正解保証を伴う思考プロセス付き合成データ生成による日本語大規模言語モデルの数学推論能力向上

岡田龍樹　平川雅人　大葉大輔
株式会社 ELYZA
{tatsuki.okada, masato.hirakawa, daisuke.oba}@elyza.ai

概要

日本語に特化した大規模言語モデル（日本語 LLM）の日本語数学能力を改善するには、高品質な学習データを大量に用意することが必要である。本研究では、任意の英語数学問題-回答ペアをシードに、日本語で記述された思考過程付きの数学学習データを、出力の正解を保証しながら半自動で合成する方法を提案した。また、実際に 英語数学データに対して提案手法を適用し、約 17 万件の学習データを合成した。PRM800K および GSM8K の日本語翻訳版を用いた評価では、提案手法により合成された学習データは、日本語 LLM (e.g., Llama-3-ELYZA-JP-8B) の日本語数学推論能力を確かに改善することを示した。その過程で、学習データが日本語で記述されていることの有用性、データ合成時における正解保証の有用性を示した。

1　はじめに

近年、大規模言語モデル（LLM）の高い汎用性が注目され、自然言語処理のみならず、多様な分野での応用が期待されている。しかし、数学タスクのように多段階の論理的推論を要する問題は依然として難易度が高く、主に英語圏において、英語数学タスクの性能向上を図る試みが盛んである [1][2][3][4][5][6]。

一方、日本語特化の LLM（日本語 LLM）が、日本語を用いて数学タスクを解く能力については、十分に検証されていない。実際、予備実験の結果（表 1）や、日本語 LLM による Japanese MT-Bench の数学カテゴリの性能分布[1)]と、英語圏モデルによる MT-Bench の性能分布 [2)]を見ても、日本語 LLM の日本語で数学を解く性能水準は、英語圏の LLM が英語で数学を解くものには達していないと推測される。

この性能差の大きな要因として、日本語の数学データ（特に思考過程を含むデータ）が英語に比べて圧倒的に不足していることが挙げられる。英語数学データを対象に、推論の思考過程を付記した合成データを作成する手法がある [1][2] が、合成されたデータの出力内容の正確性を保証していない。一方、出力が正しいかは数学に大きく影響を与えることが考えられる。

本研究では、英語数学データセットである PRM800K[7] および GSM8K[8] を日本語に翻訳したデータをシードとして利用し、合成内容の正確性を保証しながら思考過程付き日本語数学データセットを合成する手法を提案する。実験では、約 17 万件の日本語数学タスクの高品質なデータセットを実際に構築し、日本語 LLM である Llama-3-ELYZA-JP-8B[3)] を追加学習した結果、データが思考過程を含む日本語で記述されていること、およびデータ合成時に正解を保証することの有用性を示した。

2　関連研究

日本語にも対応した数学タスク評価データに MGSM [9] がある。これは、GSM8K を 10 カ国語に翻訳したデータセットであるが、各言語あたり約 250 件とサンプル数が少なく、問題の難易度も小学生レベルに留まっている。本論文ではより高度な数学的推論も対象にした、PRM800K も用いて検証を行う。

数学を含む科学領域のデータセットを合成する手法に MAmmoTH2 [1] がある。これは、Web の数学関連文書を抽出し、複数 LLM で QA 形式に整形・補完する。ただし、合成出力の正確性には課題が残る。

1） https://swallow-llm.github.io/evaluation/index.ja.html
2） https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard
3） https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

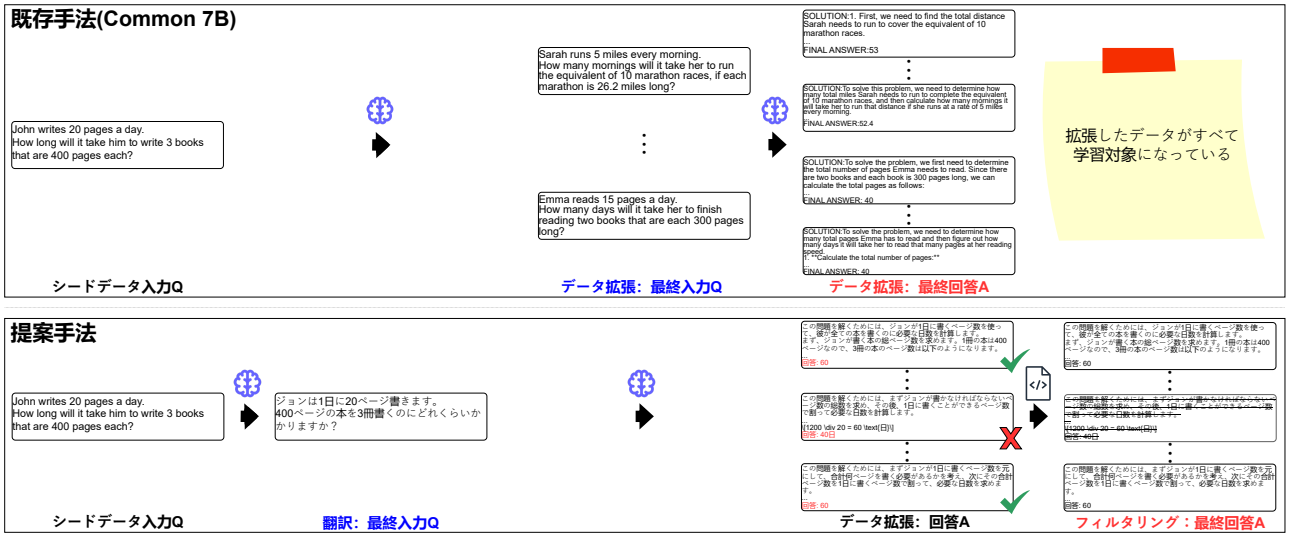

図 1　既存手法と提案手法の比較概念図。既存手法では生成された出力の品質保証が明確でないのに対し、提案手法では生成された出力を数式や計算過程をもとに検証する仕組みを導入している。

3　提案手法

提案手法は、英語の数学データセットから、正解が保証された思考過程付きの日本語数学データを生成するものである。具体的には、データセットの構築 (3.1 節) と学習形式の設計 (3.2 節) からなる。

3.1　データセットの構築

日本語 LLM の日本語数学タスク性能を向上させるため、Chain-of-Thought (CoT) 形式を用いた高品質な日本語数学データセットを新規に構築した。このデータセットは、以下の作成ステップである。

1. **英語データセットの選定と翻訳**
　ベースとなる英語データセットとして、英語数学タスクで用いられる PRM800K および GSM8K を選定した。これらのデータセットは、代数、幾何、微積分などの高度な数学タスクから四則演算などの基礎的なタスクまでを網羅している。英語から日本語への翻訳には、Apache ライセンス且つ日本語数学の性能が高い、Qwen2-7B-Instruct [4] を用い、高品質な翻訳を生成するようにプロンプトを工夫した。翻訳には few-shot サンプルを用いて、文脈と一貫性を重視した。
2. **Chain-of-Thought (CoT) 形式の応答生成**
　翻訳後のデータセットに対して、「英語データセットの選定と翻訳」同様に Qwen2-7B-Instruct を用いてシステムプロンプトに最終的な回答を 1 番最後の行に生成する CoT 形式であることを指定したうえで、few-shot を与え、思考過程を含む応答を生成した。生成の際、生成パラメータの temperature を 0.7 に生成件数の n に 20 を調整し、多様性と正確性をバランスよく確保した。
3. **生成結果の正解性確保**
　生成された応答データの正確性を確保するため、データセットに付属する正解と一致する応答のみを選別した。選別には、ルールベースで最終回答を抽出し、SymPy [5] を用いて、LaTeX などの複数の表現の中で一致するものがあるかを判定した。その結果、学習対象として利用可能なデータは約 17 万件となった。実際に学習に用いたデータ件数は、表 5 の EpochData に示す。

3.2　学習形式の設計

構築したデータセットを用いて、以下の学習形式を設計した。特に、日本語および英語での学習と評価、QA 形式と CoT 形式の比較、Rejection sampling Fine-Tuning (RFT) [10] を用いた実験を行った。

1. **日本語および英語データを用いた学習**
　日本語 LLM の性能向上を目指す中で、英語データの活用も重要視した。PRM800K を日本語に翻訳したデータセットと、元の英語データセットをそれぞれ用いて学習を行い、モデルが

4)　https://huggingface.co/Qwen/Qwen2-7B-Instruct

5)　https://github.com/sympy/sympy

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 実験全体の表。MaxSteps は 1QA の最大学習可能推論過程数、EpochData は 1Epoch での学習データ数、CoT は Chain of Thought 形式のデータを用いているか、Lang は学習データの言語、P800K@1 は PRM800K-Accuracy@1(%)、P800K@10 は PRM800K-Accuracy@10(%)、G8K@1 は GSM8K-Accuracy@1(%)、G8K@10 は GSM8K-Accuracy@10(%)、JPRate は生成される日本語の割合 (%) を示している。*は英語版の GSM8K を解いたときの結果を示す。

Model	MaxSteps	EpochData	CoT	Lang	P800K@1	P800K@10	G8K@1	G8K@10	JPRate
Qwen2-7B-Instruct		0			-	-	63.00*	-	-
Meta-Llama-3-8B-Instruct		0			-	-	69.01*	-	-
Llama-3-ELYZA-JP-8B		0			12.50	35.40	39.20	79.00	100.00
		12419	FALSE	EN	25.60	51.78	42.13	81.03	0.00
		12419		JA	16.28	45.57	32.38	79.38	100.00
	1	7180	TRUE	EN	33.14	64.95	51.94	85.52	0.00
	2	13655			36.99	65.83	57.17	87.42	0.00
	3	19616			36.69	67.76	57.25	85.90	0.00
	4	25139			41.72	67.02	58.69	86.59	0.00
	5	30299			41.13	67.16	60.51	86.59	0.00
	1	5044		JA	24.27	50.00	49.66	84.16	100.00
	2	9539			25.60	49.86	47.69	82.49	100.00
	3	13704			28.26	51.63	52.70	83.02	100.00
	4	17559			27.52	50.74	49.59	83.48	100.00
	5	21163			28.26	50.89	48.98	83.78	100.00

どの言語で学習した場合に性能が向上するかを比較した。また、英語と日本語のデータセットを組み合わせて学習させることで、学習データの多様性を確保しつつ、日本語 LLM の性能改善を図った。

2. **QA 形式と CoT 形式の比較**

   データセットの応答形式として、QA 形式と、CoT 形式の 2 つの形式を採用した。QA 形式では最終的な回答のみを学習データに含める一方、CoT 形式では回答までの思考過程を学習させた。これにより、日本語数学タスクにおける思考過程の重要性を評価し、CoT 形式がモデル性能に与える影響を実験的に検証した。

3. **CoT 形式データの Rejection sampling Fine-Tuning (RFT)**

   CoT 形式のデータセットでは、生成された応答の品質を保証するために Rejection sampling Fine-Tuning (RFT) を採用した。RFT は、同じ入力に対して N 個の生成を行い、評価指標に基づいてスコア付けを行い、その結果に基づいて正解データを選定した上で Fine-Tuning を行う手法である。

   本研究では、RFT を用いて入力に対して複数の応答を生成し、データセットに付属する正解と一致する応答のみを選別して学習した。また、1 つの QA ペアに対して学習可能な件数を N と設定し、生成結果の品質と多様性のバランスを評価することで、モデルが正確性と推論能力を向上できるよう設計した。

4. **評価の設定**

   評価には PRM800K および GSM8K の日本語翻訳版をテストセットとして使用し、モデルの性能を詳細に比較・分析した。評価指標としては、モデルに 10 件の回答を生成させ、その中で 1 件でも最終的な回答（最後の行）がデータセット作成時に付属された正解と一致するかどうかで正答率を計測した。

4 評価実験

本章では、構築したデータセットを用いて日本語および英語データでモデルを学習させた結果を報告する。特に、日本語翻訳済みデータの学習結果、英語データの学習結果、それらを用いた CoT 形式、正解保証の有無での性能比較を行う。

4.1 日本語翻訳済みデータの学習結果

日本語翻訳済みデータセットを用いた学習の結果を表 1 に示す。Llama-3-ELYZA-JP-8B を日本語の PRM800K で学習させると、QA ペアの学習可能件数が増えるにつれ、P800K@1 は 16.28%から 28.26%に改善された。さらに、学習データ量が増加すると、GSM8K に対する精度（G8K@1）も 49.66%から 52.70%に向上した。この結果は、日本語翻訳データが Llama-3-ELYZA-JP-8B の日本語数学タスク性能向上に寄与することを示す。ただし、学習可能件数を増やしても一部の評価指標で飽和が見られ、データの多様性や品質の向上が性能改善の鍵となる。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　正解保証のないデータが混ざった場合のモデル性能

DataSet	EpochData	P800K@1	P800K@10	G8K@1	G8K@10	JPRate
PRM800K (True)	21163	28.26	50.89	48.98	83.78	100.00
PRM800K (True/False)	21681	20.72	46.45	28.51	69.98	100.00

表 3　学習言語による性能比較

Lang	EpochData	P800K@1	P800K@10	G8K@1	G8K@10	JPRate
JA	21163	28.26	50.89	48.98	83.78	100.00
EN	30299	41.13	67.16	60.51	86.59	0.00
JA/EN	22205	17.31	45.86	31.77	75.44	99.92

4.2　英語データの学習結果

英語の PRM800K を用いた学習の結果を表 1 に示す。Llama-3-ELYZA-JP-8B を英語データで学習させた場合、日本語データと比較して P800K@1 が最大 41.13%と大幅に向上した。また、G8K@1 も 60.51%と、英語データがモデルの性能向上に寄与することが分かる。

これは、英語データセットが日本語データセットよりもタスクの多様性や品質が高い可能性を示唆している。ただし、生成された応答が英語であるため、日本語生成における一貫性を維持するためには追加の適応学習が必要である。

4.3　日本語翻訳済みデータと英語データの比較

日本語翻訳済みデータと英語データを用いた学習の性能を比較すると、英語データで学習したモデルは全体的に高い精度を示す。一方、日本語翻訳済みデータを用いたモデルは、CLD3[6] を用いて計測した生成される日本語の割合が 100%であり、一貫性が高いことが特徴である（表 1）。特に、G8K@1 では、英語データの 60.51%に対し、日本語データは 52.70%に留まったものの、日本語生成の品質を重視する応用においては日本語データの利用が有利である。

また、CoT 形式の応答を用いた Rejection sampling Fine-Tuning (RFT) の適用により、性能がさらに向上する傾向が見られた。例えば、日本語データセットを用いた場合でも、生成結果の選別により正解率の向上が確認された。

4.4　正解保証のないデータが混ざった場合の影響

実験では、Llama-3-ELYZA-JP-8B モデルを使用し、最大ステップ数を 5、言語設定を日本語と固定した。表 2 に示すように、正解保証のないデータが含まれる場合、P800K@1 が 28.26%から 20.72%へ、G8K@1 が 48.98%から 28.51%へと大幅に低下した。この結果は、学習データの正確性がモデル性能に直接的な影響を与えることを示している。

特に、誤った応答が混在する場合、モデルが不正確な推論を学習するリスクが高まるため、データ品質の確保が不可欠であることが分かった。

4.5　学習言語によるの性能比較

実験では、Llama-3-ELYZA-JP-8B モデルを使用し、最大ステップ数を 5、CoT を有効にした状態で、データの言語の割合を英語と日本語でほぼ 1 対 1 に調整して評価を行い。日本語以外のデータを活かすことができるかどうかを検証した。表 3 に示すように、日本語データを使用した場合の性能は P800K@1 で 28.26%、英語データでは 41.13%、日本語と英語を組み合わせた場合は 17.31%となった。(英語データを使用した場合には日本語での生成が不可能となっているため課題が残る。)

この結果から、日本語と英語を組み合わせた場合には性能劣化を引き起こしてしまうことがわかった。

5　おわりに

本研究では、日本語数学タスクの改善をするために、英語数学データセットを日本語に翻訳し、思考過程を含めた日本語数学データを作成する方法を提案した。約 17 万件のデータを使い日本語 LLM を追加学習させた結果、数学タスクの推論性能が向上した。特に、思考過程付きデータがモデルの正確性向上に貢献することを示した。

6）https://github.com/google/cld3

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. **arXiv preprint arXiv:2405.03548**, 2024.

[2] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. **arXiv preprint arXiv:2403.04706**, 2024.

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[4] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. **arXiv preprint arXiv:2309.12284**, 2023.

[5] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. **arXiv preprint arXiv:2402.06332**, 2024.

[6] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. **arXiv preprint arXiv:2308.09583**, 2023.

[7] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. **arXiv preprint arXiv:2305.20050**, 2023.

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

[9] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. **arXiv preprint arXiv:2210.03057**, 2022.

[10] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.

[11] Han Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会 (NLP2024), 2024.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A 分析

A.1 データセット拡張による性能変化

表 4 データセット拡張による性能変化

DataSet	EpochData	P800K@1	G8K@1
PRM800K	21163	28.26	48.98
+ GSM8K	46745	29.54	57.47

表 4 に示す通り、Llama-3-ELYZA-JP-8B モデルを使用し、最大ステップ数を 5、言語設定を日本語に固定して実験をした。PRM800K のみに基づく学習データと、PRM800K と GSM8K を組み合わせたデータセットで学習を行った場合の性能を比較した。PRM800K のみを使用した場合と PRM800K + GSM8K の組み合わせを使用した場合、P800K@1 や G8K@1 において大きな性能向上は見られなかった。

この結果から、PRM800K に加えて GSM8K を追加しても、学習可能なデータの増加がそのまま性能向上に繋がるわけではないことが示唆される。一方で、生成データの多様性や応用可能性の向上に寄与する可能性があるため、さらなる評価が必要である。

PRM800K と GSM8K は、それぞれ異なる難易度の問題を含んでおり、この差が性能に影響を与える要因となる可能性がある。特に、PRM800K は高度な推論を要する数学タスクが多く含まれる一方で、GSM8K は比較的基礎的な数学問題が中心である。そのため、両データセットを組み合わせることで全体のデータの多様性は向上するが、モデルが異なる難易度の問題を均等に学習することが難しくなる可能性がある。

これを踏まえると、データセットの拡張に際しては、単純なデータ量の増加だけでなく、難易度バランスやデータの多様性、学習プロセスの調整が重要である。例えば、段階的な学習戦略や難易度に応じた重み付けを導入することで、拡張されたデータセットの効果を最大限に引き出すことができると考えられる。

A.2 日本語のみ学習モデルと英語も学習したモデルの性能比較

この実験では、Llama-3-ELYZA-JP-8B と cyberagent/open-calm-7b [7] を使用し、最大ステップ数を 5、CoT を有効にして性能を比較した。言語設定は日本語（JA）と英語（EN）、およびその両方（JA/EN）で行った。

表 1 を見ると、英語で学習した方が日本語よりも最終的な回答のスコアが高くなっている。この結果は、事前学習時の言語が影響している可能性があると考えられるため、日本語データのみで学習したモデルと open-calm-7b を用いて比較を行った。

表 5 日本語のみ学習モデルと英語も学習したモデルの性能比較

Model	EpochData	Lang	P800K@10	G8K@10	JPRate
Llama-3-ELYZA	21163	JA	50.89	83.78	100.00
-JP-8B	30299	EN	67.16	86.59	0.00
			35.40	79.00	100.00
open-calm-7b	21163	JA	13.62	10.16	100.00
	30299	EN	11.40	8.80	0.26
			0.15	0.00	100.00

open-calm-7b は日本語のみで事前学習を行ったモデルであり、日本語データ（PRM800K）を用いた場合、P800K@10 が 13.62%、G8K@10 が 10.16%と非常に低い性能を示した。同モデルを英語データで学習した場合も、P800K@10 が 11.40%、G8K@10 が 8.80%にとどまり、日本語データでの学習の方が良い結果を得ることができた。

open-calm-7b の性能が Llama-3-ELYZA-JP-8B に比べて低くなっていることは、他のベンチマークでも確認されている。具体的には、open-calm-7b は llm-jp-eval(v1.2.0) の STS を除く平均において 0.224 ポイントを記録しており [11]、全体的な言語理解タスクでの基本性能が他のモデルと比較して劣っていることが示されている。一方、Llama-3-ELYZA-JP-8B モデルは同様のベンチマークで 0.662 ポイントを達成しており、この基本性能の差が本研究で使用した日本語数学タスクにおける性能差の一因と考えられる。このように、open-calm-7b と Llama-3-ELYZA-JP-8B の間で基本的な性能の違いが他のベンチマークでも報告されているため、それが日本語数学タスクの結果にも反映されているのは自然である。

7） https://huggingface.co/cyberagent/open-calm-7b

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).