deepseek

DeepSeekMath: 数学の限界に挑戦 オープン言語モデルにおける推論

Zhihong Shao1,2 †、 Peiyi Wang1,3 †、 Qihao Zhu1,3 †、 Runxin Xu1 、JunxiaoSong1 Y. Wu1 、Daya シャオビ1 、Haowei Zhang1 、Mingchuan Zhang1 、YK Li1 1DeepSeek-AI、 2、Guo1

清華大学、 3北京大学

{zhihongshao,wangpeiyi,zhuqh,guoday}@deepseek.com https://github.com/deepseek-ai/DeepSeek-Math

抽象的な

数学的推論は、その複雑で構造化された性質のため、言語モデルにとって大きな課題となります。本稿では、Common Crawl から取得した 1200 億の数学関連トークンと自然言語およびコード データを使用して DeepSeek-Coder-Base-v1.5 7B の事前トレーニングを継続する DeepSeekMath 7B を紹介します。DeepSeekMath 7B は、外部ツールキットや投票技術に頼ることなく、競争レベルの MATH ベンチマークで 51.7% という印象的なスコアを達成し、Gemini-Ultra や GPT-4 のパフォーマンス レベルに近づいています。DeepSeekMath 7B の 64 サンプルにわたる自己一貫性は、MATH で 60.9% を達成しています。

DeepSeekMath の数学的推論能力は、2 つの重要な要素に起因しています。まず、綿密に設計されたデータ選択パイプラインを通じて、公開されている Web データの大きな可能性を活用します。次に、近似ポリシー最適化 (PPO) のバリエーションであるグループ相対ポリシー最適化 (GRPO) を導入し、数学的推論能力を強化すると同時に PPO のメモリ使用量を最適化します。

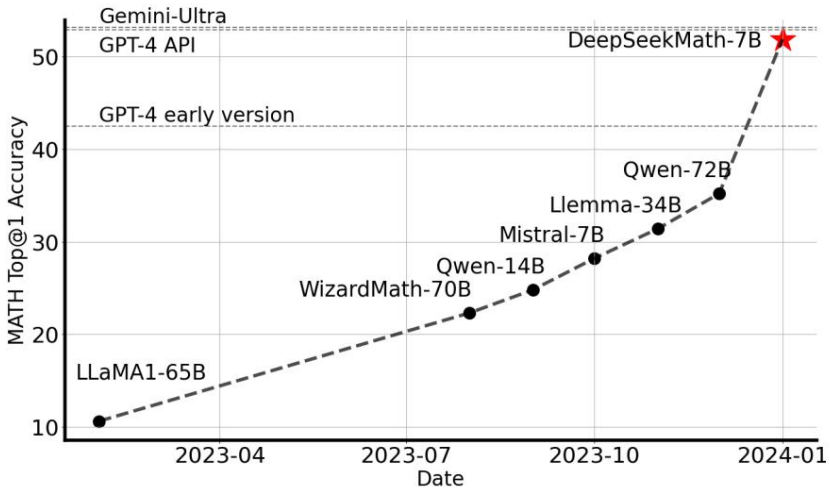

図 1 外部ツールキットや投票技術を使用せずに、競争レベルの MATH ベンチマーク(Hendrycks ら、2021) におけるオープンソース モデルのトップ 1 精度。

コア貢献者。 † DeepSeek-AI でのインターンシップ中に行われた作業。

1. はじめに

大規模言語モデル (LLM) は、人工知能における数学的推論へのアプローチに革命をもたらし、定量的推論ベンチマーク (Hendrycks ら、2021 年) と幾何学的推論ベンチマーク (Trinh ら、2024 年) の両方で大きな進歩を促進しました。
さらに、これらのモデルは、人間が複雑な数学的問題を解決するのを支援する上で役立つことが証明されています (Tao、2023)。ただし、GPT-4 (OpenAI、 2023) や Gemini-Ultra (Anil ら、2023) などの最先端のモデルは公開されておらず、現在アクセス可能なオープンソース モデルはパフォーマンスの面で大幅に遅れをとっています。

本研究では、オープンソースモデルの数学的機能を大幅に上回り、学術ベンチマークでGPT-4のパフォーマンスレベルに近づくドメイン固有言語モデルであるDeepSeekMathを紹介します。これを実現するために、 1200億の数学トークンを含む大規模で高品質の事前トレーニングコーパスであるDeepSeek-Math Corpusを作成します。このデータセットは、fastTextベースの分類器（Joulin et al.、 2016)を使用してCommon Crawl（CC)から抽出されます。最初の反復では、分類器はOpenWebMath（Paster et al.、2023)のインスタンスを正の例としてトレーニングされ、他のさまざまなWebページが負の例として組み込まれます。次に、分類器を使用してCCから追加の正のインスタンスをマイニングし、人間による注釈によってさらに洗練されます。次に、分類器はこの強化されたデータセットで更新され、パフォーマンスが向上します。評価結果によると、大規模コーパスは高品質であり、ベースモデル DeepSeekMath-Base 7B は GSM8K で 64.2% (Cobbe et al., 2021)、競技レベルの MATH データセットで 36.2% (Hendrycks et al., 2021) を達成し、Minerva 540B (Lewkowycz et al., 2022a) を上回っています。さらに、DeepSeekMath コーパスは多言語であるため、中国語の数学ベンチマークで改善が見られます(Wei et al., 2023; Zhong et al., 2023)。数学データ処理における当社の経験は研究コミュニティの出発点であり、将来的には大幅な改善の余地があると考えています。

DeepSeekMath-Base は DeepSeek-Coder-Base-v1.5 7B (Guo et al., 2024) で初期化されます。これは、一般的な LLM と比較して、コード トレーニング モデルから開始する方が優れた選択であることがわかったためです。さらに、数学トレーニングによって MMLU (Hendrycks et al., 2020) および BBH ベンチマーク (Suzgun et al., 2022) のモデル機能も向上することがわかり、モデルの数学的能力を強化するだけでなく、一般的な推論能力も増幅することがわかります。

事前トレーニング後、思考連鎖 (Wei et al., 2022)、思考プログラム (Chen et al., 2022; Gao et al., 2023)、ツール統合推論 (Gou et al., 2023) データを使用して、DeepSeekMath-Base に数学的命令チューニングを適用します。結果として得られるモデル DeepSeekMath-Instruct 7B は、7B のすべての同等モデルを上回り、70B オープンソース命令チューニングモデルに匹敵します。

さらに、近似ポリシー最適化（PPO） （Schulman et al.、2017)の変形強化学習（RL)アルゴリズムであるグループ相対ポリシー最適化（GRPO)を紹介します。
GRPO は批評家モデルを放棄し、代わりにグループスコアからベースラインを推定することで、トレーニングリソースを大幅に削減します。英語の指示チューニングデータのサブセットのみを使用することで、GRPO は、強化学習フェーズ中にドメイン内 (GSM8K: 82.9% → 88.2%、MATH: 46.8% → 51.7%) とドメイン外の数学的タスク (例: CMATH: 84.6% → 88.8%) の両方を含め、強力な DeepSeekMath-Instruct よりも大幅に改善されています。また、 Rejection Sampling Fine-Tuning (RFT) (Yuan et al., 2023a)、Direct Preference Optimization (DPO) (Rafailov et al., 2023)、PPO、GRPO などのさまざまな方法を理解するための統一されたパラダイムも提供します。このような統一されたパラダイムに基づいて、これらすべての方法が直接的または簡略化されたRL技術として概念化されていることがわかりました。また、オンライントレーニングとオフライントレーニング、結果とプロセスの監督、シングルターンと反復RLなど、広範な実験も行っています。

このパラダイムの重要な要素を深く調査します。最後に、なぜ私たちの強化学習が命令調整モデルのパフォーマンスを向上させるのかを説明し、さらにこの統一されたパラダイムに基づいてより効果的な強化学習を実現するための潜在的な方向性をまとめます。

1.1. 貢献

私たちの貢献には、スケーラブルな数学の事前トレーニング、強化学習の調査と分析が含まれます。

大規模な数学の事前トレーニング

- 私たちの研究は、公開されている Common Crawlデータに数学的な目的に有用な情報が含まれているという説得力のある証拠を提供しています。綿密に設計されたデータ選択パイプラインを実装することで、数学的なコンテンツでフィルタリングされた Web ページからの 1200 億トークンの高品質データセットである DeepSeekMath Corpus の構築に成功しました。これは、Minerva (Lewkowycz ら、2022a) が使用する数学 Web ページのサイズのほぼ 7 倍、最近リリースされた OpenWebMath (Paster ら、2023) のサイズの 9 倍です。 •事前トレーニング済みの基本モデル DeepSeekMath-Base 7B は、Minerva 540B (Lewkowycz ら、2022a) と同等のパフォーマンスを実現しており、
パラメーターの数が数学的推論機能の唯一の重要な要素ではないことを示しています。高品質のデータで事前トレーニングされた小規模なモデルでも、優れたパフォーマンスを実現できます。

- 数学トレーニング実験から得られた知見を共有します。数学トレーニングの前にコードトレーニングを行うと、ツールの使用の有無にかかわらず、モデルが数学の問題を解く能力が向上します。これは、コードトレーニングによって推論能力が向上するかという長年の疑問に対する部分的な答えを提供します。少なくとも数学的推論に関しては、向上すると考えています。 • arXiv 論文でのトレーニングは、特に多くの数学関連の論文で一般的ですが、この論文で採用されているすべての数学ベンチマークで顕著な改善はもたらされません。

強化学習の探究と分析

- 効率的で効果的な強化学習アルゴリズムであるグループ相対ポリシー最適化 (GRPO) を紹介します。GRPO は批評モデルを使わず、代わりにグループ スコアからベースラインを推定し、近似ポリシー最適化 (PPO) と比較してトレーニング リソースを大幅に削減します。

- GRPO は、命令チューニングデータのみを使用して、命令チューニングモデル DeepSeekMath-Instructのパフォーマンスを大幅に向上させることを実証しました。さらに、強化学習プロセス中にドメイン外パフォーマンスが向上することも確認しました。

- RFT、DPO、 PPO、GRPOなどのさまざまな方法を理解するための統一されたパラダイムを提供します。また、オンライントレーニングとオフライントレーニング、結果監視とプロセス監視、シングルターン強化学習と反復強化学習など、広範な実験を実施して、このパラダイムの重要な要素を深く調査します。 •統一されたパラダイムに基づいて、強化学習の有効性の背後にある理由を探り、LLMのより効果的な強化学習を実現するためのいくつかの潜在
的な方向性をまとめます。

1.2. 評価と指標の概要

- 英語と中国語の数学推論 :英語と中国語のベンチマークでモデルの包括的な評価を実施し、数学の問題をカバーしています。

小学校レベルから大学レベルまで、幅広い英語のベンチマークテストを実施しています。英語のベンチマークには、GSM8K (Cobbe et al., 2021)、MATH (Hendrycks et al., 2021)、SAT (Azerbayev et al., 2023)、OCW Courses (Lewkowycz et al., 2022a)、MMLU-STEM (Hendrycks et al., 2020) などがあります。中国語のベンチマークには、MGSM-zh (Shi et al., 2023)、CMATH (Wei et al., 2023)、Gaokao-MathCloze (Zhong et al., 2023)、Gaokao-MathQA (Zhong et al., 2023) などがあります。ツールを使用せずに自己完結型のテキストソリューションを生成するモデルの能力と、 Python を使用して問題を解決する能力を評価します。

英語のベンチマークでは、DeepSeekMath-Base はクローズドソースの Minerva 540B (Lewkowycz et al., 2022a) と競合し、数学の事前トレーニングを受けているかどうかに関係なく、すべてのオープンソースの基本モデル (例: Mistral 7B (Jiang et al., 2023) や Llemma-34B (Azerbayev et al., 2023)) を大幅に上回ります。特に、 DeepSeekMath-Base は中国語のベンチマークで優れていますが、これはおそらく、英語のみの数学の事前トレーニング データを収集し、高品質の非英語のデータも含めるという以前の研究 (Azerbayev et al., 2023; Lewkowycz et al., 2022a)に従わないためと考えられます。数学的命令の調整と強化学習により、結果として得られた DeepSeekMath-Instruct と DeepSeekMath-RL は強力なパフォーマンスを発揮し、オープンソース コミュニティ内で初めて、競技レベルの MATH データセットで 50%を超える精度を達成しました。

- 形式数学: 証明アシスタントとして Isabelle (Wenzel et al., 2008) を選択し、miniF2F (Zheng et al., 2021) 上の (Jiang et al., 2022) の非公式から形式への定理証明タスクを使用して DeepSeekMath-Base を評価しました。DeepSeekMath-Base は強力な少数ショットの自動形式化パフォーマンスを示します。

- 自然言語の理解、推論、コード:モデルの一般的な理解、推論、コーディング機能の包括的なプロファイルを構築するために、多様な主題をカバーする 57 の多肢選択タスクを含む Massive Multitask Language Understanding (MMLU) ベンチマーク (Hendrycks 他、2020 年)、主に解決に複数ステップの推論を必要とする 23 の難しいタスクで構成される BIG-Bench Hard (BBH) (Suzgun 他、2022 年)、コード言語モデルの評価に広く使用されている HumanEval (Chen 他、2021 年) および MBPP (Austin 他、2021 年) で DeepSeekMath - Base を評価します。数学の事前トレーニングは、言語理解と推論パフォーマンスの両方に役立ちます。

2. 数学の事前トレーニング

2.1. データ収集と除染

このセクションでは、Common Crawl から DeepSeekMath コーパスを構築するプロセスの概要を説明します。図 2 に示すように、シード コーパス (たとえば、数学関連のデータセットの小規模だが高品質なコレクション) から始めて、Common Crawl から大規模な数学コーパスを体系的に収集する方法を示す反復パイプラインを示します。このアプローチは、コーディングなどの他のドメインにも適用できることは注目に値します。

まず、高品質の数学ウェブテキストのコレクションであるOpenWebMath（Paster et al., 2023)を最初のシードコーパスとして選択します。このコーパスを使用して、fastTextモデル（Joulin et al., 2016)をトレーニングし、OpenWebMathに似た数学ウェブページをより多く思い出せるようにします。具体的には、シードコーパスから500,000のデータポイントをランダムに選択してポジティブトレーニングサンプルとし、 Common Crawlから別の500,000のウェブページをネガティブサンプルとして選択します。トレーニングにはオープンソースライブラリ1を使用し、ベクトル次元を256、学習率を0.1、最大長を1000に設定しました。

1https://fasttext.cc

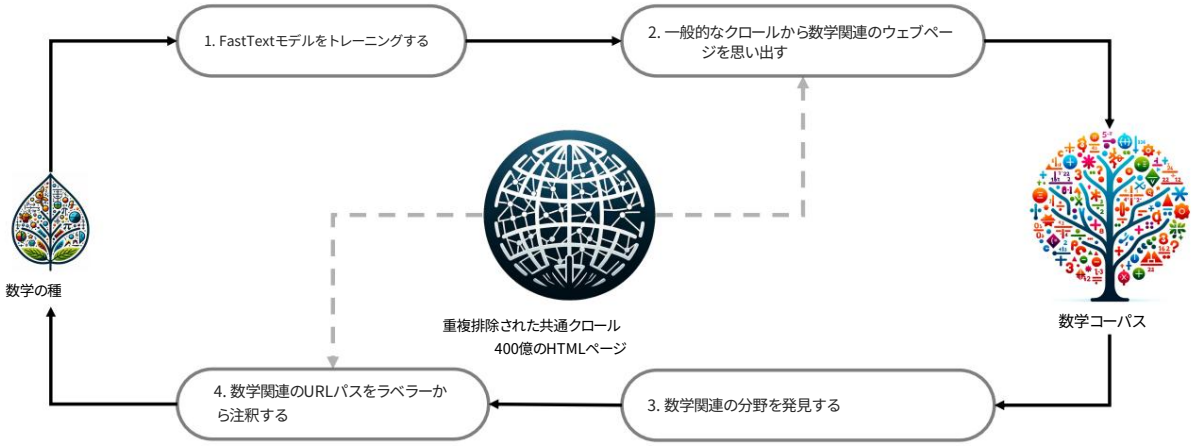

図 2 | Common Crawl から数学的な Web ページを収集する反復パイプライン。

単語 n-gram を 3、単語出現の最小数を 3、トレーニング エポック数を3 に設定しました。元の Common Crawl のサイズを縮小するために、URL ベースの重複排除およびほぼ重複排除の手法を採用し、400 億の HTML Web ページを作成しました。次に、重複排除された Common Crawl から fastText モデルを使用して数学的な Web ページを呼び出します。
低品質の数学コンテンツを除外するために、収集したページを fastText モデルによって予測されたスコアに従ってランク付けし、上位ランクのページのみを保存します。保存されるデータの量は、上位 400 億、800 億、1200 億、および 1600 億のトークンに対する事前トレーニング実験を通じて評価されます。最初の反復では、上位 400 億のトークンを保持することを選択します。

データ収集の最初の反復の後、多くの数学的な Web ページが収集されずに残ります。これは主に、fastText モデルが十分な多様性を欠く一連の正の例でトレーニングされているためです。そのため、シード コーパスを充実させるために追加の数学的な Web ソースを特定し、fastText モデルを最適化できるようにします。具体的には、まず Common Crawl 全体を別々のドメインに整理します。ドメインは、同じベース URL を共有する Web ページとして定義されます。
各ドメインについて、最初の反復で収集された Web ページの割合を計算します。
ウェブページの 10% 以上が収集されているドメインは、数学関連として分類されます(例: mathoverflow.net)。次に、特定されたドメイン内の数学コンテンツに関連付けられた URL (例: mathoverflow.net/questions) に手動で注釈を付けます。

これらの URL にリンクされているがまだ収集されていない Web ページは、シード コーパスに追加されます。このアプローチにより、より多くの肯定的な例を収集できるため、後続の反復でより多くの数学データを思い出すことができる改善された fastText モデルをトレーニングできます。データ収集を 4 回反復した後、3,550 万の数学 Web ページ、合計 1,200 億トークンが完成しました。4 回目の反復では、データのほぼ 98% が3 回目の反復ですでに収集されていることがわかったため、データ収集を停止することにしました。

ベンチマークの汚染を避けるため、Guo et al. (2024) に従い、英語の数学ベンチマークである GSM8K (Cobbe et al., 2021) や MATH (Hendrycks et al., 2021) および中国語ベンチマークである CMATH (Wei et al., 2023) や AGIEval (Zhong et al., 2023) からの質問や回答を含む Web ページを除外します。フィルタリング基準は次のとおりです。評価ベンチマークの任意の部分文字列と完全に一致する 10 グラムの文字列を含むテキストセグメントは、数学トレーニングコーパスから削除されます。10 グラム未満だが 3 グラム以上あるベンチマークテキストについては、完全一致を使用して汚染された Web ページを除外します。

2.2. DeepSeekMathコーパスの品質の検証

DeepSeekMathコーパスがどのように比較されるかを調べるために事前トレーニング実験を実施しました。
最近リリースされた数学トレーニングコーパス:

- MathPile (Wang et al., 2023c): 複数のソースから収集されたコーパス（89億トークン）
  教科書、Wikipedia、ProofWiki、CommonCrawl、StackExchange、arXivなど、
  大部分（85%以上)はarXivから取得されています。
- OpenWebMath (Paster et al., 2023): 数学コンテンツでフィルタリングされたCommonCrawlデータ、
  合計136億トークン。
- Proof-Pile-2 (Azerbayev et al., 2023): OpenWeb-Math、AlgebraicStack (103億トークンの数学コード)、
  arXiv論文 (280億トークン) からなる数学コーパス。Proof-Pile-2の実験では、Azerbayev et al. (2023) に従い、
  arXiv:Web:コードの比率は2:4:1です。

2.2.1. トレーニング設定

我々は、13億個のパラメータを持つ一般的な事前学習済み言語モデルに数学学習を適用し、
DeepSeek LLM（DeepSeek-AI、2024)と同じフレームワークを共有しており、DeepSeek- LLM 1.3Bと表記されています。各数学コーパスで1500億トークンのモデルを個別にトレーニングします。
効率的で軽量なHAI-LLM（High-flyer、2023)を使用して実験を実施します。
トレーニングフレームワーク。DeepSeek LLMのトレーニングプラクティスに従って、AdamW
最適化器（LoshchilovとHutter、2017） = 0.95、 weight_decay = 0.1、
学習率が2,000回後にピークに達する多段階学習率スケジュールを採用
ウォームアップステップは、トレーニングプロセスの80%後に31.6%に減少し、さらに減少します。
トレーニングプロセスの90%後のピークの10.0%。学習率の最大値を設定します。
5.3e-4 に準拠し、4K コンテキスト長で 4M トークンのバッチ サイズを使用します。

数学コーパス	サイズ	英語ベンチマーク							中国のベンチマーク		
		GSM8K	数学	OCW	SAT	MMLU	幹		数学	高考 数学穴埋め	高考 数学QA
数学のトレーニングなし	N/A	2.9%	3.0%	2.9%	15.6%	19.5%	12.3%			0.8%	17.9%
数学パイル	89億	2.7%	3.3%	2.2%	12.5%	15.7%	1.2%			0.0%	2.8%
オープンウェブ数学	136億	11.5%	8.9%	3.7%	31.3%	29.6%	16.8%	519億	14.3%	11.2%	3.7% 0.0% 14.2%
証拠の山-2	43.8%	29.2%	19.9%							5.1%	11.7%
DeepSeekMathコーパス	120.2B	23.8%	13.6%	4.8%	56.3%	33.1%	41.5%			5.9%	23.6%

表1 |さまざまな数学コーパスでトレーニングされたDeepSeek-LLM 1.3Bのパフォーマンス。数ショットの思考連鎖プロンプトを使用して評価。コーパスのサイズは、トークナイザーを使用して計算されます。
語彙数は10万語です。

2.2.2. 評価結果

DeepSeekMathコーパスは高品質で、多言語の数学コンテンツをカバーしており、
サイズは最大です。

- 高品質: 8つの数学的ベンチマークを使用して下流のパフォーマンスを評価します。
  数ショットの思考連鎖を促すWei et al. (2022)。表1に示すように、明確な
  DeepSeekMath Corpusで訓練されたモデルのパフォーマンスリード。図3は、
  DeepSeekMath Corpusで訓練されたモデルは、

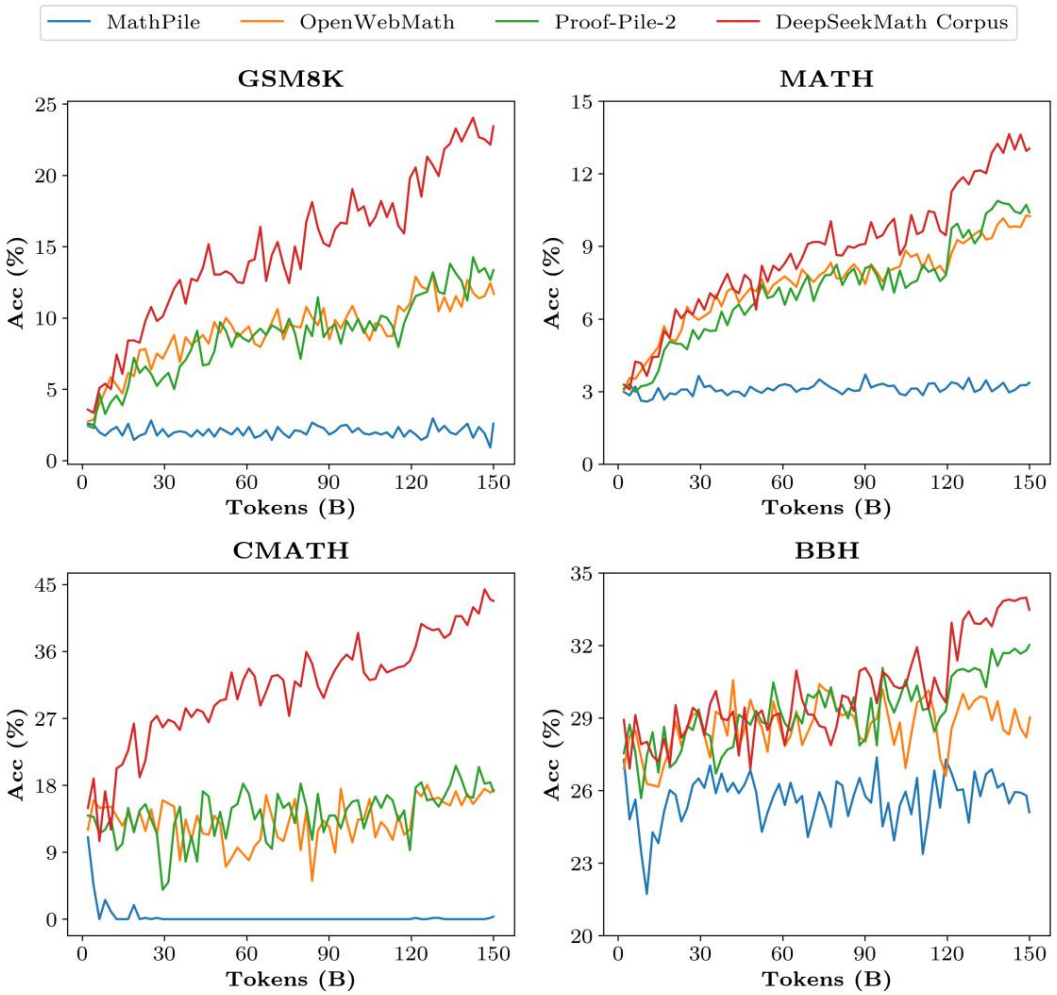

図 3 |さまざまな数学コーパスでトレーニングされた DeepSeek-LLM 1.3B のベンチマーク曲線。

Proof-Pile-2 は 500 億トークン (Proof-Pile-2 の 1 つの完全なエポック) であり、DeepSeekMath Corpus の平均品質が高いことを示しています。

•多言語: DeepSeekMath コーパスには複数の言語のデータが含まれており、最も多く使用されている言語は主に英語と中国語です。表 1 に示すように、DeepSeekMath コーパスでのトレーニングにより、英語と中国語の両方で数学的推論のパフォーマンスが向上します。対照的に、主に英語中心の既存の数学コーパスでは、中国語での数学的推論のパフォーマンスの向上は限られており、パフォーマンスを妨げる可能性もあります。

•大規模: DeepSeekMath コーパスは、既存の数学コーパスよりも数倍大きいです。図 3 に示すように、DeepSeek- Math コーパスでトレーニングされた DeepSeek-LLM 1.3B は、より急峻な学習曲線とより持続的な改善を示しています。対照的に、ベースライン コーパスははるかに小さく、トレーニング中にすでに複数回の繰り返しが行われているため、結果として得られるモデルのパフォーマンスはすぐに頭打ちになっています。

2.3. DeepSeekMath-Base 7B のトレーニングと評価

このセクションでは、特に数学において強力な推論能力を備えた基本モデルであるDeepSeekMath-Base 7Bを紹介します。このモデルはDeepSeek-Coder-Base-v1.5 7Bで初期化されています。

（Guo et al., 2024)500Bトークンでトレーニングしました。データの分布は次のとおりです：56%
1%はDeepSeekMath Corpusから、4%はAlgebraicStackから、10%はarXivから、20%はGithubから
残りの10%はCommon Crawlからの英語と日本語の自然言語データです。
中国語。主にセクション2.2.1で指定されたトレーニング設定を採用しますが、
学習率の最大値は 4.2e-4 に設定し、バッチ サイズは 10M トークンを使用します。

DeepSeekMath-Base 7Bの数学的な能力について、外部の計算に頼らずに自己完結的な数学的ソリューションを生成する能力に焦点を当てて総合的に評価します。
外部ツールについて学び、ツールを使用して数学の問題を解き、正式な定理証明を実施します。
数学以外にも、ベースモデルのより一般的なプロファイルも提供しており、
自然言語理解、推論、プログラミングスキルのパフォーマンス。

段階的な推論による数学的問題解決DeepSeekMathBaseの、数ショットの思考連鎖プロンプト（Wei et al., 2022)を使用した数学的問題の解決パフォーマンスを、英語と中国語の8つのベンチマークで評価しました。これらのベンチマークには、定量的推論（例：GSM8K（Cobbe et al., 2021）、MATH（Hendrycks et al., 2021）、

およびCMATH（Wei et al., 2023）)および多肢選択問題（例：MMLU-STEM（Hendrycks et al., 2020) および Gaokao-MathQA (Zhong et al., 2023))、数学の多様な分野をカバー
初歩的なレベルから大学レベルの複雑さまで。

表2に示すように、DeepSeekMath-Base 7Bは、オープンソースベースモデル（広く使用されている一般モデルMistralを含む)の中で、 8つのベンチマークすべてでパフォーマンスをリードしています。
7B（Jiang et al., 2023)と最近発表されたLlemma 34B（Azerbayev et al., 2023)は、
（Azerbayev et al., 2023）)Proof-Pile-2で数学のトレーニングを受けた。特に、競技レベルのMATHデータセットでは、DeepSeekMath-Baseは既存のオープンソースベースモデルを
10%絶対であり、クローズドソースベースであるMinerva 540B（Lewkowycz et al., 2022a)よりも優れています。
PaLM（Lewkowycz et al., 2022b)に基づいて構築され、さらに訓練された77倍のモデル
数学のテキストについて。

モデル	サイズ	英語ベンチマーク				中国のベンチマーク		
		GSM8K	数学	OCW	SAT	MMLU 幹	数学	高考 数学穴埋め	高考 数学QA
クローズドソースベースモデル									
ミネルヴァ	7B	16.2%	14.1%	7.7%	-	35.6%	-	-	-
ミネルヴァ	620億	52.4%	27.6%	12.0%	-	5400億 58.8%	-	-	-
ミネルヴァ		33.6%	17.6%	-		63.9%	-	-	-
オープンソースベースモデル									
ミストラル	7B	40.3%	14.3%	9.2%	71.9%	51.1%	44.9%	5.1%	23.4%
レマ	7B	37.4%	18.1%	6.3%	59.4%	43.1%	43.4%	11.9%	23.6%
レマ	34B	54.0%	25.3%	10.3%	71.9%	52.9%	56.1%	11.9%	26.2%
DeepSeekMath ベース	7B	64.2%	36.2%	15.4%	84.4%	56.5%	71.7%	20.3%	35.3%

表2 | DeepSeekMath-Base 7Bと英語およびフランス語の強基本モデルの比較
中国の数学ベンチマーク。モデルは思考連鎖プロンプトで評価されます。
Minervaの結果はLewkowycz et al. (2022a)から引用されています。

ツールを使った数学の問題解決プログラム支援による数学の問題解決を評価します
GSM8KとMATHでの推論を、数回の思考プログラムプロンプトを使用して行う（Chen et al.、
モデルは、Pythonプログラムを記述することで各問題を解くように促される。
複雑な計算にはmathやsympyなどのライブラリを利用できます。実行
プログラムの結果が答えとして評価される。表3に示すように、DeepSeekMath-Base 7B
従来の最先端のLlemma 34Bよりも優れた性能を発揮します。

モデル	サイズ	ツールを使った問題解決		非公式から公式への証明	
		GSM8K+Python	MATH+Python	miniF2F 有効	miniF2F テスト
ミストラル	7B	48.5%	18.2%	18.9%	18.0%
コードラマ	7B	27.1%	17.2%	16.3%	17.6%
コードラマ	34B	52.7%	23.5%	18.5%	18.0%
レマ	7B	41.0%	18.6%	20.6%	22.1%
レマ	34B	64.6%	26.3%	21.0%	21.3%
DeepSeekMath ベース 7B		66.9%	31.4%	25.8%	24.6%

表3 |ツールを使用して数学的問題を解く基本モデルの能力の少数ショット評価
Isabelle で非公式から公式までの定理証明を実行する機能。

形式数学形式証明の自動化は、数学的証明の正確性と信頼性を確保し、効率性を高めるのに有益であり、近年注目を集めています。
我々は、DeepSeekMath-Base 7Bを、(Jiang et al.,
2022年)は、非公式な声明に基づいて正式な証明を生成するものであり、正式な対応物である
ステートメントの公式と非公式の証明。我々は、オリンピックレベルの数学のベンチマークであるminiF2F（Zheng
et al., 2021)で評価し、Isabelleで各ステートメントの公式の証明を生成する。
少数ショットのプロンプトの問題。Jiang et al. (2022) に従って、モデルを活用して生成します。
証明スケッチを作成し、市販の自動証明器 Sledgehammer (Paulson, 2010) を実行する
不足している詳細を埋めるために、表3に示すように、DeepSeekMath-Base 7Bは強力な
証明の自動形式化におけるパフォーマンス。

モデル	サイズ	MMLU	BBH	HumanEval (Pass@1)	MBPP (Pass@1)
ミストラル	7B	62.4%	55.7%	28.0%	41.4%
DeepSeek-Coder-Base-v1.5 †	7B	42.9%	42.9%	40.2%	52.6%
DeepSeek-Coder-Base-v1.5	7B	49.1%	55.2%	43.2%	60.4%
DeepSeekMath-Base	7B	54.9%	59.5%	40.9%	52.6%

表 4 |自然言語理解、推論、コードベンチマークの評価。
DeepSeek-Coder-Base-v1.5 †は学習率の減衰直前のチェックポイントであり、
DeepSeekMath-Base をトレーニングします。MMLU と BBH では、数回の思考連鎖プロンプトを使用します。
HumanEvalとMBPPでは、ゼロショット設定と
それぞれ少数ショット設定です。

自然言語理解、推論、コードモデルのパフォーマンスを評価します
MMLU（Hendrycks et al., 2020)による自然言語理解、BBH（Suzgun
et al., 2022)、HumanEval (Chen et al., 2021) および MBPP (Austin et al.,

表4に示すように、DeepSeekMath-Base 7Bは、その前身であるDeepSeek-Coder-Base-v1.5（Guo et al., 2024)と比較して、MMLUとBBHのパフォーマンスが大幅に向上しており、数学のトレーニングが言語の理解と推論にプラスの影響を与えることを示しています。

さらに、継続的なトレーニング用のコード トークンを含めることで、DeepSeekMath-Base 7B は、 2 つのコーディング ベンチマークで DeepSeek-Coder-Base-v1.5 のパフォーマンスを効果的に維持します。全体として、DeepSeekMath-Base 7B は、3 つの推論およびコーディング ベンチマークで一般モデル Mistral 7B (Jiang ら、 2023) を大幅に上回ります。

3. 教師あり微調整

3.1. SFTデータのキュレーション

私たちは、さまざまな数学分野からさまざまな複雑さのレベルの英語と中国語の問題をカバーする数学の指導調整データセットを構築しました。問題は、思考連鎖 (CoT) (Wei et al., 2022)、思考プログラム (PoT) (Chen et al., 2022; Gao et al., 2023)、およびツール統合推論形式 (Gou et al., 2023) でソリューションとペアになっています。トレーニング例の総数は776K です。

- 英語の数学データセット： GSM8K および MATH の問題にツール統合ソリューションを注釈付けし、問題がCoTまたは PoT で解決される Lila-OOD (Mishra et al., 2022) のトレーニング セットとともに MathInstruct (Yue et al., 2023) のサブセットを採用しています。英語のコレクションは、代数、確率、数論、微積分、幾何学など、数学のさまざまな分野をカバーしています。 •中国語の数学データセット：線形方程式などの 76 のサブトピックにまたがる中国語の K-12 数学の問題を収集し、ソリューションは CoT とツール統合推論形式の両方で注釈付けされています。

3.2. DeepSeekMath-Instruct 7Bのトレーニングと評価

このセクションでは、 DeepSeekMath-Base に基づいて数学命令のチューニングを行った DeepSeekMath-Instruct 7B を紹介します。トレーニング サンプルは、最大コンテキスト長 4K トークンに達するまでランダムに連結されます。バッチサイズ 256、一定の学習率 5e-5で 500 ステップにわたってモデルをトレーニングします。

私たちは、英語と中国語の 4 つの定量的推論ベンチマークで、ツールの使用の有無にかかわらずモデルの数学的パフォーマンスを評価します。私たちは、当時の主要なモデルに対してモデルをベンチマークします。

- クローズドソースのモデルには、(1) GPTファミリー（その中で最も優れたGPT-4（OpenAI、2023)とGPT-4コードインタープリター）[2]、(2) Gemini UltraとPro（Anil et al.、2023）、(3) Inflection-2（Inflection AI、2023）、(4) Grok-1 [3]、および中国企業が最近リリースしたモデル[5] （5)Baichuan-3、(6)最新のGLM-4 GLMファミリー（Du et al.、2022)が含まれます。これらのモデルは汎用性が高く、そのほとんどは一連の調整手順を経ています。 •オープンソースモデルには、(1) [4] DeepSeek-LLM-Chat 67B (DeepSeek- AI、2024)、(2) Qwen 72B (Bai et al.、2023)、(3) SeaLLM-v2 7B (Nguyen et al.、2023)、(4) などの一般的なモデルが含まれます。

2https://openai.com/blog/chatgpt-plugins#コードインタープリター
3https://x.ai/model-card
4https://www.baichuan-ai.com
5https://open.bigmodel.cn/dev/api#glm-4

ChatGLM3 6B（ChatGLM3チーム、2023年）、およびInternLM2に基づいて構築され、数学の強化を受けた（5)InterLM2-Math 20Bトレーニングとそ 数学6のモデル。れに続く指導の調整、(6)PPOを適用するMath-Shepherd-Mistral 7Bを含むics 訓練（Schulman et al., 2017)からミストラル7B（Jiang et al., 2023)まで、プロセス監視型報酬モデル、(7)数学的学習を改善するWizardMathシリーズ（Luo et al., 2023）Mistral 7BとLlama-2 70B（Touvron et al., 2023)におけるevolve-instruct（すなわち、AI進化命令を使用した命令チューニングのバージョン)とPPOトレーニング訓練問題は主にGSM8KとMATHから入手した。(8) MetaMath 70B (Yu et al., 2023年)は、GSM8KとMATHの拡張バージョンで微調整されたLlama-2 70Bです。（9)ToRA 34B Gou et al.（2023)は、ツール統合を行うために微調整されたCodeLlama 34Bです。数学的推論、(10)MAmmoTH 70B（Yue et al., 2023)はLlama-2 70Bである MathInstruct で命令調整されています。

表5に示すように、ツールの使用が禁止されている評価設定では、DeepSeekMath- Instruct 7Bはステップバイステップの推論において優れたパフォーマンスを示しています。特に、競争レベルのMATHデータセットでは、私たちのモデルはすべてのオープンソースモデルとほとんどの独自モデル（Inflection-2やGemini Proなど)を少なくとも9%上回っています。これは、かなり大規模なモデル（例：Qwen 72B)や、数学に重点を置いた強化学習によって特別に強化されたモデル（例：WizardMath-v1.1 7B)にも当てはまります。DeepSeekMath-Instructは、MATHにおいて中国独自のモデルGLM-4やBaichuan-3に匹敵する。それでも、GPT-4 や Gemini Ultra よりパフォーマンスは劣ります。

モデルが自然言語推論とプログラムベースのツールの使用を統合して問題解決を行うことが許可されている評価設定では、DeepSeekMath-Instruct 7Bは次のようなアプローチを採用しています。MATHの精度は60%で、既存のオープンソースモデルを上回っています。他のベンチマークでは、私たちのモデルは、先行する最先端のDeepSeek-LLM-Chat 67Bと競合しています。10倍の大きさです。

4. 強化学習

4.1. グループ相対ポリシー最適化

強化学習（RL)は、教師あり微調整（SFT)段階の後にLLMの数学的推論能力をさらに向上させるのに効果的であることが証明されています（Luo et al.、2023; 王ら、2023b）。このセクションでは、効率的で効果的なRLアルゴリズムであるグループ相対ポリシー最適化 (GRPO)。

4.1.1. PPO から GRPO へ

近接ポリシー最適化（PPO）（Schulman et al., 2017)は、アクター•クリティックRLアルゴリズムであり、LLMのRL微調整段階で広く使用されています（Ouyang et al., 2022）。特に、次の代理目標を最大化することにより、LLM を実現します。

$$J(\,) = E[\quad (\,), \qquad (|\,)]\,|\ \frac{1}{|}\sum_{=1}^{|\,|} 分 \quad \frac{(|、<)}{(|、<)} \ 、クリップ \ \frac{|、<)}{((\ |、<)} 、1- \ 、1+ \qquad 、 \qquad (1)$$

ここで、およびは現在の政策モデルと古い政策モデルであり、およびは質問と出力である。、質問データセットと古いポリシーからそれぞれサンプリングされた。クリッピング関連の PPOで訓練を安定させるために導入されたハイパーパラメータ。これは利点であり、一般化優位性推定（GAE）（Schulman et al., 2015)を適用して計算され、

6https://github.com/InternLM/InternLM-Math

モデル	サイズ	英語ベンチマーク		中国語ベンチマーク	
		GSM8K	数学	MGSM-zh	CMATH
思考連鎖推論					
クローズドソースモデル					
Gemini Ultra	-	94.4%	53.2%	-	-
GPT-4	-	92.0%	52.9%	-	86.0%
Inflection-2	-	81.4%	34.8%	-	-
GPT-3.5	-	80.8%	34.1%	-	73.8%
Gemini Pro	-	86.5%	32.6%	-	-
Grok-1	-	62.9%	23.9%	-	-
白川-3	-	88.2%	49.2%	-	-
GLM-4	-	87.6%	47.9%	-	-
オープンソースモデル					
InternLM2-Math	20B	82.6%	37.7%	-	-
Qwen	72B	78.9%	35.2%	-	-
Math-Shepherd-Mistral	7B	84.1%	33.0%	-	-
WizardMath-v1.1	7B	83.2%	33.0%	-	-
DeepSeek-LLM-Chat	67B	84.1%	32.6%	74.0%	80.3%
MetaMath	70B	82.3%	26.6%	66.4%	70.9%
SeaLLM-v2	7B	78.2%	27.5%	64.8%	-
ChatGLM3	6B	72.3%	25.7%	-	-
WizardMath-v1.0	70B	81.6%	22.7%	64.8%	65.4%
ディープシーク数学-命令	7B	82.9%	46.8%	73.2%	84.6%
ディープシーク数学-RL	7B	88.2%	51.7%	79.6%	88.8%
ツール統合推論					
クローズドソースモデル					
GPT-4 コードインタープリター	-	97.0%	69.7%	-	-
オープンソースモデル					
インターンLM2-数学	200億	80.7%	54.3%	-	-
DeepSeek-LLM-チャット	670億	86.7%	51.1%	76.4%	85.4%
トラ	340億	80.7%	50.8%	41.2%	53.4%
マンモス	700億	76.9%	41.8%	-	-
ディープシーク数学-命令	7B	83.7%	57.4%	72.0%	84.3%
ディープシーク数学-RL	7B	86.7%	58.8%	78.4%	87.6%

表5 |思考連鎖とクローズドソースモデルの両方を使用したオープンソースモデルとクローズドソースモデルのパフォーマンス英語と中国語のベンチマークにおけるツール統合推論。灰色のスコアは多数派を表す。32の候補から投票。その他はトップ1のスコアです。DeepSeekMath-RL 7Bは、7Bから70Bまでのすべてのオープンソースモデルと、クローズドソースモデルの大多数に勝っています。

DeepSeekMath-RL 7Bは、思考連鎖形式の命令チューニングデータでのみさらにトレーニングされます。

GSM8K と MATH では、すべてのベンチマークで DeepSeekMath-Instruct 7B よりも性能が向上しています。

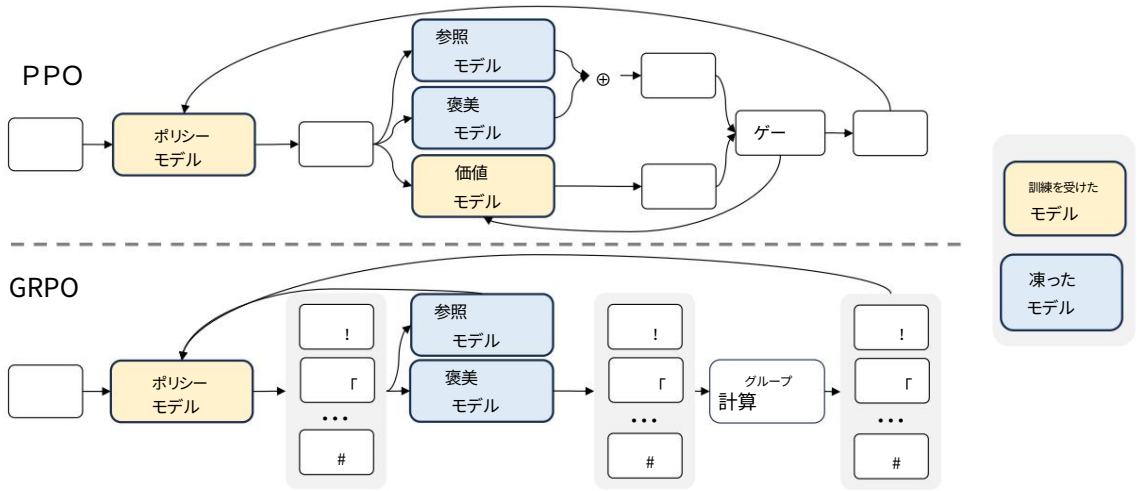

図4 | PPOとGRPOのデモンストレーション。GRPOは価値モデルを放棄し、代わりに
グループスコアからベースラインを推定し、トレーニングリソースを大幅に削減します。

報酬{ ≥ }と学習された価値関数に基づいている。したがって、PPOでは、価値関数は
政策モデルと並行して訓練され、報酬モデルの過剰最適化を緩和する。
標準的なアプローチは、参照モデルからトークンごとのKLペナルティを報酬に加えることである。
各トークン（Ouyang et al., 2022）、すなわち、

$$= (, \leq) - \log \frac{(| 、 <)}{(| 、 <)} 、 \quad (2)$$

ここで は報酬モデル、は KL ペナルティ　　参照モデルは、通常は初期のSFTモデルです。
の係数です。

PPOで使用される価値関数は通常、同等のサイズの別のモデルであるため、
ポリシーモデルでは、かなりのメモリと計算負荷がかかります。さらに、
RLトレーニング中、価値関数は優位性の計算におけるベースラインとして扱われる。
分散の削減のために。LLMコンテキストでは、通常、最後のトークンにのみ
報酬モデルによる報酬スコアは、価値関数の訓練を複雑にする可能性がある。
それぞれのトークンで正確である。これに対処するために、図4に示すように、グループ相対ポリシーを提案する。
最適化（GRPO)により、追加の価値関数近似の必要性が排除される。
PPOでは、代わりに、応答として生成された複数のサンプル出力の平均報酬を使用します。
同じ質問に対して、ベースラインとして、GRPOは各質問に対して、
古いポリシーからの出力のグループ{ 1, 2, 、・・}を選択し、ポリシーモデルを最適化する

以下の目標を最大化することにより：

$$J ( ) = E[\quad ( ), \{ \} \quad {}_{=1} \sim \quad (|)]$$

$$\frac{1}{} \sum_{=1} \frac{1}{| |} \sum_{=1}^{| |} 分 \frac{、| 、、<)}{((、| 、、<)} 、_{クリップ} \frac{、| 、、<)}{((、| 、、<)} 、1 - 、1 + \quad 、 \quad ーダ \quad 、 \quad (3)$$

ここで、およびはハイパーパラメータであり、は相対的な、
GRPOが利点を計算するために利用するグループ相対的な方法は、各グループ内の出力の報酬の
みであり、次のサブセクションで詳しく説明します。
報酬モデルはデータセットで訓練されることが多いため、報酬モデルの比較的性質がある。
同じ質問に対する出力の比較。また、KLを追加する代わりに、
GRPOは報酬にペナルティを加えることで正規化を行い、
訓練されたポリシーと参照ポリシーを損失に適用することで、計算の複雑さを回避します。　　、。

アルゴリズム 1反復グループ相対ポリシー最適化

入力初期ポリシーモデル;報酬モデル;タスクプロンプトD;ハイパーパラメータ、 1:ポリシーモデル← 2:反復= 1、 ...、I do 、

初期化

3: 参照モデル ←ステップ = 1, . . . , M

4: の場合

5: DからバッチDをサンプリングする

6: 古いポリシーモデルを更新 ← サンプル出力 { }

7: =1 (· |) は、各質問 ∈ D について、グループの相

8: 報酬を計算する { } =1 対的優位性の推定を通じて、番目のトークンに対し

9: GRPO反復 = 、て実行することにより、各サンプリングされた出力に対して実行されます。

10: 1、 ...、を計算し、

11: GRPO目標を最大化するように政策モデルを更新する（式21）

12:リプレイメカニズムを使用した継続的なトレーニングを通じて更新します。

出力

また、（２)で使用したKLペナルティ項とは異なり、KLダイバージェンスは
次の不偏推定量（Schulman、2020）：

$$デ = \frac{、|、、<)}{((、|、、<)} - ログ \frac{、|、、<)}{((、1, | , ,<)} - \quad (4)$$

必ずプラスになるはずです。

4.1.2. GRPOによる成果監督RL

形式的には、各質問に対して、出力のグループ{ 1, 2, ・・・ }が古いものからサンプリングされる。
ポリシーモデル報酬モデルは出力にスコアを付け、報酬を生み出すために使われる。
r = { 1, 2, ・・・ , }となる。その後、これらの報酬は減算して正規化される。
グループ平均をグループ標準偏差で割ったもの。アウトカム監視は
各出力の終わりに正規化された報酬を設定し、すべてのトークンの利点を設定します。 、
出力は正規化された報酬、すなわち式(3)で定義された 、 $= = \frac{-平均(r)}{std(r)、}$ そして、政策を最適化するために
目的を最大化する。

4.1.3. GRPO によるプロセス監視 RL

アウトカム監督は、各アウトプットの最後に報酬を与えるだけであり、
複雑な数学的課題における政策を監督するのに十分かつ効率的である。王氏に続いて
（2023b)では、プロセスの最後に報酬を与えるプロセス監督についても検討しています。
各推論ステップ。形式的には、質問とサンプル出力{ 1, 2, ・・・ , }が与えられた場合、
プロセス報酬モデルは出力の各ステップにスコアを付け、対応する報酬を生み出すために使用されます。
R = {{は $_1$ $^{(1)}$、 $^{\cdots}$ 、 $_1$ $^{(1)}$ }, ・・・ , { $^{(1)}$、 $^{\cdots}$ 、 $^{(\ )}$ }}、どこ ( )は終了トークンインデックスです
-番目のステップのものであり、 -番目の出力におけるステップの総数である。これらも正規化する。
平均と標準偏差で報酬を与える、つまりプロセス監視は各トークンの $^{()} = \frac{^{()}-平均(R)}{標準(R)}$ その後、
利点を正規化された平均と標準偏差の合計として計算する。
次のステップから報酬を得て、次に政策を最適化する。$^{()}$、 ( ) ≥
式(3)で定義された目的を最大化する。

4.1.4. GRPO による反復強化学習

強化学習のトレーニング プロセスが進むにつれて、古い報酬モデルでは現在のポリシー モデルを監督するのに十分ではなくなる可能性があります。そのため、 GRPO を使用した反復 RL も検討します。アルゴリズム 1 に示すように、反復 GRPO では、ポリシー モデルからのサンプリング結果に基づいて報酬モデルの新しいトレーニング セットを生成し、履歴データの 10% を組み込んだ再生メカニズムを使用して古い報酬モデルを継続的にトレーニングします。次に、参照モデルをポリシー モデルとして設定し、新しい報酬モデルを使用してポリシー モデルを継続的にトレーニングします。

4.2. DeepSeekMath-RLのトレーニングと評価

DeepSeekMath-Instruct 7B に基づいて RL を実行します。RL のトレーニング データは、約 144K の質問で構成される SFT データからの GSM8K および MATH に関連する連鎖思考形式の質問です。RL フェーズ全体でデータが不足しているベンチマークに対する RL の影響を調査するため、他の SFT の質問を除外します。報酬モデルのトレーニング セットは、(Wang et al., 2023b) に従って構築します。初期報酬モデルは、学習率 2e-5 で DeepSeekMath-Base 7B に基づいてトレーニングします。GRPO の場合、ポリシー モデルの学習率を1e-6 に設定します。KL 係数は 0.04 です。質問ごとに、64 個の出力をサンプリングします。最大長は 1024 に設定され、トレーニング バッチ サイズは 1024 です。ポリシー モデルは、各探索段階の後に 1 回のみ更新されます。DeepSeekMath-RL 7B を、DeepSeekMath-Instruct 7B に続くベンチマークで評価します。DeepSeekMath-RL 7B の場合、GSM8K と思考連鎖推論を備えた MATH はドメイン内タスクと見なすことができ、他のすべてのベンチマークはドメイン外タスクと見なすことができます。

表 5 は、英語と中国語のベンチマークで、思考の連鎖とツール統合推論の両方を備えたオープンソース モデルとクローズド ソース モデルのパフォーマンスを示しています。次のことがわかります。1) DeepSeekMath-RL 7B は、思考の連鎖推論を利用して、GSM8K と MATH でそれぞれ 88.2% と 51.7% の精度を達成しています。このパフォーマンスは、7B から 70B の範囲のすべてのオープンソースモデル、および大多数のクローズド ソース モデルの性能を上回っています。2) 重要なのは、DeepSeekMath-RL 7B は、DeepSeekMath-Instruct 7B から始まる、GSM8K と MATH の思考の連鎖形式の命令チューニング データでのみトレーニングされていることです。トレーニング データの範囲が制限されているにもかかわらず、すべての評価基準で DeepSeekMath-Instruct 7B を上回っており、強化学習の有効性が示されています。

5. 議論

このセクションでは、事前トレーニングと RL 実験で得られた結果を共有します。

5.1. 事前トレーニングで学んだ教訓

まず、事前トレーニングの経験を共有します。特に指定がない限り、セクション 2.2.1 で概説されているトレーニング設定に従います。このセクションで DeepSeekMath Corpus を参照する場合、データ収集プロセスの 2 回目の反復から 890 億トークンのデータセットを使用していることに注意してください。

5.1.1. コードトレーニングは数学的推論に役立つ

よく知られているが検証されていない仮説では、コードトレーニングは推論力を向上させると示唆している。私たちは、特に数学の分野で、これに対する部分的な回答を提供しようとしている。コードトレーニング

トレーニング設定	トレーニングトークン			ツールを使用しない		ツール使用あり	
	一般コード	数学	GSM8K	MATH	CMATH	GSM8K+Python	MATH+Python
継続的なトレーニングなし	–	–	–	2.9%	3.0%	12.3%	2.7%	2.3%
2段階トレーニング							
ステージ1: 一般トレーニング	400B	–	–	2.9%	3.2%	14.8%	3.3%	2.3%
ステージ2: 数学トレーニング	–	–	150B	19.1%	14.4%	37.2%	14.3%	6.7%
ステージ1: コードトレーニング	–	400B	–	5.9%	3.6%	19.9%	12.4%	10.0%
ステージ2: 数学トレーニング	–	–	150B	21.9%	15.3%	39.7%	17.4%	9.4%
ワンステージトレーニング							
数学トレーニング	–	–	150B	20.5%	13.1%	37.6%	11.4%	6.5%
コードと数学の混合トレーニング	–	4000億	1500億	17.6%	12.1%	36.3%	19.7%	13.5%

表6 |異なるトレーニングにおけるコードが数学的推論にどのように影響するかの調査
設定。DeepSeek-LLM 1.3Bを試し、その数学的推論を評価します。
数ショットの思考連鎖プロンプトと数ショットの思考連鎖プロンプトによるツール使用の有無のパフォーマンス
それぞれ思考プログラムの促進です。

ツールの使用の有無にかかわらず、モデルの数学的推論能力が向上します。

コードのトレーニングが数学的推論にどのような影響を与えるかを調べるために、私たちは
次の 2 段階トレーニングと 1 段階トレーニングの設定:

2段階トレーニング

- 400Bトークンのコードトレーニング→ 150Bトークンの数学トレーニング: DeepSeekをトレーニングします。
  400B コード トークンとそれに続く 150B 数学トークンの LLM 1.3B。
- 400Bトークンの一般トレーニング→ 150Bトークンの数学トレーニング :コントロールとして
  実験では、一般的なトークン（大規模な一般的なトークンからサンプリングしたもの)も実験します。
  トレーニングの最初の段階では、コードトークンの代わりにDeepSeek-AIによって作成されたコーパスを使用して、
  コードトークンが一般トークンよりも優れている点を調査し、改善を試みる
  数学的推論。

ワンステージトレーニング

- 1500億トークンの数学トレーニング: 1500億の数学トークンに対してDeepSeek-LLM 1.3Bをトレーニングします。
- 400Bコードトークンと150B数学トークンの混合トレーニング :コードトレーニングの後に数学トレーニングを行うと、コーディングのパフォーマンスが低下します。コードトークン、
  1段階のトレーニングで数学トークンと混ぜると、数学のスキルは向上する。
  推論能力を高め、破滅的な忘却の問題も軽減します。

結果表6と表7は、異なるトレーニングにおける下流のパフォーマンスを示しています。
設定。

コードトレーニングは、2段階のプログラム支援による数学的推論に役立ちます。
トレーニングと1段階トレーニングの設定。表6に示すように、2段階トレーニングでは
設定により、コードトレーニングだけでもGSM8Kを解決する能力が大幅に向上し、
Python を使用した数学の問題。第 2 段階の数学トレーニングにより、さらなる向上が実現します。
興味深いことに、1段階のトレーニング設定では、コードトークンと数学トークンを混ぜると、 2段階のトレーニングから生じる壊滅的な忘却の問題が効果的に軽減され、また、
コーディング（表7)とプログラム支援による数学的推論（表6)を相乗効果で組み合わせます。

トレーニング設定	トレーニングトークン			MMLU	BBH	HumanEval (Pass@1)	MBPP (Pass@1)
	一般的なコード	数学					
継続的なトレーニングなし	–	–	–	24.5%	28.1%	12.2%	13.0%
2段階トレーニング							
ステージ1: 一般トレーニング	400B	–	–	25.9%	27.7%	15.2%	13.6%
ステージ2: 数学トレーニング	–	–	150B	33.1%	32.7%	12.8%	13.2%
ステージ1: コードトレーニング	–	400B	–	25.0%	31.5%	25.0%	40.0%
ステージ2: 数学トレーニング	–	–	150B	36.2%	35.3%	12.2%	17.0%
ワンステージトレーニング							
数学トレーニング	–	–	150B	32.3%	32.5%	11.6%	13.2%
コードと数学の混合トレーニング	–	4000億	1500億	33.5%	35.6%	29.3%	39.4%

表7 |コードと数学のトレーニングの異なる設定が言語理解、推論、コーディングのモデルのパフォーマンスにどのように影響するかの調査。DeepSeek-LLMで実験します。
1.3B. 少量の思考連鎖プロンプトを使用して、MMLU と BBH のモデルを評価します。
HumanEval と MBPP では、それぞれゼロショット評価と少数ショット評価を実施します。

モデル	ArXivコーパスのサイズ	英語ベンチマーク						中国のベンチマーク		
		GSM8K	数学	OCW	SAT	MMLU	幹	数学	高考 数学穴埋め	高考 数学QA
ディープシークLLM	13億	数学の訓練なし	2.9%	3.0%	2.9%	15.6%	19.5%	12.3%	0.8%	17.9%
		MathPile	2.7%	3.3%	2.2%	12.5%	15.7%	1.2%	0.0%	2.8%
		ArXiv-RedPajama	3.3%	3.4%	4.0%	9.4%	9.0%	7.4%	0.8%	2.3%
		数学トレーニングなし	29.0%	12.5%	6.6%	40.6%	38.1%	45.9%	5.9%	21.1%
ディープシーク コーダ ベース v1.5 7B		MathPile	23.6%	11.5%	7.0%	46.9%	35.8%	37.9%	4.2%	25.6%
		ArXiv-RedPajama	28.1%	11.1%	7.7%	50.0%	35.2%	42.6%	7.6%	24.8%

表8 |異なるarXivデータセットにおける数学トレーニングの効果。モデルのパフォーマンスは次のように評価されます。
数回のショットで思考の連鎖を促す。

ArXiv コーパス	miniF2F 有効	miniF2F テスト
数学の訓練なし	20.1%	21.7%
数学パイル	16.8%	16.4%
ArXiv-RedPajama	14.8%	11.9%

表 9 |さまざまな arXiv コーパスに対する数学トレーニングの効果。ベース モデルは DeepSeek- Coder-Base-v1.5 7B です。Isabelle で非公式から公式への証明を評価します。

コードトレーニングは、ツールを使わずに数学的推論力を向上させる。2段階のトレーニング設定では、コード トレーニングの初期段階ですでに中程度の機能強化が得られます。また、その後の数学のトレーニングの効率も向上し、最終的には最高のパフォーマンス。しかし、コードトークンと数学トークンを1段階のトレーニングに組み合わせると、ツールを使用せずに数学的推論を実行できなくなる。1つの推測は、DeepSeek-LLM 1.3B、規模が限られているため、コードと数学的データの両方を完全に統合する能力が欠けている同時に。

5.1.2. ArXiv論文は数学的推論力の向上には効果がないようだ

ArXiv論文は、数学の事前トレーニングデータの構成要素としてよく含まれています（アゼルバイエフ et al., 2023; Lewkowycz et al., 2022a; Polu and Sutskever, 2020; Wang et al., 2023c)。しかし、

数学的推論への影響に関する詳細な分析は、広範囲に行われていません。直感に反するかもしれませんが、私たちの実験によると、arXivの論文は数学的推論の改善には効果がないようです。私たちは、さまざまな処理パイプラインを経たarXivコーパスを使用して、DeepSeek-LLM 1.3BやDeepSeek-Coder-Base-v1.5 7B（Guo et al.、2024)を含むさまざまなサイズのモデルで実験しました。

- MathPile (Wang et al., 2023c): クリーニングとフィルタリングを施して開発された89億トークンのコーパス
  ヒューリスティックルール（その85%以上は科学的なarXiv論文です）
- ArXiv-RedPajama （コンピュータ、2023年）：arXiv LaTeXファイル全体とプリアンブル、
  コメント、マクロ、参考文献が削除され、合計 280 億トークンになりました。

実験では、各 arXiv コーパスで DeepSeek-LLM 1.3B を 1500 億トークン、DeepSeek-Coder-Base-v1.5 7B を 400 億トークンで別々にトレーニングしました。arXiv 論文は数学的推論の改善には効果がないようです。arXiv のみのコーパスでトレーニングした場合、両方のモデルは、この研究で採用されているさまざまな複雑度のさまざまな数学ベンチマークで顕著な改善は見られず、悪化さえも示しませんでした。これらのベンチマークには、 GSM8K や MATH などの定量的推論データセット (表 8)、MMLU-STEM などの多肢選択チャレンジ (表8)、miniF2F などの形式数学 (表 9) が含まれます。

しかし、この結論には限界があり、鵜呑みにしてはいけません。私たちはまだ以下のことを研究していません。

- arXiv トークンが、この研究には含まれていない特定の数学関連のタスク（定理の非公式化、つまり正式な記述や証明を非公式のバージョンに変換することなど)に与える影響。

- arXiv トークンを他の種類のデータと組み合わせた場合の効果。 • arXiv 論文の利点がより
  大きなモデル規模で現れるかどうか。

したがって、さらなる調査が必要であり、それは将来の研究に残されます。

5.2. 強化学習の洞察

5.2.1. 統一パラダイムに向けて

このセクションでは、SFT、RFT、DPO、PPO、GRPO などのさまざまなトレーニング方法を分析するための統一パラダイムを提供し、さらに統一パラダイムの要因を調査するための実験を実施します。一般に、トレーニング方法のパラメーターに関する勾配は次のように記述できます。

$$\nabla_{\theta} J_{\mathcal{A}}(\theta) = \mathbb{E}[(q,o) \sim \mathcal{D}] \left(\frac{1}{|o|} \sum_{t=1}^{|o|} \mathcal{A}(o, q, t, \pi_{\theta_{old}}) \nabla_{\theta} \log \pi_{\theta}(o_t | q, o_{<t}) \right). \tag{5}$$

3つの主要なコンポーネントが存在します: 1)トレーニングデータを決定するデータソースD; 2)
トレーニング報酬信号のソースとなる報酬関数。3) アルゴリズムA:トレーニング データと報酬信号を、データに対するペナルティまたは強化の大きさを決定する勾配係数に処理します。このような統一されたパラダイムに基づいて、いくつかの代表的な方法を分析します。

- 教師あり微調整（SFT）： SFTは、人間が選択したSFTに基づいて事前学習済みモデルを微調整します。
  データ。

方法	データソース	報酬関数の勾配係数	
スフト	、 ～（、）	-	1
RFT 個人情報保護方針	～（）、～（｜）（）、（｜） ～ +、 − ～	ルール ルール	式10 式14
オンラインRFT ＰＰＯ GRPO	～ （）、～（｜）（）、～ ～ （｜）（）、{ } ～ =1 ～ （｜）	ルール モデル モデル	式10 式18 式21

表10 |異なる方法のデータソースと勾配係数。データを示す

教師あり微調整データセットの分布。そして教師あり微調整データセットを表す。
それぞれ、オンライン トレーニング プロセス中にモデルとリアルタイム ポリシー モデルを実行します。

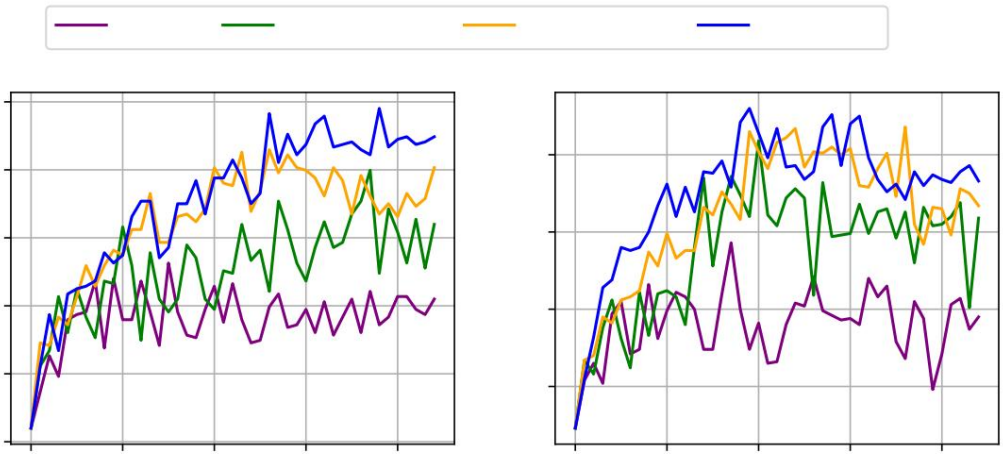

図5 |さらにトレーニングされたDeepSeekMath-Instruct 1.3Bモデルのパフォーマンス
2 つのベンチマークでさまざまな方法を使用しました。

- 拒絶サンプリング微調整（RFT）： RFTはSFTモデルをさらに微調整し、
  SFTの質問に基づいてSFTモデルからサンプリングされたフィルタリングされた出力。RFTは
  回答の正確さに基づいて出力します。
- 直接選好最適化（DPO）： DPOは、微調整によってSFTモデルをさらに改良します。
  ペアワイズ DPO 損失を使用して、SFT モデルからサンプリングされた拡張出力に適用します。
- オンライン拒否サンプリング微調整（オンラインRFT）： RFTとは異なり、オンラインRFT
  SFTモデルを使用して政策モデルを開始し、微調整することでそれを改良します。
  リアルタイム ポリシー モデルからサンプリングされた拡張出力。
- PPO/GRPO: PPO/GRPOはSFTモデルを使用してポリシーモデルを初期化し、
  リアルタイム ポリシー モデルからサンプリングされた出力と比較します。

これらの方法の構成要素を表10にまとめました。詳細については付録A.1を参照してください。
より詳細な導出プロセス。

データソースに関する考察データソースは、オンラインサンプリングとオフラインサンプリングの2つのカテゴリに分けられます。オンラインサンプリングは、トレーニングデータがリアルタイムトレーニングポリシーモデルの探索結果から取得されることを意味し、オフラインサンプリングは、

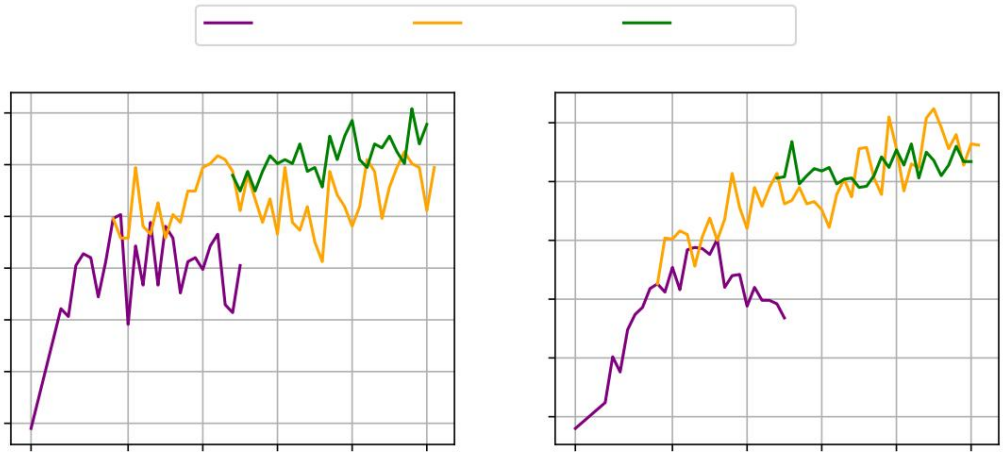

図 6 | 2 つのベンチマークにおける DeepSeekMath-Instruct 7B を使用した反復強化学習のパフォーマンス。

トレーニング データは、初期 SFT モデルのサンプリング結果から取得されます。RFT と DPO はオフライン スタイルに従いますが、オンライン RFT と GRPO はオンライン スタイルに従います。

図 5 に示すように、オンライン RFT は 2 つのベンチマークで RFT を大幅に上回っていることがわかります。具体的には、オンライン RFT はトレーニングの初期段階では RFT に匹敵しますが、後期段階では絶対的な優位性を獲得し、オンライン トレーニングの優位性を実証しています。
これは直感的です。初期段階では、アクターと SFT モデルはよく似ており、サンプリングされたデータにはわずかな違いしかありません。しかし、後の段階では、アクターからサンプリングされたデータにはより大きな違いが見られ、リアルタイムのデータ サンプリングにはより大きな利点があります。

勾配係数に関する観察アルゴリズムは、報酬信号を勾配係数に処理して、モデル パラメーターを更新します。実験では、報酬関数を「ルール」と「モデル」に分けます。ルールとは、回答の正確さに基づいて応答の品質を判断することを指し、モデルとは、各応答にスコアを付ける報酬モデルをトレーニングすることを意味します。報酬モデルのトレーニング データは、ルールの判断に基づいています。式 10と 21 は、GRPO とオンライン RFT の主な違いを強調しています。GRPO は、報酬モデルによって提供される報酬値に基づいて、勾配係数を一意に調整します。これにより、応答のさまざまな大きさに応じて、応答の差別的な強化とペナルティが可能になります。

対照的に、オンライン RFT にはこの機能がありません。誤った回答をペナルティにすることはなく、すべての回答を同じレベルの強度で正解で均一に強化します。

図 5 に示すように、GRPO はオンライン RFT を上回っており、正と負の勾配係数を変更する効率が強調されています。さらに、GRPO+PS はGRPO+OS と比較して優れたパフォーマンスを示しており、きめ細かいステップ認識勾配係数を使用する利点を示しています。さらに、反復 RL を調査し、実験では 2 ラウンドの反復を実行します。図 6 に示すように、反復 RL によって、特に最初の反復でパフォーマンスが大幅に向上することがわかります。

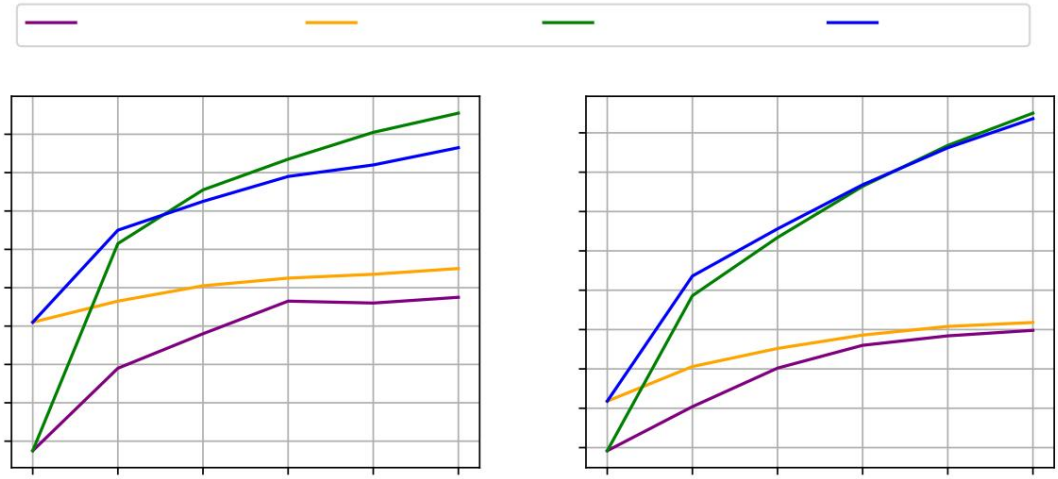

図 7 | GSM8K および MATH (温度 0.7) における SFT および RL DeepSeekMath 7B の Maj@K と Pass@K。RL はMaj@K を向上させますが、Pass@K は向上しないことがわかりました。

5.2.2. RL が機能する理由

本稿では、命令チューニングデータのサブセットに基づいて強化学習を実施し、命令チューニングモデルの大幅なパフォーマンス向上を実現します。
強化学習が機能する理由をさらに説明します。2つのベンチマークで、Instruct モデルと RL モデルのPass@K と Maj@K の精度を評価します。図 7 に示すように、RL は Maj@K のパフォーマンスを向上させますが、Pass@K のパフォーマンスは向上しません。これらの結果は、RL が出力分布をより堅牢にすることでモデルの全体的なパフォーマンスを向上させることを示しています。言い換えれば、改善は基本的な機能の向上ではなく、TopK からの正しい応答の向上に起因しているようです。同様に、(Wang et al., 2023a) は、SFT モデル内の推論タスクにおける不整合の問題を特定し、一連の好みの調整戦略を通じて SFT モデルの推論パフォーマンスを改善できることを示しました (Song et al., 2023; Wang et al., 2023a; Yuan et al., 2023b)。

5.2.3. より効果的な RL を実現するには?

RL は数学的推論タスクで非常にうまく機能することを実証しました。また、さまざまな代表的なトレーニング方法を理解するための統一されたパラダイムも提供します。このパラダイムでは、すべての方法が直接または簡略化された RL テクニックとして概念化されています。式 5 にまとめられているように、データ ソース、アルゴリズム、報酬関数という 3 つの主要なコンポーネントが存在します。
これら 3 つのコンポーネントに関する将来の方向性をいくつか示します。

データソースデータソースは、すべてのトレーニング方法の原材料です。RLの文脈では、データソースとは、ポリシーモデルからサンプリングされた出力を持つラベルなしの質問を指します。この論文では、指示チューニング段階からの質問とナイーブな核サンプリングのみを使用して出力をサンプリングします。これが、RLパイプラインがMaj@Kのパフォーマンスのみを向上させる潜在的な理由であると考えています。今後は、ツリー検索法に基づくもの（Yao et al.、2023)などの高度なサンプリング（デコード)戦略と組み合わせて、分布外の質問プロンプトでRLパイプラインを探索します。また、効率的な推論技術（Kwon et al.、2023; Leviathan et al.、2023; Xia et al.、2023、2024)は、

政策モデルの探索効率も非常に重要な役割を果たします。

アルゴリズムアルゴリズムは、データと報酬信号を勾配係数に処理して、モデルパラメータを更新します。 式5に基づくと、ある程度、すべての方法は、特定のトークンの条件付き確率を増加または減少させる報酬関数の信号を完全に信頼するようになりました。 ただし、特に非常に複雑なタスクでは、報酬信号が常に信頼できることを保証することは不可能です。 たとえば、十分に訓練された注釈者によって慎重に注釈が付けられたPRM800Kデータセット（Lightman et al.、2023)でさえ、約20%の誤った注釈7がまだ含まれています。 この目的のために、ノイズの多い報酬信号に対して堅牢な強化学習アルゴリズムを検討します。 このようなWEAK-TO-STRONG （Burns et al.、 2023)アライメント方法は、学習アルゴリズムに根本的な変化をもたらすと考えています。

報酬関数報酬関数はトレーニング信号のソースです。RL では、報酬関数は通常、ニューラル報酬モデルです。報酬モデルには、次の 3 つの重要な方向性があると考えています。1) 報酬モデルの一般化能力を高める方法。報酬モデルは、分布外の質問や高度なデコード出力を処理できるように効果的に一般化する必要があります。そうしないと、強化学習は LLM の分布を安定させるだけで、基本的な機能を向上させることはできません。2)報酬モデルの不確実性を反映する方法。不確実性は、弱い報酬モデルと弱から強への学習アルゴリズム間の橋渡しとして機能する可能性があります。3)推論プロセスにきめ細かいトレーニング信号を提供できる高品質のプロセス報酬モデルを効率的に構築する方法(Lightman et al.、2023; Wang et al.、2023b)。

6. 結論、限界、今後の課題

我々は、競争レベルの MATH ベンチマークですべてのオープンソース モデルを上回り、クローズド モデルのパフォーマンスに近づく DeepSeekMath を紹介します。DeepSeekMath はDeepSeek-Coder-v1.5 7B で初期化され、500B トークンの継続的なトレーニングを受けます。トレーニング データの重要なコンポーネントは、Common Crawl から取得した 120B の数学トークンです。我々の広範なアブレーション スタディでは、Web ページは高品質の数学データに大きな可能性を秘めている一方で、arXiv は期待したほど有益ではない可能性があることが示されています。我々は、メモリ消費を抑えながら数学的推論機能を大幅に向上させることができる、Proximal Policy Optimization (PPO) のバリアントである Group Relative Policy Optimization (GRPO)を紹介します。実験結果は、DeepSeekMath-Instruct 7B がベンチマークで高スコアに達した場合でも GRPO が有効であることを示しています。また、一連の方法を理解し、より効果的な強化学習のいくつかの潜在的な方向性をまとめるための統一されたパラダイムも提供します。

DeepSeekMathは定量的推論ベンチマークで印象的なスコアを達成していますが、幾何学と定理証明の能力はクローズドモデルに比べて比較的弱いです。たとえば、ドライランでは、モデルは三角形と楕円に関連する問題を処理できず、事前トレーニングと微調整でデータ選択の偏りを示している可能性があります。さらに、モデル規模によって制限されているため、DeepSeekMathは少数ショット機能の点でGPT-4よりも劣っています。GPT-4は少数ショット入力でパフォーマンスを向上させる可能性がありますが、DeepSeekMathはゼロショットと少数ショットの評価で同様のパフォーマンスを示します。今後は、エンジニアリングされたデータ選択パイプラインをさらに改善して、より高品質の事前トレーニング済みコーパスを構築します。さらに、 LLMのより効果的な強化学習の潜在的な方向性（セクション5.2.3)を検討します。

7https://github.com/openai/prm800k/issues/12#issuecomment-1728491852

参考文献

R. アニル、S. ボルゴー、Y. ウー、J. アララック、J. ユウ、R. ソリカット、J. シャルクウィク、AM ダイ、A. ハウト、 K. ミリカン、D. シルバー、S. ペトロフ、M. ジョンソン、I. アントノグルー、J. シュリットヴィーザー、A. グレース、J. チェン、 E. ピトラー、TP リリックラップ、A. ラザリドゥ、 O. フィラット、J. モロイ、M. アイサード、PR バーハム、T. ヘニガン、 B. リー、F. ヴィオラ、M. レイノルズ、Y. シュー、R. ドハティ、E. コリンズ、C. マイヤー、E. ラザフォード、E. モレイラ、 K. アユーブ、M. ゴエル、G. タッカー、E. ピケラス、M. クリクン、I. バー、N.サヴィノフ、I. ダニヘルカ、B. Roelofs、A. White、A. Andreassen、T. von Glehn、L. Yagati、M. Kazemi、L. Gonzalez、M. Khalman、J. Sygnowski、およびet al。Gemini：非常に優れたマルチモーダルモデルのファミリー。CoRR 、 abs / 2312.11805、2023。doi：10.48550 / ARXIV.2312.11805。URL https：//doi.org/10.48550/arXiv.2312.11805 。

J. オースティン、A. オデナ、M. ナイ、M. ボスマ、H. ミカレフスキー、D. ドーハン、E. ジャン、C. カイ、M. テリー、 Q. ル、他。大規模な言語モデルを使用したプログラム合成。 arXiv プレプリント arXiv:2108.07732、2021 。

Z. アゼルバエフ、H. シェルコップフ、K. パスター、MD サントス、S. マカリール、AQ ジャン、J. デン、S. ビダーマン、および S. ウェレック。 Llemma: 数学のオープン言語モデル。 arXiv プレプリント arXiv:2310.10631、2023。

J. Bai、S. Bai、Y. Chu、Z. Cui、K. Dang、X. Deng、Y. Fan、W. Ge、Y. Han、F. Huang、他クウェン技術レポート。arXivプレプリントarXiv:2309.16609、2023。

C. Burns、P. Izmailov、JH Kirchner、B. Baker、L. Gao、L. Aschenbrenner、Y. Chen、A. Ecoffet、 M. Joglekar、J. Leike、他「弱から強への一般化：弱い監督による強い能力の引き出し」arXiv プレプリント arXiv:2312.09390、2023 年。

ChatGLM3 チーム。Chatglm3 シリーズ: オープンなバイリンガル チャット LMS、2023 年。URL https://github.com/THUDM/ChatGLM3 。

M. チェン、J. トゥーレク、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. ブロックマン、A. レイ、R. プリ、G. クルーガー、M. ペトロフ、H. クラーフ、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. ババリアン、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. バーンズ、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、I. 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、I. 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.

W. Chen、X. Ma、X. Wang、WW Cohen。思考を促すプログラム：数値推論タスクのための計算と推論の分離。CoRR 、 abs / 2211.12588、2022年。doi： 10.48550/ARXIV.2211.12588。URL https://doi.org/10.48550/arXiv.2211.12588。

K. Cobbe、V. Kosaraju、M. Bavarian、M. Chen、H. Jun、L. Kaiser、M. Plappert、J. Tworek、J. Hilton、R. Nakano、他「数学の文章問題を解くための検証者のトレーニング」 arXiv プレプリント arXiv:2110.14168、2021 年。

T. Computer. Redpajama: 大規模言語モデルのトレーニングのためのオープンデータセット、2023年10月。URL https://github.com/togethercomputer/RedPajama-Data。

DeepSeek-AI。Deepseek LLM: 長期主義によるオープンソース言語モデルのスケーリング。CoRR 、 abs/2401.02954、2024。doi: 10.48550/ARXIV.2401.02954。URL https://doi.org/10.485 50/arXiv.2401.02954。

Z. Du、Y. Qian、X. Liu、M. Ding、J. Qiu、Z. Yang、J. Tang。Glm: 自己回帰空白補完による一般言語モデルの事前学習。計算言語学会第60回年次会議論文集（第1巻：長文論文）、 320～335ページ、 2022年。

L. Gao、A. Madaan、S. Zhou、U. Alon、P. Liu、Y. Yang、J. Callan、および G. Neubig。PAL: プログラム支援言語モデル。A. Krause、E. Brunskill、K. Cho、B. Engelhardt、S. Sabato、および J. Scarlett 編、 International Conference on Machine Learning、ICML 2023、2023 年 7 月 23 ～ 29 日、米国ハワイ州ホノルル、Proceedings of Machine Learning Researchの第 202 巻、 10764 ～ 10799 ページ。PMLR、2023 年。URL https://proceedings.mlr.press/v202/gao23f.html。

Z. Gou、Z. Shao、Y. Gong、Y. Shen、Y. Yang、M. Huang、N. Duan、W. Chen。Tora：数学的問題解決のためのツール統合推論エージェント。CoRR 、 abs / 2309.17452、2023。doi：10.48550 / ARXIV.2309.17452。URL https://doi.org/10.48550/arXiv.2309.17452.

D. Guo、Q. Zhu、D. Yang、Z. Xie、K. Dong、W. Zhang、G. Chen、X. Bi、Y. Wu、YK Li、F. Luo、 Y. Xiong、W. Liang。Deepseek-coder: 大規模言語モデルとプログラミングが出会うとき - コードインテリジェンスの台頭、2024年。

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。

D. Hendrycks、C. Burns、S. Kadavath、A. Arora、S. Basart、E. Tang、D. Song、J. Steinhardt。数学データセットによる数学問題解決の測定。arXivプレプリント arXiv:2103.03874、2021年。

ハイフライヤー。 Hai-llm:高効および量的大模型工具、 2023。URL https://www.high-flyer.c n/en/blog/hai-llm。

Inflection AI. Inflection-2、2023年。URL https://inflection.ai/inflection-2。

AQ Jiang、S. Welleck、JP Zhou、W. Li、J. Liu、M. Jamnik、T. Lacroix、Y. Wu、および G. Lample。ドラフト、スケッチ、および証明: 非公式証明による正式な定理証明者のガイド。arXivプレプリント arXiv:2210.12283、2022 年。

AQ Jiang、A. Sablayrolles、A. Mensch、C. Bamford、DS Chaplot、D. dl Casas、F. Bressand、 G. Lengyel、G. Lample、L. Saulnier、他。ミストラル7b。 arXiv プレプリント arXiv:2310.06825、2023。

A. ジューリン、E. グレイブ、P. ボジャノフスキー、M. ドゥーズ、H. ジェグー、T. ミコロフ。ファストテキスト。 zip:テキスト分類モデルを圧縮します。 arXiv プレプリント arXiv:1612.03651、2016。

W. Kwon、Z. Li、S. Zhuang、Y. Sheng、L. Zheng、CH Yu、JE Gonzalez、H. Zhang、I. Stoica。ページアテンションによる大規模言語モデルサービングの効率的なメモリ管理。2023年の ACM SIGOPS 29th Symposium on Operating Systems Principles の Proceedings に掲載。

Y. Leviathan、M. Kalman、および Y. Matias。投機的デコードによるトランスフォーマーからの高速推論。International Conference on Machine Learning、19274～19286 ページ。PMLR、 2023 年。

A. Lewkowycz、A. Andreassen、D. Dohan、E. Dyer、H. Michalewski、V. Ramasesh、A. Slone、C. Anil、I. Schlag、T. Gutman-Solo、他「言語モデルによる定量的推論問題の解決」神経情報処理システムの進歩、35:3843–3857、2022a。

A. Lewkowycz、A. Andreassen、D. Dohan、E. Dyer、H. Michalewski、VV Ramasesh、A. Slone、 C. Anil、I. Schlag、T. Gutman-Solo、Y. Wu、B. Neyshabur、G. Gur-Ari、および V. Misra。言語モデルによる定量的推論問題の解決。S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho、および A. Oh 編、 Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022、NeurIPS 2022、米国ルイジアナ州ニューオーリンズ、2022 年 11 月 28 日 - 12 月 9 日、 2022b。URL http://papers.nips。 cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstr act-Conference.html.

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。

I. Loshchilov と F. Hutter。分離重み減衰正規化。arXivプレプリント arXiv:1711.05101、2017年。

H. Luo、Q. Sun、C. Xu、P. Zhao、J. Lou、C. Tao、X. Geng、Q. Lin、S. Chen、および D. Zhang。Wizardmath: 強化された evol-instruct による大規模言語モデルの数学的推論の強化。arXiv プレプリント arXiv:2308.09583、2023。

S. ミシュラ、M. フィンレイソン、P. ルー、L. タン、S. ウェレック、C. バラル、T. ラジプロヒット、O. タフィヨルド、A. サブハルワル、P. クラーク、および A. カリヤン。 LILA: 数学的推論のための統一ベンチマーク。Y. Goldberg、Z. Kozareva、Y. Zhang 編著、 2022 年自然言語処理における経験的手法に関する会議の議事録、EMNLP 2022、アラブ首長国連邦アブダビ、2022 年 12 月 7 ～ 11 日、 5807 ～ 5832 ページ。計算言語学協会、 2022 年。doi: 10.18653/V1/2022.EMNLP-MAIN.392。URL https://doi.org/10.18653/v1/ 2022.emnlp-main.392。

X. Nguyen、W. Zhang、X. Li、MM Aljunied、Q. Tan、L. Cheng、G. Chen、Y. Deng、S. Yang、C. Liu、H. Zhang、およびL. Bing。Seallms - 東南アジア向け大規模言語モデル。CoRR 、 abs / 2312.00738、2023。doi :10.48550 / ARXIV.2312.00738。URL https://doi.org/10.485 50 / arXiv.2312.00738。

OpenAI。GPT4技術レポート。arXivプレプリントarXiv:2303.08774、2023。

L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、 K. Slama、A. Ray、他「人間から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」神経情報処理システムの進歩、35:27730–27744、2022年。

K. Paster、MD Santos、Z. Azerbayev、J. Ba。Openwebmath: 高品質の数学ウェブテキストのオープンデータセット。CoRR 、 abs/2310.06786、2023。doi: 10.48550/ARXIV.2310.06786。URL https://doi.org/10.48550/arXiv.2310.06786。

LC Paulson。自動定理証明器と対話型定理証明器を実際につなぐ sledgehammer の 3 年間の経験。RA Schmidt、S. Schulz、B. Konev 編著、自動推論の実際的側面に関する第 2 回ワークショップの議事録、PAAR-2010、エジンバラ、スコットランド、英国、2010 年 7 月 14 日、 EPiC Series in Computingの第 9 巻、 1 ～ 10 ページ。EasyChair、2010年。doi: 10.29007/TNFD。URL https://doi.org/10.29007/tnfd。

S. Polu と I. Sutskever。自動定理証明のための生成言語モデリング。CoRR 、 abs/2009.03393、2020 年。URL https://arxiv.org/abs/2009.03393。

R. Rafailov、A. Sharma、E. Mitchell、S. Ermon、CD Manning、C. Finn。直接的な選好最適化: 言語モデルは実は報酬モデルです。2023 年。

J. シュルマン。KL ダイバージェンスの近似、2020 年。URL http://joschu.net/blog/kl-app
rox.html を参照してください。

J. Schulman、P. Moritz、S. Levine、M. Jordan、および P. 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。

J. シュルマン、F. ウォルスキ、P. ダリワル、A. ラドフォード、O. クリモフ。近接政策最適化
アルゴリズム。 arXiv プレプリント arXiv:1707.06347、2017。

F. Shi、M. Suzgun、M. Freitag、X. Wang、S. Srivats、S. Vosoughi、HW Chung、Y. Tay、S. Ruder、 D. Zhou、D. Das、および J. Wei。言語モデルは、多言語の思考連鎖推論システムです。
第11回国際学習表現会議、ICLR 2023、ルワンダ、キガリ、2023年5月1日～5日。OpenReview.net 、2023年。URL https://openreview.net/pdf?id= fR3wGCk-IXp。

F. Song、B. Yu、M. Li、H. Yu、F. Huang、Y. Li、H. Wang。選好順位最適化
人間のアライメント。arXivプレプリントarXiv:2306.17492, 2023。

M. Suzgun、N. Scales、N. Schärli、S. Gehrmann、Y. Tay、HW Chung、A. Chowdhery、QV Le、 EH Chi、D. Zhou、他「挑戦的なビッグベンチタスクと、思考連鎖で解決できるかどうか」arXiv プレプリント arXiv:2210.09261、2022 年。

T. Tao。変化を受け入れ、期待をリセットする、2023年。URL https://unlocked.micro soft.com/ai-anthology/terence-tao/。

H. トゥヴロン、L. マーティン、K. ストーン、P. アルバート、A. アルマハイリ、Y. ババエイ、N. バシリコフ、S. バトラ、 P. バルガヴァ、S. ボサレ、D. ビケル、L. ブレッチャー、C. カントンフェレール、M. チェン、G. ククルル、D. エシオブ、 J. フェルナンデス、J. フー、W. フー、B. フラー、 C. ガオ、V. ゴスワミ、N. ゴヤル、A. ハーツホルン、 S. ホセイニ、R. ホウ、H. イナン、M. カルダス、V. ケルケス、M. カブサ、I. クルーマン、A. コレネフ、PS コウラ、 M. ラショー、T. ラブリル、J. リー、D. リスコビッチ、Y. ルー、Y. マオ、X. マーティネット、T.ミハイロフ、P.ミシュラ、 I. Molybog、Y. Nie、A. Poulton、J. Reizenstein、R. Rungta、K. Saladi、A. Schelten、R. Silva、EM
スミス、R. スブラマニアン、XE タン、B. タン、R. テイラー、A. ウィリアムズ、JX クアン、P. シュー、Z. ヤン、 I. ザロフ、Y. チャン、A. ファン、M. カンバドゥル、S. ナラン、A. ロドリゲス、R. ストジニッチ、S. エドゥノフ、 T. シャロム。 Llama 2: オープンな基盤と微調整されたチャット モデル。 CoRR、 abs/2307.09288、2023。doi: 10.48550/arXiv.2307.09288。 URL https://doi.org/10.48550/arXiv.2307。 09288.

TH Trinh、Y. Wu、QV Le、H. He、T. Luong。人間なしでオリンピック幾何学を解く
デモンストレーション。ネイチャー、625（7995）：476–482、2024年。

P. Wang、L. Li、L. Chen、F. Song、B. Lin、Y. Cao、T. Liu、Z. Sui。大規模言語モデルの作成
アライメントによる推論精度の向上。arXiv プレプリント arXiv:2309.02144, 2023a。

P. Wang、L. Li、Z. Shao、R. Xu、D. Dai、Y. Li、D. Chen、Y. Wu、および Z. Sui。Math-shepherd：人間の注釈なしで、ステップバイステップでリソースを検証および強化します。CoRR、abs / 2312.08935、2023b。

Z. Wang、R. Xia、P. Liu。数学のための生成 AI: パート I - mathpile: 数学のための 10 億トークン規模の事前トレーニング コーパス。CoRR 、 abs/2312.17120、2023c。doi: 10.48550/ARXIV.2312.17120。
URL https://doi.org/10.48550/arXiv.2312.17120.

J. Wei、X. Wang、D. Schuurmans、M. Bosma、B. Ichter、F. Xia、EH Chi、QV Le、および D. Zhou。
思考連鎖の促進が大規模言語モデルにおける推論を引き出す。NeurIPS 、 2022年。
URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf 4f15af0f7b31abca4-Abstract-Conference.html.

T. Wei、J. Luan、W. Liu、S. Dong、B. Wang。Cmath: 言語モデルは中国語に合格できるか
小学校の算数のテスト？、2023年。

M. ウェンゼル、LC ポールソン、T. ニプコウ。イザベルフレームワーク。OA モハメッド、CA
Muñoz、S. Tahar 編、「Theorem Proving in Higher Order Logics」、第 21 回国際会議、TPHOLs 2008、カナダ、モントリオール、2008 年 8 月 18 ～ 21 日。議事録、Lecture Notes in Computer Scienceの第 5170 巻、33 ～ 38 ページ。Springer、2008 年。doi: 10.1007/978-3-540-7 1067-7_7。URL https://doi.org/10.1007/978-3-540-71067-7_7。

H. Xia、T. Ge、P. Wang、S.-Q. Chen、F. Wei、および Z. Sui。投機的デコード：投機的実行を利用して seq2seq 生成を高速化する。H. Bouamor、J. Pino、および K. Bali編、Findings of the Association for Computational Linguistics: EMNLP 2023、3909 ～ 3925 ページ、シンガポール、2023 年 12 月。Association for Computational Linguistics。doi: 10.18653/v1/20 23.findings-emnlp.257。URL https://aclanthology.org/2023.findings-emnlp.257。

H. Xia、Z. Yang、Q. Dong、P. Wang、Y. Li、T. Ge、T. Liu、W. Li、および Z. Sui。大規模言語モデル推論における効率性の解放：投機的デコードの包括的な調査。arXivプレプリント arXiv:2401.07851、2024。

S. Yao、D. Yu、J. Zhao、I. Shafran、TL Griffiths、Y. Cao、K. Narasimhan。思考のツリー：大規模言語モデルによる意図的な問題解決。arXivプレプリント arXiv:2305.10601、2023 。

L. Yu、W. Jiang、H. Shi、J. Yu、Z. Liu、Y. Zhang、JT Kwok、Z. Li、A. Weller、W. Liu。
Metamath: 大規模言語モデルのための独自の数学的質問をブートストラップします。CoRR 、 abs/2309.12284、2023。doi: 10.48550/ARXIV.2309.12284。URL https://doi.org/10.485 50/arXiv.2309.12284。

Z. Yuan、H. Yuan、C. Li、G. Dong、C. Tan、C. Zhou。大規模言語モデルによる数学的推論の学習におけるスケーリング関係。arXivプレプリント arXiv:2308.01825, 2023a。

Z. Yuan、H. Yuan、C. Tan、W. Wang、S. Huang、F. Huang。Rrhf: 応答をランク付けして、涙なしで言語モデルを人間のフィードバックに合わせます。arXivプレプリント arXiv:2304.05302、2023b 。

X. Yue、X. Qu、G. Zhang、Y. Fu、W. Huang、H. Sun、Y. Su、W. Chen。マンモス：ハイブリッド命令チューニングによる数学ジェネラリストモデルの構築。CoRR 、 abs / 2309.05653、2023。doi： 10.48550/ARXIV.2309.05653。URL https://doi.org/10.48550/arXiv.2309.05653。

K. Zheng、JM Han、S. Polu。Minif2f: 公式オリンピックレベルのクロスシステムベンチマーク
数学。arXivプレプリントarXiv:2109.00110、2021年。

W. Zhong、R. Cui、Y. Guo、Y. Liang、S. Lu、Y. Wang、A. Saied、W. Chen、および N. Duan。 AGIEval:基礎モデルを評価するための人間中心のベンチマーク。 CoRR、 abs/2304.06364、2023。土井: 10.48550/arXiv.2304.06364。 URL https://doi.org/10.48550/arXiv.2304.06364。

A. 付録

A.1. 強化学習の分析

データソースと勾配係数（アルゴリズムと
SFT、RFT、オンラインRFT、DPO、PPOなど、さまざまな方法で報酬関数を実装し、
GRPO。

A.1.1. 教師あり微調整

教師あり微調整の目的は、次の目標を最大化することです。

$$J(\ ) = E[\ ,\quad \sim (、\ ) \ ]\quad \frac{1}{\ }\sum_{=1}^{|\ |} ログ(|, <)。 \tag{6}$$

J ( ) の勾配は次のようになります。

$$\nabla J = E[\ ,\quad \sim (、\ ) \ ]\quad \frac{1}{|\ |}\sum_{=1}^{|\ |} \nabla \log(|, <) です。 \tag{7}$$

データソース: SFTに使用されたデータセット。報酬関数: これは人間の
選択。勾配係数: 常に 1 に設定されます。

A.1.2. 拒否サンプリングの微調整

拒否サンプリング微調整は、まず教師あり微調整された出力から複数の出力をサンプリングする。
各質問に対して LLM を実行し、正解を含むサンプル出力で LLM をトレーニングします。
正式には、RFT の目的は次の目標を最大化することです。

$$J(\ ) = E[\ \ (),\ \ (\qquad |)]\quad \frac{1}{|\ |}\sum_{=1}^{|\ |} I(\ ) ログ (\ |\ , <)\qquad 。 \tag{8}$$

J ( ) の勾配は次のようになります。

$$\nabla J(\ ) = E[\ \ (),\ \ (\qquad |)]\quad \frac{1}{|\ |}\sum_{=1}^{|\ |} I(\ ) \nabla \log (\qquad |、 <)。 \tag{9}$$

データソース: SFT モデルからサンプリングされた出力を含む SFT データセット内の質問。報酬関数:
ルール（答えが正しいかどうか）。勾配係数：

$$(、、) = 私 (\ ) = \begin{cases} 1\ o の答えが正解です \\ 0\ o の答えは不正解です \end{cases} \tag{10}$$

A.1.3. オンライン拒否サンプリングの微調整

RFTとオンラインRFTの唯一の違いは、オンラインRFTの出力がサンプリングされることです。
リアルタイム政策モデルではなくSFTモデルから得られる。したがって、勾配
オンラインRFTは次のとおりです。

$$\nabla J(\ ) = E[\sim (\ ), \sim (\ |\ )]\quad \frac{1}{\ }\sum_{=1}^{|\ |} I(\ ) \nabla \log (\ |\ , <)\qquad 。 \tag{11}$$

A.1.4. 直接選好最適化（DPO）

DPO の目的は次のとおりです。

$$J(\ ) = E[\ \ (\ ),\ \ {}^{+}、{}^{-}\ \ (\ |\ )]\ \log \frac{1}{|+|}\sum_{=1}^{|\ ^{+}|} \log \frac{{}^{+}|、{}^{+}_{<})}{(参照({}^{+}|、{}^{+}_{<})} - \frac{1}{|-|}\sum_{=1}^{|-|} \log \frac{(\ _{<}|、{}^{-}_{<})}{参照(\ _{<}|、{}^{-}_{<})} \tag{12}$$

J ( ) の勾配は次のようになります。

$$\nabla J(\ ) = E[\ \ (\ ),\ \ {}^{+}、{}^{-}\sim\ \ (\ |\ )\ ]\ \frac{1}{|+|}\sum_{=1}^{|\ ^{+}|} (\ ,\ ,\ )\nabla \log(\ \ {}^{+}|、{}^{+}_{<})$$
$$- \frac{1}{|-|}\sum_{=1}^{|-|} (\ ,\ ,\ )\nabla \log(-|\ \ 、{}^{-}_{<}) \tag{13}$$

データソース: SFT モデルからサンプリングされた出力を含む SFT データセット内の質問。報酬関数: 一般的な領域における人間の好み（数学的な課題では「ルール」となる）。勾配係数：

$$(、、)= \log \frac{(-|,\ {}^{-}_{<})}{ref(-|,\ {}^{-}_{<})} - \log \frac{{}^{+}|、{}^{+}_{<})}{(参照({}^{+}|、{}^{+}_{<})} \tag{14}$$

A.1.5. 近接ポリシー最適化 (PPO)

PPO の目的は次のとおりです。

$$J(\ ) = E[\ \ (\ ),\ \ (\ |\ )]\ |\ \frac{1}{|}\sum_{=1}^{|\ |} 分\ \frac{(|、<)}{(|、<)}\ _{、クリップ}\ \frac{(|、<)}{(|、<)}、1-\ 、1+\ 。 \tag{15}$$

分析を簡略化するために、モデルは各更新の後に1回だけ更新されると仮定します。
探査段階において、 = 。この場合、 minとclipを削除できます
手術：

$$J(\ ) = E[\ \ (\ ),\ \ (\ |\ )]\ |\ \frac{1}{|}\sum_{=1}^{|\ |} \frac{|、<)}{(\ (\ |、<)}\ 。 \tag{16}$$

J ( ) の勾配は次のようになります。

$$\nabla J(\ ) = E[\ \ (\ )、\ \ (\ |\ )]\ |\ \frac{1}{|}\sum_{=1}^{|\ |} \nabla \log(\ \ |、<) \tag{17}$$

データソース: ポリシーモデルからサンプリングされた出力を含む SFT データセット内の質問。報酬関数: 報酬モデル。勾配係数:

$$(、、、\ \ ) = \ \ 、 \tag{18}$$

ここで、は一般化優位性推定を適用して計算される優位性である。
(GAE) (Schulman et al., 2015)は、報酬{ ≥ }と学習された価値関数に基づいています。

A.1.6. グループ相対ポリシー最適化 (GRPO)

GRPOの目的は（仮定） = 簡易分析用):

$$J(\ ) = E[\ \ (\ )、\{\ \}_{=1}\sim\ \ (\ |\ )\ ]$$
$$- \sum_{=1} \frac{1}{|\ |}\sum_{=1}^{|\ |} \frac{、|、、<)}{(\ (\ 、|)、、<}\ 、- \left(\frac{、|、、<)}{(\ (、|、、<)} - \log \frac{、|、、<)}{(\ (、|、、<)} - 1\right)。 \tag{19}$$

Machine Translated by Google

J ( ) の勾配は次のようになります。

$$\nabla J ( ) = E[\quad ( ), \{ \}_{=1} \sim \quad (|) \;]$$

$$\frac{1}{} \sum_{=1} \frac{1}{| \; |} \sum_{=1}^{| \; |} \text{、}^{+} \frac{\text{、}| \; \text{、}<)}{((\text{、}| \; \text{、}<)} - 1 \nabla \log (\quad \text{、}|\text{、} \; \text{、}<)\text{。} \tag{20}$$

データソース: ポリシーモデルからサンプリングされた出力を含む SFT データセット内の質問。報酬関数: 報酬モデル。勾配係数:

$$(\text{、}\text{、}\text{、} \quad) = \quad \text{、}^{+} \frac{\text{、}| \; \text{、}<)}{((\text{、}| \; \text{、}<)} - 1 \;\; \text{、} \tag{21}$$

どこ 、グループ報酬スコアに基づいて計算されます。