アンサンブル蒸留と学習ベース集計を用いた数学的推論プロセスの検証と性能分析

榎本倫太郎 [1,3]　栗田修平 [2,3]　河原大輔 [1,3]
[1] 早稲田大学 [2] 国立情報学研究所 [3] 国立情報学研究所 大規模言語モデル研究開発センター
{re9484@akane.,dkw@}waseda.jp,　skurita@nii.ac.jp

概要

大規模言語モデルの数学的推論能力向上には，推論過程を適切に評価するプロセス報酬モデル（PRM）が有効であるが，構築時の報酬信号生成に膨大なモンテカルロ・ロールアウト（MCR）を要する点が課題である．本論文では，MCR を必要としない解答強制フレーズを用いた他系統モデルからのアンサンブル蒸留と，統計的特徴量に基づく学習ベース集計を組み合わせた高効率な PRM 構築フレームワークを提案する．検証実験の結果，提案手法は難関タスク AIME 24 において，従来の結果報酬モデル（ORM）や強力なベースラインである多数決を上回る最高精度（20.67%）を達成した．分析により，提案モデルは論理の収束や自己矛盾を検知し，強化学習における密な報酬信号として極めて有望であることを示唆している．

1　はじめに

大規模言語モデル（LLM）は，Chain-of-Thought（CoT）プロンプティング [1] により複雑な推論を可能としたが，推論の長期化に伴うエラー伝播の問題を依然として抱えている [2, 3]．これに対し，複数の推論パスから有望なものを選択する Best-of-N サンプリングが有効であり，特にステップ単位で正当性を評価するプロセス報酬モデル（Process Reward Model; PRM）は，結果のみを評価する結果報酬モデル（Outcome Reward Model; ORM）[4] よりも高い評価精度が期待される [5].

しかし，PRM の構築には，膨大な計算コストが障壁となる．既存手法 [6, 7, 8] の多くは，各ステップの報酬信号として最終的な正解到達率を得るために，追加のサンプリングを伴うモンテカルロ・ロールアウト（MCR）を必要とする．本論文では，報酬信号生成時の追加ロールアウトを必要としない高効率な PRM 構築フレームワークを提案する．具体的には，推論パスを生成したモデル（7B 級）自身，あるいはそれと同等の規模を持つ他系統のモデルを「指導モデル」として位置づける．これらに対し，解答を強制するフレーズを挿入する「Prompted Forcing」を導入し，その条件下での正答生成確率を報酬信号として利用する．

さらに，PRM の運用において，推論ステップごとに得られる複数の報酬スコアをいかにしてパス全体の最終的な評価値へと集計するかという課題がある．この課題に対して従来の最小値や平均値といった単純な集計には限界がある．そこで，PRM の出力に含まれるノイズやモデル特性を考慮し，全ステップの統計的特徴量に基づく「学習ベース集計（Learned Aggregation）」を導入することで，推論パスの正誤識別能力の最大化を図る．

本研究の主な貢献は 3 点である．**第一に，複数の指導モデルから得られる報酬信号の統合学習（アンサンブル蒸留）を安定化する Prompted Forcing の実証である．**報酬蒸留を行う際，特定の強制フレーズ挿入による文脈制御が，多様な指導モデル群からの信号の識別性を高めることを示した．**第二に，学習ベース集計による PRM の潜在能力の抽出である．**学習ベースの集計器の導入により，ORM のような均一な報酬列では困難な，推論の進行に伴って正答への確信が高まっていく「論理の収束」が PRM において検知可能であることを示した．これにより，難関タスク AIME 24 において多数決や ORM を凌駕するピーク性能を達成し，PRM が持つ高密度な報酬信号の有用性を実証した．**第三に，指導モデルの系統性が与える影響の解明である．**学習データ生成モデルとは異なる系統（DeepSeek, Llama）の指導モデルを用いることで，生成時のバイアスを排除した客観的な報酬信号が得られることを確認した．また，アンサンブルによる一般化性能と特定モデルによる

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

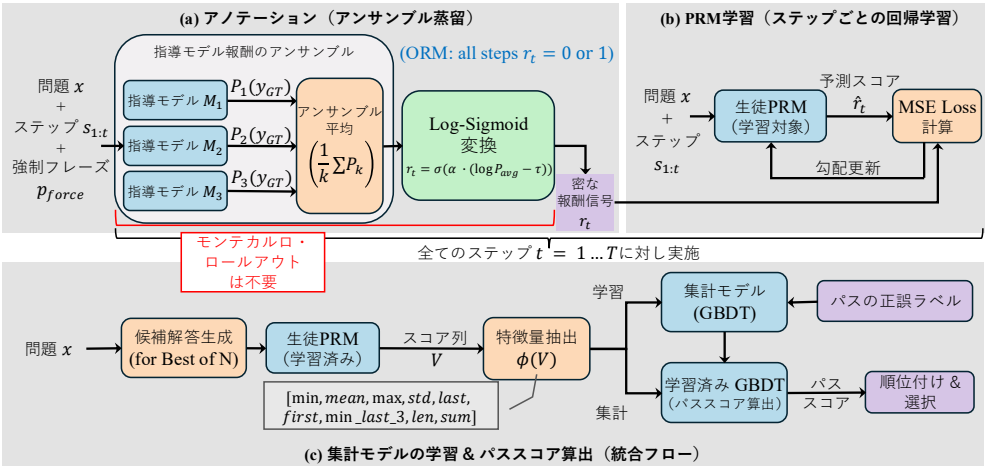


図 1: 提案する PRM フレームワークの概要.

ピーク性能のトレードオフを明らかにした.

2 関連研究

プロセス報酬データの自動構築と効率性　推論パスを密に評価する PRM は，ORM では困難な早期エラー検知を可能にする [5, 6]. PRM 作成における人手アノテーション [9] のコスト回避のため，Math-Shepherd [6] や OmegaPRM [7] 等の自動構築手法が提案されているが，報酬信号生成に多大な MCR を要する点がボトルネックとなる.

モデル内在的確率と自己評価の限界　一方，強化学習では計算コスト削減のため，モデル自身の生成確率を報酬とする LaSeR [10]，RLPR [11]，PACR [12] 等が提案されている．これらは解答生成後の特殊トークンや，特定の接頭辞を挿入して正答トークンの予測確率を算出する手法を用いる．しかし，学習モデル自身の主観的な確率に依存するため，モデル固有の過信やバイアスを排除しきれない課題がある．提案する「Prompted Forcing」と「指導モデル群の蒸留」は，複数の強力な指導モデルによるアンサンブル結果を軽量な 1.5B モデルに統合する．これにより，自己報酬バイアスの軽減と推論リソース削減を同時に実現する.

検証スコアの集計手法　PRM による推論パス評価では，可変長のスコア列を集計する必要がある．従来は積や最小値等の単純な経験的な手法が標準的であったが [9, 6]，比較的小規模な言語モデル (Small Language Model; SLM) の出力はノイズを含みやすく，評価を誤るリスクがある．本研究では，スコア列の分布特徴を考慮する「学習ベース集計」を導入しこれに対処する.

強化学習への応用　PRM は GRPO [13] 等の強化学習において，スパースな結果報酬を補完する有効な手段である．密なフィードバックは学習効率を向上させることが示唆されており [12]，本研究におけるステップ検証性能と報酬軌跡の分析によって，将来的な高密度報酬信号としての有効性が示される.

3 提案手法

本研究では，高効率な PRM 作成のため，図 1 に示す (a) 指導モデル群による報酬アノテーション，(b) 生徒モデル (PRM) の学習，(c) GBDT によるスコア集計と順位付けの 3 段階からなるフレームワークを提案する.

3.1 確率ベースのプロセス報酬モデル

本手法では指導モデルの確信度を蒸留し，高品質なプロセス報酬を学習する (図 1(a), (b)).

Prompted Forcing による確信度抽出　学習データ生成モデルで数学問題 x に対する推論パス s を生成し，$s = (s_1, \ldots, s_T)$ の各ステップ s_t までを評価するため，指導モデル群 $\mathcal{M} = \{M_1, \ldots, M_K\}$ を利用する．評価の安定化のため，文脈 $(x, s_{1:t})$ の直後に接続フレーズ p_{force}（“\nThe final answer is \boxed{”）を挿入し，正解トークン y_{GT} の生成確率 $P_k(s_{1:t}) = P_{M_k}(y_{GT} \mid x, s_{1:t}, p_{force})$ を算出する．これで，$s_{1:t}$ が正解へ接続可能かを測定する．また，モデル固有の過信を抑制するため，指導モデル群の平均確信度 $P_{avg}(s_{1:t}) = \frac{1}{K}\sum_{k=1}^{K} P_k(s_{1:t})$ を教師信号として用いる．比較として，単一モデルによる蒸留も行う.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Log-Sigmoid 変換による報酬設計　極めて小さな値をとりうる確率を回帰学習に適した範囲へ写像するため，Log-Sigmoid 変換を導入する．ステップ報酬 r_t を $r_t = \sigma(\alpha \cdot (\log P_{avg}(s_{1:t}) - \tau))$ と定義する．ここで τ は中心化パラメータ，α はスケーリング係数である．生徒モデル（PRM）は，この r_t を平均二乗誤差（MSE）により予測するように学習する．比較用の ORM 報酬信号は，パス全体の正誤（1 または 0）を全ステップに一律に付与する．

3.2　学習ベースのスコア集計

可変長のスコア列 $V = (v_1, \ldots, v_T)$ を単一指標に統合する際，学習ベース集計を導入する (図 1(c))．

特徴量抽出と集計モデル　スコア列 V から以下の 9 次元の特徴量 $\phi(V)$ を抽出する：

- **基本統計量**: Min (最小値), Mean (平均), Max (最大), Std (標準偏差), Sum (合計)
- **位置・局所指標**: First (ステップ 1), Last (最後のステップ), Min-Last-3 (終盤 3 ステップの最小値)
- **パス属性**: Length (ステップ長)

これらを入力とし，パスの正否を二値分類する勾配ブースティング木（GBDT）を学習する．推論時には GBDT が出力する正解確率を最終的なパス評価値として採用し，Best-of-N の順位付けを行う．

4　実験

4.1　実験設定

PRM の構築　学習データとして NuminaMath [14] から 15,000 問を抽出し，Qwen2.5-Math-7B-Instruct [15] を用いて各 8 パス（計約 12 万パス）を生成した．各パスは二重改行でステップ分割し，Qwen2.5-Math-7B-Instruct，DeepSeek-Math-7B-Instruct [13]，Llama-3.1-8B-Instruct [16] の 3 つの指導モデルにより，Prompted Forcing に基づく報酬信号 r_t を算出した．Log-Sigmoid 変換のパラメータは $\tau = -5.0, \alpha = 1.0$ とした．生徒モデルには Qwen2.5-Math-1.5B-Instruct [15] を採用し，1 エポックの回帰学習を行った．合計学習データ量は約 100 万サンプル（問題数 * 8 パス * 平均ステップ数）である．

評価設定　評価用データセットとして，Math500 [17, 9]，AIME 24 [18]，AIME 25 [19] を使用する．Qwen2.5-Math-1.5B-Instruct を用いて Best-of-N 評価（$N = 16, 64$）を，5 つの独立したシードで実施し，その平均正答率を最終結果とした．

表 1: 各ベンチマークにおける順位付け性能（正答率, %）の比較. $N = 16, 64$ のうち高い方の値を記載.

BM	手法	単純集計 †	学習ベース集計
MATH500	多数決	62.48	—
	Avg@N	58.49	—
	ORM	61.20 (MEAN)	59.96
	PRM (Qwen)	60.36 (LAST)	60.24
	PRM (Llama)	60.84 (LAST)	**61.68**
	PRM (DeepSeek)	60.68 (MEAN)	60.88
	PRM (Ensemble)	60.60 (MEAN)	60.56
AIME24	多数決	18.00	—
	Avg@N	10.69	—
	ORM	20.00 (MIN)	12.67
	PRM (Qwen)	16.00 (MEAN)	13.33
	PRM (Llama)	16.67 (MEAN)	14.67
	PRM (DeepSeek)	16.00 (MEAN)	**20.67**
	PRM (Ensemble)	14.67 (LAST)	17.33
AIME25	多数決	19.33	—
	Avg@N	10.58	—
	ORM	13.33 (SUM)	10.67
	PRM (Qwen)	14.00 (LAST)	12.67
	PRM (Llama)	12.00 (LAST)	10.67
	PRM (DeepSeek)	12.00 (SUM)	9.33
	PRM (Ensemble)	14.67 (LAST)	**16.00**

† Min, Mean, Last, Sum の中での最高精度を記載.

比較対象として，Avg@N (複数回生成のうちランダムに選んだ時の平均性能), 多数決，ORM，および PRM の固定集計（Min, Mean, Last, Sum）を設定した．学習ベース集計では，NuminaMath で学習した GBDT を用いた．

4.2　実験結果

表 1 に, 各ベンチマークにおける Best-of-N による正答率の比較を示す. なお, 今後表と本文に示す PRM（モデル名）はモデル名を指導モデルとした PRM を使った結果である．

主要な性能傾向　AIME 24 において，PRM (DeepSeek) に学習ベース集計を適用したモデルが 20.67% を記録し，ORM (MIN) の 20.00% および多数決 (18.00%) を上回る最高精度を達成した．一方，MATH500 および AIME 25 では 多数決 が最も高い正答率を維持しており，タスクや N によって最適な検証戦略が異なることが確認された．

学習ベース集計の有効性と限界　提案する学習ベース集計の有効性は，報酬モデルの性質とタスクに強く依存することが確認された．PRM においては，学習ベース集計が特定の難関設定で最高性能を

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

引き出す鍵となっている．AIME 24 における PRM (DeepSeek) は，固定指標（16.00%）から 20.67% へと飛躍的な向上を見せ，AIME 25 の PRM (Ensemble) も 14.67% から 16.00% へと改善した．これは，ステップごとの報酬変化を伴う PRM の特性により，スコア分布 $\phi(V)$ に正誤識別のための有益な信号が含まれていることを示唆する．

一方で，ORM は学習ベース集計の適用によって，AIME 24 で 20.00% から 12.67% へと大幅に悪化した．これは，ORM が全ステップに同一報酬信号を付与する性質上，スコア列が均一化されており，統計的特徴に基づく識別が困難であるためと考えられる．学習ベース集計が PRM の特定設定においてのみ有効であった事実は，本手法が汎用的な解法ではないものの，従来の経験的な手法では到達不可能なピーク性能を特定の条件下で突破し得る強力な手法であることを示している．

指導モデルとアンサンブル効果　指導モデルの分析では，学習データの生成や報酬算出に用いた Qwen 系列とは異なる系列（DeepSeek や Llama）の確信度を蒸留した PRM が，高い検証性能を示す傾向が確認された．具体的には，AIME 24 における PRM (DeepSeek) の最高値（20.67%）や，MATH 500 における PRM (Llama) の PRM 内最高性能（61.68%）がこれに該当する．

また注目すべきはアンサンブルの有効性である．AIME 25 において，単一の指導モデルを用いた PRM はいずれも 14.00% 以下に留まる中，PRM (Ensemble) は学習ベース集計との組み合わせで 16.00% と全報酬モデル中で最高の精度を達成した．PRM (DeepSeek) が AIME 24 で最高精度を記録しながら AIME 25 では 12.00% と失速し，学習ベース集計 (9.33%) でも改善が見られないなど，単独モデルにはタスク依存の性能のばらつきが見られる．対照的に，アンサンブルは多様な指導モデルによる合議を通じてこれらの欠点を補完し，難関タスクにおいて一貫して堅牢な報酬信号を提供できている．

5　追加分析と考察

提案手法の妥当性を多角的に検証する．まず，定性分析およびスコア軌跡の分析（詳細は付録 A, B）により，提案 PRM が論理の自己矛盾や収束プロセスを鋭敏に検知可能であることを確認した．

解答強制の効果　アブレーション研究（表は付録 C 参照）の結果，解答強制の導入により全タスクで最高精度が向上し，特に AIME 25 では +4.0% の大幅な改善を確認した．これは，現在の推論ステップが最終的な正答を得るための有効な足がかりとなっているかを評価する設計が，識別精度の向上に直結していることを示している．

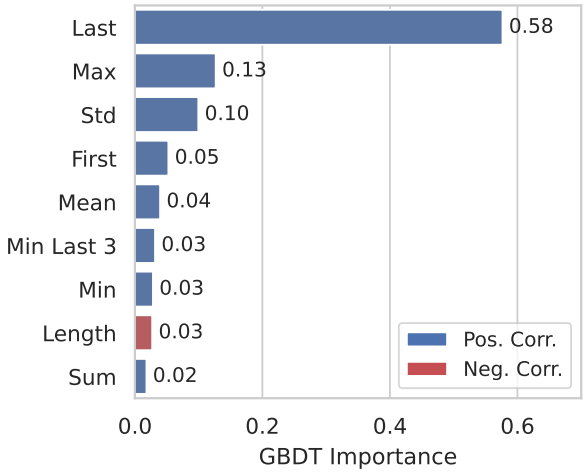


図 2: PRM 特徴量の重要度と正答との相関．棒の長さは寄与度，色は相関係数の正負（青：正，赤：負）．

集計モデルの特性分析　GBDT による特徴量分析の結果を図 2 に示す．分析により，以下の知見を得た．(1) 最終ステップ報酬 **Last** が約 0.58 と最大の重要度を示し，正答と正の相関（+0.36）を持つ．これは PRM が推論全体の帰結を最も支配的な判断基準としていることを裏付けている．(2) PRM 特有の動態指標である報酬の分散 **Std** は，0.10 の高い重要度と正の相関（+0.23）を示した．正解パスにおける報酬の変動は，難所での一時的な低下とその後の回復という「思考のプロセス」(付録 B, 図 3(b) 参照) を反映しており，集計器はこの変動を正の信号として捉えている．(3) 補助的なメタ特徴量である **Length** は，負の相関（-0.10）を示した．これは数学的推論における不必要なステップ長が論理の停滞や迷走を意味していることを示唆する．

6　おわりに

本論文では，MCR に依存しない高効率な PRM 構築法を提案し，AIME 24 において多数決を凌駕する最高精度（20.67%）を達成した．分析の結果，本手法は論理の自己矛盾を特定ステップで敏感に検知可能であり，ORM では困難なステップ単位の密なフィードバックを提供し得ることを実証した．今後は，本信号を強化学習の即時報酬として統合し，自律的な推論能力のさらなる向上を目指す．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は, 文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです. また, 産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」の支援を受けて利用しました.

参考文献

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, Vol. 35, pp. 24824–24837, 2022.

[2] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. **arXiv preprint arXiv:2310.01798**, 2023.

[3] Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 13894–13908, 2024.

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[5] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. **arXiv preprint arXiv:2211.14275**, 2022.

[6] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9426–9439, 2024.

[7] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. **arXiv preprint arXiv:2406.06592**, 2024.

[8] Lingxiao Du, Fanqing Meng, Zongkai Liu, Zhixiang Zhou, Ping Luo, Qiaosheng Zhang, and Wenqi Shao. Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision. **arXiv preprint arXiv:2505.13427**, 2025.

[9] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In **The Twelfth International Conference on Learning Representations**, 2023.

[10] Wenkai Yang, Weijie Liu, Ruobing Xie, Yiju Guo, Lulu Wu, Saiyong Yang, and Yankai Lin. Laser: Reinforcement learning with last-token self-rewarding. **arXiv preprint arXiv:2510.14943**, 2025.

[11] Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, et al. Rlpr: Extrapolating rlvr to general domains without verifiers. **arXiv preprint arXiv:2506.18254**, 2025.

[12] Eunseop Yoon, Hee Suk Yoon, Jaehyun Jang, SooHwan Eom, Qi Dai, Chong Luo, Mark A Hasegawa-Johnson, and Chang D Yoo. Pacr: Progressively ascending confidence reward for llm reasoning. **arXiv preprint arXiv:2510.22255**, 2025.

[13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint arXiv:2402.03300**, 2024.

[14] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

[15] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. **arXiv preprint arXiv:2409.12122**, 2024.

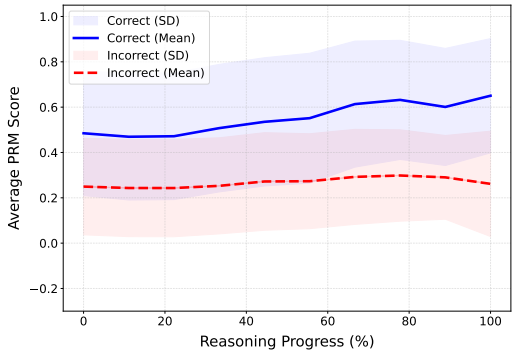
[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[17] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. **arXiv preprint arXiv:2103.03874**, 2021.

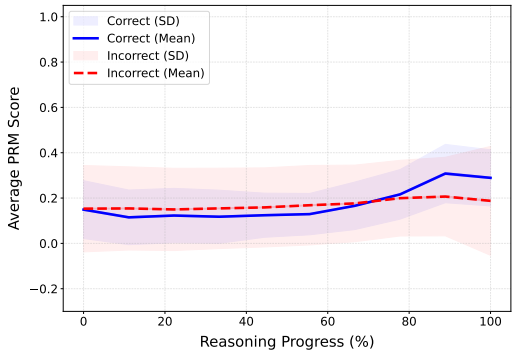
[18] HuggingFaceH4. Huggingfaceh4/aime_2024. Hugging Face Dataset, 2024. https://huggingface.co/datasets/HuggingFaceH4/aime_2024.

[19] math ai. math-ai/aime25. Hugging Face Dataset, 2025. https://huggingface.co/datasets/math-ai/aime25.

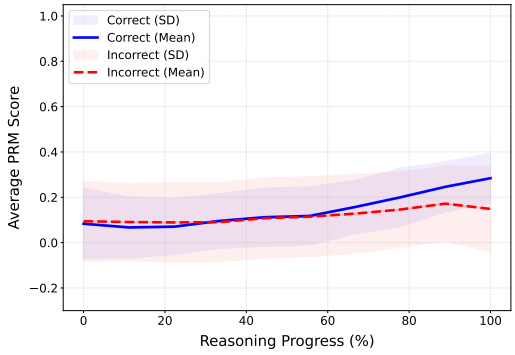
This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).



(a) MATH500



(b) AIME 24



(c) AIME 25

図 3: 推論進行度に対する平均 PRM スコアの推移. 正解パス（青・実線）と不正解パス（赤・破線）の比較を示す．なお, 線の背景薄色は標準偏差である.

A　推論プロセスの監視能力

MATH500 における事例（図 4）は，PRM が文脈内の論理的整合性を監視し得ることを示す興味深い一例である．モデルは **Step 2** で「$\angle T = 90°$」という誤った前提を生成したが，PRM はこの時点ではスコアを維持した (0.159)．しかし，**Step 3** で「RT が斜辺」と記述し，直前の前提（$T = 90°$ なら斜辺は RS）と幾何学的に矛盾した瞬間，スコアは –0.81 へ急落した．この挙動は，PRM が単なる事実との照合以上に，推論パス内部の自己矛盾（Self-Consistency）に対して感度良く反応する可能性を示唆している．

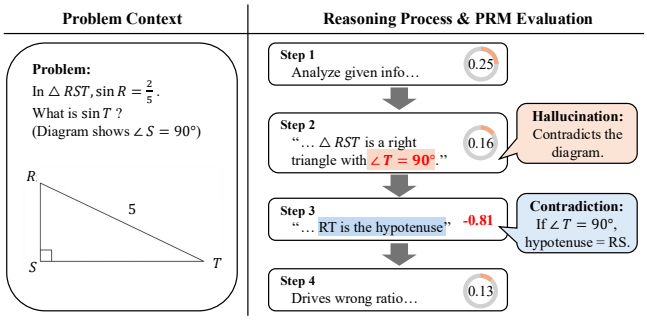


図 4: MATH500 における事例分析例. Step 2 で図と矛盾する前提（幻覚）を生成し, Step 3 でその前提と整合しない記述（自己矛盾）を行った瞬間, PRM スコアが急落している.

表 2: 解答強制（Prompted Forcing）の有無によるアブレーション研究（正答率, %）. $N \in \{16, 64\}$ のうち高い方の値を採用し, PRM 4 モデル（Qwen, Llama, DeepSeek, Ensemble）に基づき算出.

Benchmark	Metric	w/o Forcing	w/ Forcing	Δ
MATH500	Best	59.32	**61.68**	+2.36
	Avg	58.62	**60.84**	+2.22
AIME 24	Best	18.67	**20.67**	+2.00
	Avg	**16.83**	16.50	-0.33
AIME 25	Best	12.00	**16.00**	+4.00
	Avg	8.33	**12.17**	+3.83

B　推論進行度とスコア軌跡

推論進行度に対する PRM (ensemble) のスコア推移（図 3）を分析した．MATH500 (a) では正解パスが初期から高スコアを維持する理想的な挙動を示したが，難関 AIME 24 (b) では序盤に正解スコアが下回る「逆転現象」が見られた．これは難問の非自明な論理に対する一時的な保守的評価を示唆するが，進行度 70% 以降で急上昇して逆転しており，最終的な論理整合性は捕捉できている．この終盤の急峻な報酬勾配は，強化学習における高密度なフィードバック信号として極めて有望な特性である．この傾向は AIME 25（c）においても同様である.

C　Prompted Forcing の効果

表 2 に，解答強制（Prompted Forcing）の導入効果を示す．全タスクで最高精度（Best）が向上し，特に最難関の AIME 25 では +4.00pt の大幅な改善を達成した．AIME 24 の平均値（Avg）で僅かな低下が見られるものの，全体的なピーク性能の向上は，各ステップから正答への到達可能性を直接評価する本手法が，難問における報酬の識別性を高めていることを裏付けている.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).