |
|               | SWE 検証済み(解決済み)            | 50.8  | 38.8 | 42.0  | 41.6 | 48.9 | 49.2  |
|               | Aider-Polyglot（Acc.）      | 45.3  | 16.0 | 49.6  | 32.9 | 61.7 | 53.3  |
| 数学            | AIME 2024（合格@1）           | 16.0  | 9.3  | 39.2  | 63.6 | 79.2 | 79.8  |
|               | MATH-500（合格@1）            | 78.3  | 74.6 | 90.2  | 90.0 | 96.4 | 97.3  |
|               | CNMO 2024（合格@1）           | 13.1  | 10.8 | 43.2  | 67.6 | -    | 78.8  |
| 中国語           | CLUEWSC（EM）               | 85.4  | 87.9 | 90.9  | 89.9 | -    | 92.8  |
|               | C評価（EM）                   | 76.7  | 76.0 | 86.5  | 68.9 | -    | 91.8  |
|               | C-SimpleQA（正解）            | 55.4  | 58.7 | 68.0  | 40.3 | -    | 63.7  |

表 4 | DeepSeek-R1 と他の代表的なモデルの比較。

## MMLU、MMLU-Pro、GPQAなどの教育向け知識ベンチマーク

ダイヤモンド、DeepSeek-R1はDeepSeek-V3と比較して優れたパフォーマンスを発揮します。この改善は主にSTEM関連の質問における精度の向上によるもので、大規模な強化学習によって大きな成果が達成されています。さらに、DeepSeek-R1は

長時間のコンテキスト依存QAタスクであるFRAMESに優れており、強力なドキュメント分析を実証しています。これは、AI駆動型検索とデータにおける推論モデルの潜在能力を浮き彫りにするものである。分析タスク。事実上のベンチマークであるSimpleQAでは、DeepSeek-R1はDeepSeek-V3よりも優れています。

事実に基づくクエリを処理する能力を実証しています。同様の傾向は、

OpenAI-o1はこのベンチマークでGPT-4oを上回っています。しかし、DeepSeek-R1はそれよりもパフォーマンスが劣っています。

DeepSeek-V3は、主に拒否傾向のため、中国のSimpleQAベンチマークで

安全強化学習の後に特定のクエリに答える。安全強化学習がなければ、DeepSeek-R1は70%以上の精度。

DeepSeek-R1は、IF-Evalというベンチマークでも素晴らしい結果を示しています。これは、モデルのフォーマット指示に従う能力。これらの改善は、

教師あり微調整 (SFT)と強化学習の最終段階での指示追従データの

トレーニング。さらに、AlpacaEval2.0とArenaHardでは顕著なパフォーマンスが観察されています。

これは、DeepSeek-R1が文章作成タスクとオープンドメインの質問応答に優れていることを示しています。

DeepSeek-V3の大幅なパフォーマンスは、大規模システムの一般化の利点を強調しています。

RLは推論能力を高めるだけでなく、多様なパフォーマンスも向上させます。

さらに、DeepSeek-R1によって生成される要約の長さは簡潔で、

ArenaHardでは平均689トークン、AlpacaEval 2.0では平均2,218文字です。これは、

DeepSeek-R1はGPTベースの評価中に長さのバイアスを導入することを回避し、さらに複数のタスクにわたる堅牢性。

数学の課題では、DeepSeek-R1はOpenAI-o1-1217と同等のパフォーマンスを発揮します。他のモデルを大きく上回っています。コーディングアルゴリズムでも同様の傾向が見られます。LiveCodeBenchやCodeforcesなどのタスクでは、推論に重点を置いたモデルが主流となっている。ベンチマーク。エンジニアリング指向のコーディングタスクでは、OpenAI-o1-1217はDeepSeek-R1よりも優れています。Aiderでは同等のパフォーマンスを達成していますが、SWE Verifiedでは同等のパフォーマンスを達成しています。DeepSeek-R1のパフォーマンスは、関連するRLの量が増えるにつれて、次のバージョンで向上するでしょう。トレーニング データは現時点では非常に限られています。

3.2. 蒸留モデルの評価

| モデル                               | EIM2024                         |       | 数学500 | GPQA ライブコード<br>ダイヤモンドベンチ |      | コードフォース |      |
|-----------------------------------|---------------------------------|-------|-------|--------------------------|------|---------|------|
|                                   | パス@1                            | 短所@64 |       | パス@1                     | パス@1 |         | 評価   |
| GPT-4o-0513                       | 9.3                             | 13.4  | 74.6  | 49.9                     | 32.9 | 759     |      |
| クロード-3.5-ソネット-1022 OpenAI-        | 16.0                            | 26.7  | 78.3  | 65.0                     | 38.9 | 717     |      |
| o1-mini QwQ-32B-プレビ               | 63.6                            | 80.0  | 90.0  | 60.0                     | 53.8 | 1820    |      |
| ュー                                | 50.0                            | 60.0  | 90.6  | 54.5                     | 41.9 | 1316    |      |
| ディープシーク-R1-Distill-Qwen-1.5B 28.9 | ディープシーク-R1-Distill-Qwen-7B 55.5 |       | 52.7  | 83.9                     | 33.8 | 16.9    | 954  |
| ディープシーク-R1-Distill-Qwen-14B 69.7  | ディープシーク-R1-Distill-Qwen-32B     |       | 83.3  | 92.8                     | 49.1 | 37.6    | 1189 |
| ディープシーク-R1-Distill-Qwen-70B       | ディープシーク-R1-Distill-Llama-8B     |       | 80.0  | 93.9                     | 59.1 | 53.1    | 1481 |
| ディープシーク-R1-Distill-Llama-70B      | 72.6                            | 83.3  | 94.3  | 62.1                     | 57.2 | 1691    |      |
|                                   | 50.4                            | 80.0  | 89.1  | 49.0                     | 39.6 | 1205    |      |
|                                   | 70.0                            | 86.7  | 94.5  | 65.2                     | 57.5 | 1633    |      |

表5 | DeepSeek-R1の抽出モデルと他の類似モデルの比較  
推論関連のベンチマーク。

表5に示すように、DeepSeek-R1の出力を単純に抽出すると、効率的なDeepSeek-R1-7B（つまり、DeepSeek-R1-Distill-Qwen-7B、以下同様に略記）は、GPT-4o-0513などの非推論モデルを全面的に上回ることができます。DeepSeek-R1-14Bは、すべての評価指標でQwQ-32B-Previewを上回り、DeepSeek-R1-32BとDeepSeek-R1-70Bは大幅に優れています。ほとんどのベンチマークでo1-miniを超えています。これらの結果は蒸留の大きな可能性を示しています。さらに、これらの蒸留モデルにRLを適用すると、さらに大きな改善が得られることがわかりました。利益。これはさらなる調査が必要であると考え、そのため、シンプルな SFT 蒸留モデルはこちら。

4. 議論

4.1. 蒸留学習と強化学習

セクション3.2では、DeepSeek-R1を抽出して、小さなモデルが印象的な結果を達成できることがわかります。しかし、まだ1つの疑問が残っています。モデルは同等のパフォーマンスを達成できるのでしょうか？論文で議論されている大規模なRLトレーニングを蒸留なしで実行できるのでしょうか？

この質問に答えるために、我々はQwen-32B-Base上で数学を用いて大規模なRLトレーニングを実施しました。コードとSTEMデータを組み合わせて1万ステップ以上のトレーニングを行い、DeepSeek-R1-Zero-Qwen-32Bが完成しました。表6に示す実験結果は、32Bベースモデルが大規模実験の後、



| モデル   | エイム2024 |            | MATH-500 GPQA ダイヤモンド LiveCodeBench |      |      |
|---|---------|------------|------------------------------------|------|------|
|   | パス@1    | 短所@64 パス@1 |                                    | パス@1 | パス@1 |
| QwQ-32B-プレビュー                                       | 50.0    | 60.0       | 90.6                               | 54.5 | 41.9 |
| DeepSeek-R1-Zero-Qwen-32B 47.0 DeepSeek-R1-Distill- |         | 60.0       | 91.6                               | 55.0 | 40.2 |
| Qwen-32B 72.6                                       |         | 83.3       | 94.3                               | 62.1 | 57.2 |

表 6 | 推論関連ベンチマークにおける蒸留モデルと RL モデルの比較。

RLトレーニングでは、QwQ-32B-Previewと同等のパフォーマンスを達成しています。ただし、DeepSeek-R1-DeepSeek-R1から抽出されたDistill-Qwen-32Bは、すべてのベンチマークにおける DeepSeek-R1-Zero-Qwen-32B。

したがって、2つの結論を導き出すことができます。まず、より強力なモデルをより小さなモデルに凝縮することで、大規模なRLに頼った小規模なモデルは優れた結果をもたらすが、この論文は膨大な計算能力を必要とし、パフォーマンスさえ達成できないかもしれない。蒸留の第二の理由は、蒸留戦略は経済的かつ効果的であるが、知能の限界を超えるには、さらに強力な基本モデルと大規模な強化学習が必要になる可能性があります。

#### 4.2. 失敗した試み

DeepSeek-R1の開発初期段階では、失敗や挫折もありました。道です。ここで失敗体験をシェアするのは、洞察力を提供するためですが、これはこれらのアプローチでは、効果的な推論モデルを開発することができません。

**プロセス報酬モデル (PRM)** PRMはモデルをより良い方向に導くための合理的な方法です。

推論課題を解決するためのアプローチ（ライトマンら、2023年；上里ら、2022年；王ら、しかし、実際には、PRMには最終的な成功を妨げる可能性のある3つの主な制限があります。第一に、一般的な推論における細かいステップを明示的に定義することは困難です。第二に、現在の中間ステップが正しいかどうかを判断するのは難しい作業です。自動化モデルを使った注釈では満足いく結果が得られない可能性があり、手動注釈ではスケールアップに役立たない。第三に、モデルベースのPRMが導入されると、必然的に報酬が増加する。ハッキング（Gao et al., 2022）であり、報酬モデルの再トレーニングには追加のトレーニングリソースが必要であるトレーニングパイプライン全体を複雑にします。結論として、PRMは優れたモデルによって生成された上位Nの回答を再ランク付けしたり、ガイド付き検索を支援したりする機能（Snell et al., 2024）によると、その利点は、追加の計算オーバーヘッドに比べて限られている。私たちの実験では、大規模な強化学習プロセス中に導入されます。

**モンテカルロ木探索 (MCTS)** AlphaGo（Silver et al., 2017b）とAlphaZero（Silver et al., 2017a）に触発され、モンテカルロ木探索（MCTS）を使用してテスト時間の短縮を検討しました。計算のスケラビリティ。このアプローチでは、回答を小さな部分に分割して、モデルは、体系的に解空間を探索します。これを促進するために、モデルに検索に必要な特定の推論ステップに対応する複数のタグを生成します。

トレーニングでは、まず収集したプロンプトを使用して、事前にトレーニングされた値によって導かれるMCTSを介して回答を見つけます。モデル。その後、得られた質問と回答のペアを使用して、俳優モデルとそして価値モデルを構築し、プロセスを反復的に改良します。

しかし、このアプローチでは、トレーニングを拡大する際にいくつかの課題に直面します。まず、探索空間が比較的明確に定義されているチェスとは異なり、トークン生成は

指数関数的に大きな検索空間。これに対処するために、各ノードに最大拡張制限を設定しましたが、これによりモデルが局所最適値で停止する可能性があります。次に、価値モデルは検索プロセスの各ステップをガイドするため、生成の品質に直接影響します。

きめ細かい価値モデルのトレーニングは本質的に難しいため、モデルを反復的に改善することが困難です。AlphaGo の成功の鍵は、価値モデルをトレーニングしてパフォーマンスを徐々に向上させることでしたが、トークン生成の複雑さにより、この原則を私たちのセットアップで再現することは困難です。

結論として、MCTS は事前トレーニング済みの価値モデルと組み合わせると推論中のパフォーマンスを向上させることができますが、自己探索を通じてモデルのパフォーマンスを反復的に向上させることは依然として大きな課題です。

## 5. 結論、限界、今後の課題

この研究では、強化学習を通じてモデルの推論能力を向上させる取り組みについて共有します。DeepSeek-R1-Zero は、コールド スタート データに依存せず、さまざまなタスクで優れたパフォーマンスを実現する純粋な RL アプローチを表しています。DeepSeek-R1 はさらに強力で、コールド スタート データと反復的な RL 微調整を活用します。最終的に、DeepSeek-R1 はさまざまなタスクで OpenAI-o1-1217 に匹敵するパフォーマンスを実現します。

さらに、推論能力を小規模の密なモデルに蒸留することを検討しました。DeepSeek-R1 を教師モデルとして使用して 80 万個のトレーニング サンプルを生成し、いくつかの小規模の密なモデルを微調整しました。結果は有望で、DeepSeek-R1-Distill-Qwen-1.5B は、数学ベンチマークで GPT-4o および Claude-3.5-Sonnet を上回り、AIME で 28.9%、MATH で 83.9% の成績を達成しました。他の密なモデルも印象的な結果を達成し、同じ基礎チェックポイントに基づく他の命令調整モデルを大幅に上回りました。

今後、DeepSeek-R1では以下の方向性で研究に投資していく予定です。

- 一般的な機能: 現在、DeepSeek-R1 の機能は、関数呼び出し、マルチターン、複雑なロールプレイング、JSON 出力などのタスクでは DeepSeek-V3 に劣ります。  
 今後は、これらの分野のタスクを強化するために CoT をどの程度活用できるかを検討する予定です。
- 言語の混在: DeepSeek-R1 は現在、中国語と英語に最適化されているため、他の言語でのクエリを処理するときに言語の混在の問題が発生する可能性があります。たとえば、クエリが英語または中国語以外の言語であっても、DeepSeek-R1 は推論と応答に英語を使用する場合があります。今後のアップデートでこの制限に対処する予定です。
- プロンプト エンジニアリング: DeepSeek-R1 を評価すると、プロンプトに敏感であることがわかります。プロンプトを数回実行すると、パフォーマンスが一貫して低下します。したがって、最適な結果を得るには、**ユーザーが問題を直接記述し、ゼロショット設定を使用して出力形式を指定することをお勧めします。**
- ソフトウェア エンジニアリング タスク: 評価時間が長く、RL プロセスの効率に影響するため、大規模 RL はソフトウェアエンジニアリング タスクにはあまり適用されていません。その結果、DeepSeek-R1 はソフトウェア エンジニアリング ベンチマークで DeepSeek-V3 に比べて大きな改善を示していません。将来のバージョンでは、ソフトウェア エンジニアリング データに拒否サンプリングを実装するか、RL プロセス中に非同期評価を組み込むことで、効率性を向上させることで、この問題に対処します。

## 参考文献

AI@Meta, Llama 3.1 モデル カード, 2024 年。URL [https://github.com/meta-llama/llama-models/blob/main/models/llama3\\_1/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md)。

アントロピック。クロード 3.5 ソネット, 2024 年。URL <https://www.anthropic.com/news/claude-3-5-sonnet>。

M. チェン, J. トウレク, H. ジュン, Q. ユアン, HP デオリベイラ ピント, J. カプラン, H. エドワーズ, Y. バーダ, N. ジョセフ, G. プロックマン, A. レイ, R. プリ, G. クルーガー, M. ペトロフ, H. クラフ, G. サストリー, P. ミシュキン, B. チャン, S. グレイ, N. ライダー, M. パブロフ, A. パワー, L. カイザー, M. パバリアン, C. ウィンター, P. ティレット, FP サッチ, D. カミングス, M. プラパート, F. チャンツィス, E. パーンズ, A. ハーバート フォス, WH ガス, A. ニコル, A. ペインオ, N. テザック, J. タン, J. バブシュキン, S. バラジ, S. ジェイン, W. サンダース, C. ヘッセ, AN カー, J. ライケ, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, J. Sutskever, W. Zaremba, コードでトレーニングされた大規模言語モデルの評価。CoRR, abs/2107.03374, 2021 年。

URL <https://arxiv.org/abs/2107.03374>。

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, 他。Llama 3 の群れのモデル。arXiv プレプリント arXiv:2407.21783, 2024。

Y. デュボア, B. ガランボシ, P. リャン, TB 橋本。長さ制御されたアルパカエバル: シンプルな自動評価者のバイアスを解消する方法。arXiv プレプリント arXiv:2404.04475, 2024。

X. Feng, Z. Wan, M. Wen, SM McAleer, Y. Wen, W. Zhang, J. Wang, Alphazero のようなツリー検索は、大規模な言語モデルのデコードとトレーニングをガイドできます。2024 年。URL <https://arxiv.org/abs/2309.17179>。

L. Gao, J. Schulman, J. Hilton。報酬モデルの過剰最適化に関するスケーリング法則, 2022 年。URL <https://arxiv.org/abs/2210.10760>。

AP Gema, JOJ Leang, G. Hon, A. Devoto, ACM Mancino, R. Saxena, X. He, Y. Zhao, X. Du, MRG Madani, C. Barale, R. McHardy, J. Harris, J. Kaddour, E. van Krieken, および P. Minervini。mmlu はもう終わりですか? CoRR, abs/2406.04127, 2024。URL <https://doi.org/10.48550/arXiv.2406.04127>。

Google。次世代モデル: Gemini 1.5, 2024 年。URL <https://blog.google/technology/ai/google-gemini-次世代モデル-2024年2月>。

Y. He, S. Li, J. Liu, Y. Tan, W. Wang, H. Huang, X. Bu, H. Guo, C. Hu, B. Zheng, 他「Chinese simpleqa: 大規模言語モデルのための中国語事実性評価」 arXiv プレプリント arXiv:2411.07140, 2024。

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt。大規模マルチタスク言語理解の測定。arXiv プレプリント arXiv:2009.03300, 2020 年。

Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, 他「C-Eval: 基礎モデルのためのマルチレベル、マルチ分野の中国語評価スイート」 arXiv プレプリント arXiv:2305.08322, 2023 年。

N. Jain, K. Han, A. Gu, W. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, I. Stoica。Livecodebench: コードの大規模言語モデルの全体的かつ汚染のない評価。CoRR, abs/2403.07974, 2024。URL <https://doi.org/10.48550/arXiv.2403.07974>。

S. クリシュナ、K. クリシュナ、A. モハナニー、S. シュワルツ、A. スタンブラー、S. ウパディヤイ、M. ファルキ。

事実、フェッチ、推論：検索強化型生成の統一評価。CoRR、[abs/2409.12941](https://arxiv.org/abs/2409.12941), 2024。doi :10.48550/ARXIV.2409.12941。URL <https://doi.org/10.48550/arXiv.2409.12941>。

A. Kumar, V. Zhuang, R. Agarwal, Y. Su, JD Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, 他「強化学習による自己修正言語モデルのトレーニング」 arXiv プレプリント arXiv:2409.12917, 2024。

H. リー、Y. チャン、F. コト、Y. ヤン、H. チャオ、Y. ゴン、N. ドゥアン、および T. ボールドウィン。CMMLU: 中国語における大量のマルチタスク言語理解を測定します。arXiv プレプリント arXiv:2306.09212, 2023。

T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, JE Gonzalez, I. Stoica. クラウドソーシングされたデータから高品質のベンチマークへ: Arena-hard およびベンチビルダー パイプライン。arXiv プレプリント arXiv:2406.11939, 2024。

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe. ステップごとに検証してみましょう。arXiv プレプリント arXiv:2305.20050, 2023。

リン著。ZeroEval: 言語モデルを評価するための統一フレームワーク, 2024 年 7 月。URL <https://github.com/WildEval/ZeroEval>。

MAA。アメリカ招待数学試験 - aime。アメリカ招待数学試験 - AIME 2024, 2024 年 2 月。URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>。

オープンAI。こんにちは、GPT-4o, 2024a。URL <https://openai.com/index/hello-gpt-4o/>。

OpenAI。LLMs による推論の学習, 2024b。URL <https://openai.com/index/learning-to-reason-with-llms/>。

OpenAI。SimpleQA の紹介, 2024c。URL <https://openai.com/index/introducing-simpleqa/>。

OpenAI。SWE-bench verified のご紹介。2024 年以降に人間が検証した SWE-bench のサブセットをリリースします。URL <https://openai.com/index/introducing-swe-bench-verified/>。

Qwen。Qwq: 未知の境界について深く考える, 2024a。URL <https://qwenlm.github.io/blog/qwq-32b-preview/>。

Qwen。Qwen2.5: 基礎モデルのパーティ, 2024b。URL <https://qwenlm.github.io/blog/qwen2.5>。

D. Rein, BL Hou, AC Stickland, J. Petty, RY Pang, J. Dirani, J. Michael, および SR Bowman。GPQA: 大学院レベルの Google 対応 Q&A ベンチマーク。arXiv プレプリント arXiv:2311.12022, 2023。

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu, および D. 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る。arXiv プレプリント arXiv:2402.03300, 2024。

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, TP Lillicrap, K. Simonyan, および D. Hassabis。汎用強化学習アルゴリズムを使用したセルフプレイによるチェスと将棋の習得。CoRR、[abs/1712.01815](https://arxiv.org/abs/1712.01815), 2017a。URL <http://arxiv.org/abs/1712.01815>。

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, TP Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, および D. Hassabis. 人間の知識なしで囲碁のゲームをマスターする。Nat. , 550(7676):354–359, 2017b. doi: 10.1038/NATURE24270. URL <https://doi.org/10.1038/nature24270>.

C. Snell, J. Lee, K. Xu, A. Kumar. LLM テスト時間の計算を最適にスケールすると、モデル パラメーターをスケールするよりも効果的になる場合があります。2024 年。URL <https://arxiv.org/abs/2408.03314>.

T. Trinh, Y. Wu, Q. Le, H. He, T. Luong. 人間なしでオリンピック幾何学を解く  
デモ。Nature, 2024年。doi: 10.1038/s41586-023-06747-5.

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, I. Higgins. プロセスと結果に基づくフィードバックを使用して数学の文章問題を解く。arXiv プレプリント arXiv:2211.14275, 2022.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, Z. Sui. Math-shepherd : 数学的推論における LMS のラベル  
フリーステップバイステップ検証ツール。arXiv プレプリント arXiv :2312.08935, \_\_\_\_\_  
2023 年。

X. ワン, J. ウェイ, D. シュールマンズ, Q. リー, E. チー, S. ナラン, A. チョードリー, および D. チョウ。  
自己一貫性は言語モデルにおける思考連鎖推論を改善する。arXiv プレプリント arXiv:2203.11171, 2022. \_\_\_\_\_

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, W. Chen. Mmlu-pro: より堅牢で挑戦的なマルチタスク言語理解ベンチマーク。  
CoRR, abs/2406.01574, 2024. \_\_\_\_\_  
URL <https://doi.org/10.48550/arXiv.2406.01574>.

CS Xia, Y. Deng, S. Dunn, L. Zhang. エージェントレス : LLM ベースのソフトウェアエンジニアリングエージェントの謎を解明。arXiv プレプリント, 2024 年。\_\_\_\_\_

H. Xin, ZZ Ren, J. Song, Z. Shao, W. Zhao, H. Wang, B. Liu, L. Zhang, X. Lu, Q. Du, W. Gao, Q. Zhu, D. Yang, Z. Gou, ZF Wu, F. Luo, および C. Ruan. Deepseek-prover-v1.5: 強化学習とモンテカルロ木探索のための証明アシスタントフィードバックの活用, 2024 年。URL <https://arxiv.org/abs/2408.08152>.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, および L. Hou. 指示に従う  
大規模言語モデルの評価。arXiv プレプリント arXiv:2311.07911, 2023. \_\_\_\_\_

## 付録

### A. 貢献と謝辞

コア貢献者

ダヤ・グオ

デジヤン・ヤン

張浩偉

ソン・ジュンシャオ

張若宇

徐潤馨

朱 奇豪

馬志龍

王 ペイイー

シャオビ

Xiaokang Zhang Xingkai

Yu Yu Wu ZF Wu

Zhibin

Gou

Zhihong Shao

Zhuoshu Li Ziyi Gao

投稿者Aixin Liu Bing

Xue Bingxuan

Wang Bochao

Wu Bei Feng Chengda

Lu Chenggang

Zhao Chengqi

Deng Chong Ruan

Damai Dai Deli Chen

Dongjie Ji Erhang Li

Fangyun Lin

Fucong Dai Fuli

Luo\* Guangbo

Hao Guanting

Chen Guowei

Li H. Zhang Hanwei

Xu Honhui Ding

Huazuo Gao

Hui Qu

ホイ・リー

ジェンジョン グオ・ジアシ

ー リー・ジン

チャン チェン・ジンヤン ユ

アン・ジンハオ トウ・ジュ

ンジェ クイ・ジ

ユンロン・リー JL

カイ・ジアチー ニ

ー・ジャン・リ

ヤン・ジン カ

イ・ドン・カイフー

\* カイチャオ ヨ

ウ・カイゲ・ガオ

カン・グアン

ケシン ホアン・クアイ

・ユー・リーワン

・レコン チャン・リヤ

ン チャオ・リートン ワ

ン・リーユエ

チャン・レイ シュー

・レイイ・シア・ミンチュア

ンZhang Minghua

Zhang Minghui

Tang Mingxu Zhou

Meng Li

Miaojun

Wang Mingming Li Ning

Tian Panpan Huang

Peng Zhang

Qiancheng Wang

Qinyu Chen

Qiusi Du Ruiqi Ge\*

Ruisong Zhang

Ruizhe Pan

Runji Wang RJ Chen

RL Jin



如儀 チェン・シ  
 ヤンハオ ルー・シャ  
 ンヤン 周シャンファン  
 チェン・シェンフェン イェ  
 ・シーユー ワン・シュ  
 イピン ユー・シュ  
 ンフェン 周 シュ  
 ティン・パン SS リー・シ  
 ユアン 周・シャオ  
 チン ウ  
 ー・シェンフェン イ  
 エ・タオ・ユン ティア  
 ン・ペイ ティアンユ  
 ー サン・テ  
 イー ワン・ワ  
 ンディング ゼン  
 ・ウエン リウ  
 ・ウエンフェン リャン・ウ  
 エンジュン  
 ガオ・ウエンチン ユー  
 \* ウエンタオ・チャ  
 ン  
 WLシャオ  
 ウエイ・アン  
 ・シャオドン、リウ・シャ  
 オハン、ワン・シャオカ  
 ン、チェン・シャオタオ、  
 ニー・シン、チェン・  
 シン、リウ・シン、  
 シー・シンチ  
 ャオ、リウ・  
 シンユー、ヤン・シン  
 ユアン、リー・シュ  
 エチェン、スー・シ  
 ュヘン、リン・XQ、リー  
 ・シャンユエ、ジン・  
 シャオジ  
 ン、シェン・シャオシ  
 ヤ、チェン・シャオウ  
 エン、サン・シャオシア  
 ン、ワン・シンナン、ソ  
 ン・シンイー、周賢津、ワン・  
 シンシア・シャン

YK リー  
 YQ ワン

YX ウエイ  
 ヤン チャン ヤ  
 ンホン シュウ  
 ヤオ・リー  
 ヤオ・チャオ  
 孫堯峰  
 王 堯慧  
 イーユー  
 張一超  
 石一凡  
 イリヤン・シオン  
 イン・ヘ  
 イシ・ピャオ  
 ワン・イーソン  
 タン・イーシュアン  
 Yiyang Ma\*  
 Yiyuan Liu  
 Yongqiang Guo Yuan  
 Ou Yudian  
 Wang Yue Gong  
 Yuheng Zou  
 Yujia He Yuunfan  
 Xiong  
 Yuxiang Luo Yuxiang  
 You Yuxuan Liu  
 Yuyang Zhou YX  
 Zhu Yanping  
 Huang Yaohui Li Yi  
 Zheng  
 Yuchen Zhu Yunxian  
 Ma Ying Tang  
 Yukun Zha  
 Yuting Yan ZZ  
 Ren Zehui Ren  
 Zhangli Sha  
 Zhe Fu Zhean  
 Xu Zhenda Xie  
 Zhengyan  
 Zhang Zhewen  
 Hao Zhicheng Ma  
 Zhigang  
 Yan Zhiyu Wu  
 Zihui Gu

朱子佳  
劉子俊\*  
リー・ジリン  
謝紫薇  
ソン・ジヤン  
パン・ジジェン

ジェン・ファン  
徐志鵬  
張中宇  
張珍

各役割内で、著者はファーストネームのアルファベット順でリストされます。\* マークの付いた名前は、チームを離れた個人を表します。