日本語大規模言語モデルの推論能力向上のための強化学習データセット構築

太田 晋 [1], 片山 結太 [1], 水木 栄 [1,2], 岡崎 直観 [1,2,3]

[1] 東京科学大学, [2] 産業技術総合研究所, [3] NII LLMC

言語処理学会第32回年次大会 併設ワークショップ
日本語言語資源の構築と利用性の向上 (JLR2026) 2026-03-13

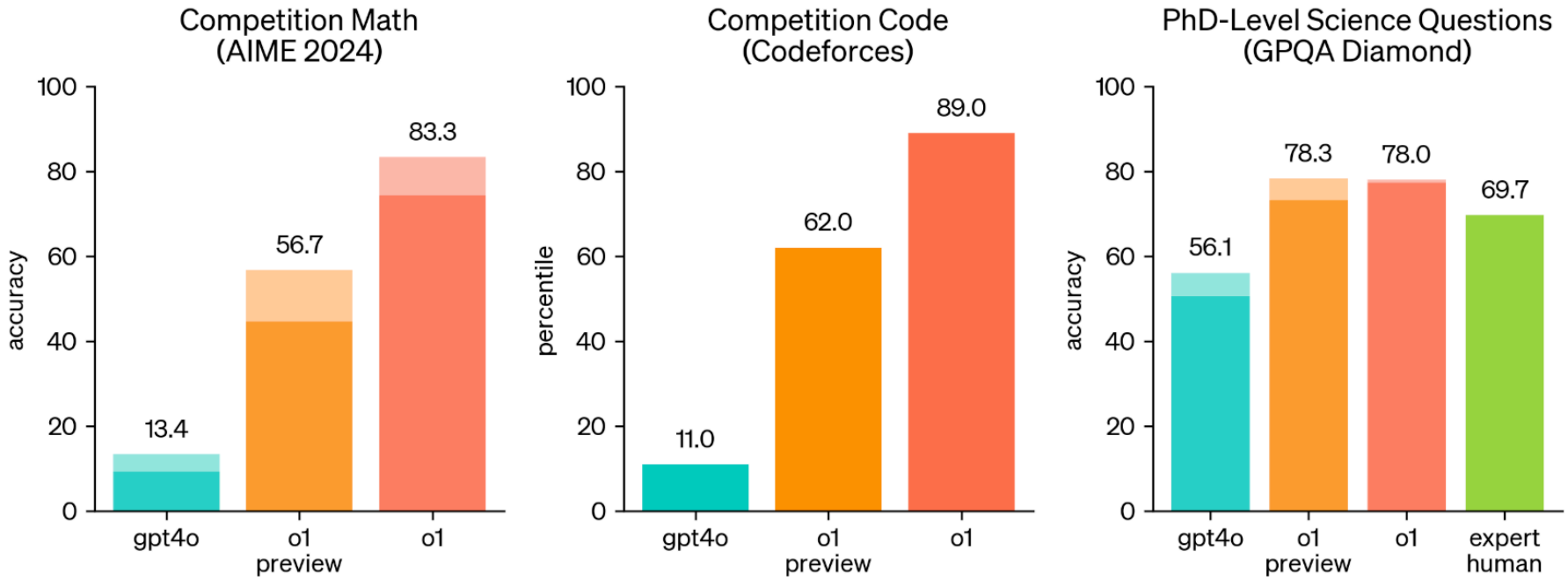
データセット公開URL:
https://huggingface.co/datasets/tokyotech-llm/s1-test-time-scaling-synth-public

概要

- 背景
  - 推論型モデルによる数学・科学・コード生成タスクの大幅な性能向上
  - 検証可能な報酬による強化学習(RLVR)による推論効率改善
- 目的
  - 日本語タスクの推論能力向上のために、英語RLVRデータセットによる学習のみで十分か、それとも日本語RLVRデータセットによる学習が必要かを明らかにする
- 方法
  - s1データセット
  - 日英RLVRデータセット構築
    - 設問の邦訳
    - 検証可能な解答の抽出
    - 設問・解答ペアがRLVRに適しているかどうか判定
- 実験
  - 構築した日・英RLVRデータセットに対してそれぞれRLVRを実施し日本語ベンチマークで評価
  - 日本語RLVRデータセットで学習したモデルが日本語ベンチマークにおいて一貫して性能向上

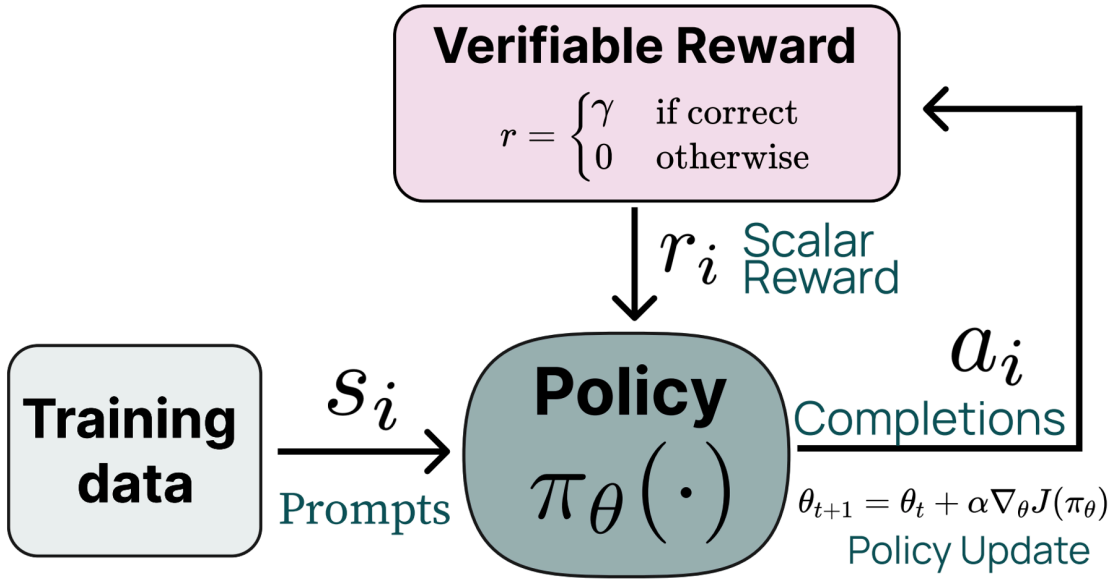
大規模言語モデル(LLM)の推論能力向上

- 推論型モデル
  - 「深い推論」(reasoning)を含む応答の生成
  - **数学・科学・コード生成**などの高度なタスクにおいて大幅な性能向上を達成 [OpenAI, 2024; Guo+, 2025]



RLVRによる推論効率改善

- 検証可能な報酬による強化学習 (RLVR) [Lambert+, 2025]
  - 従来のRLHFの報酬モデルを**検証可能な報酬関数**に置き換え
- RLVRデータセットの構成
  - 設問
  - 解答 (推論過程を含まない)
  - 検証可能な報酬関数



検証可能な報酬による強化学習(Reinforcement Learning with Verifiable Rewards; RLVR)

目的

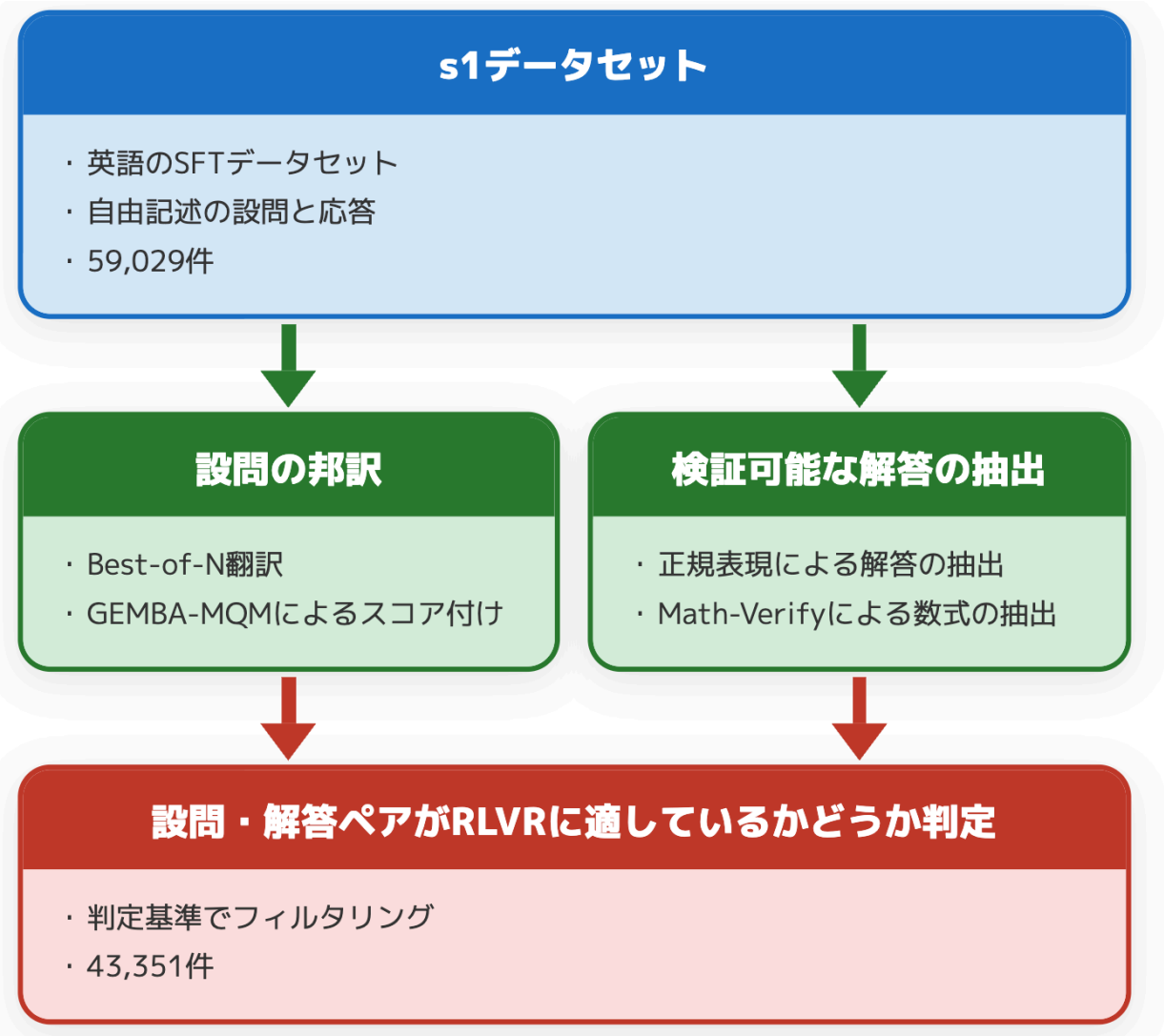
- 先行研究の課題
  - RLVRに用いられている設問・解答データセットの多くは**英語**が対象 [Team Olmo, 2025; NVIDIA, 2025]
- 本研究の目的
  - 日本語タスクの推論能力向上のために、英語RLVRデータセットによる学習のみで十分か、それとも**日本語RLVRデータセットによる学習が必要かを明らかにする**

Dolci-Think-RLデータセットの内訳 [Team Olmo, 2025]

Category	Prompt Dataset	# Prompts Used in Think RL	# Prompts Used in Instruct RL	Reference
Precise IF	IF-RLVR	30,186	38,000	Pyatkin et al. (2025)
Math	Open-Reasoner-Zero	3,000	14,000	Hu et al. (2025)
	DAPO-Math	2,584	7,000	Yu et al. (2025)
	AceReason-Math	6,602	–	Chen et al. (2025)
	Polaris-Dataset	–	14,000	An et al. (2025)
	KlearReasoner-MathSub	3,000	9,000	Su et al. (2025c)
	OMEGA-train	15,000	20,000	Sun et al. (2025)
Coding	AceCoder	9,767	20,000	Zeng et al. (2025a)
	KlearReasoner-Code	8,040	–	Su et al. (2025c)
	Nemotron Post-training Code	2,303	–	NVIDIA AI (2025)
	SYNTHETIC-2	3,000	–	PrimeIntellect (2025)
General Chat	Tulu 3 SFT	7,129	18,955	Lambert et al. (2024)
	Wildchat-4.8M	7,129	18,761	-
	Multi-Subject RLVR	7,129	12,234	Su et al. (2025b)
Total		104,869	171,950	

方法

- s1データセット
- 日英RLVRデータセット構築
  - 設問の邦訳
  - 検証可能な解答の抽出
  - 設問・解答ペアがRLVRに適しているかどうか判定



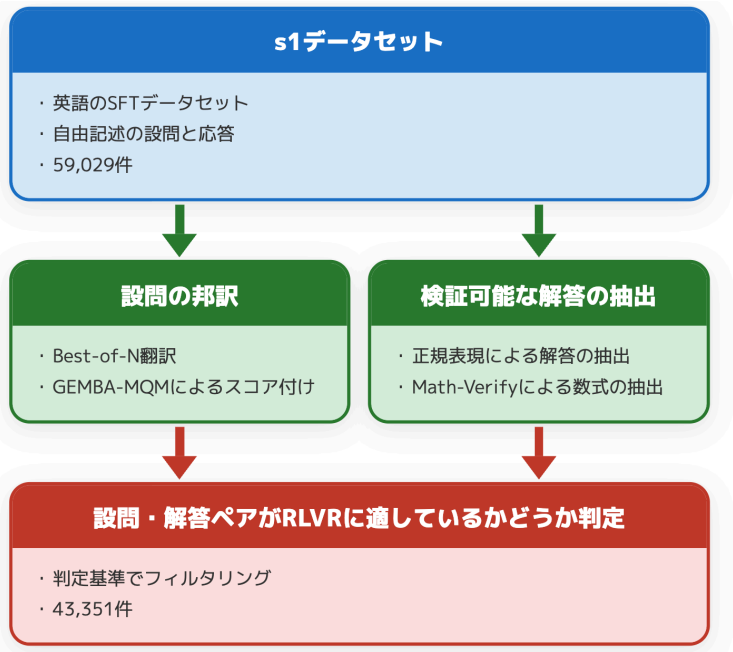
s1データセット

- s1: Simple Test-Time Scaling データセット [Muennighoff+, 2025]
  - 16個の情報源から収集された59,029件の**英語**設問を含む**SFTデータセット**
  - 問題点: RLVRの報酬関数で検証可能な解答が提供されているわけではない
    - 例: 証明問題、自由記述の応答など
    - RLVRで必要なのは、設問と検証可能な解答のペア
- データ例
  - RLVRのためには自由記述の応答から38という**解答だけを抽出する必要がある**

設問	応答
The symbol $\lfloor x \rfloor$ denotes the largest integer not exceeding x. For example, $\lfloor 3 \rfloor = 3$, and $\lfloor 9/2 \rfloor = 4$. Compute $\lfloor \sqrt{1} \rfloor + \lfloor \sqrt{2} \rfloor + \lfloor \sqrt{3} \rfloor + \cdots + \lfloor \sqrt{16} \rfloor$.	Since $1 \leq \sqrt{1} < \sqrt{2} < \sqrt{3} < 2$, the first three terms of the sum are equal to 1. Then, since $2 \leq \sqrt{4} < \sqrt{5} < \cdots < \sqrt{8} < 3$, the next five terms equal 2. Then, since $3 \leq \sqrt{9} < \sqrt{10} < \cdots < \sqrt{15} < 4$, the next seven terms equal 3. Finally, the last term equals $\lfloor 4 \rfloor = 4$. So the overall sum is $3(1) + 5(2) + 7(3) + 4 = 3 + 10 + 21 + 4 = \boxed{38}$.

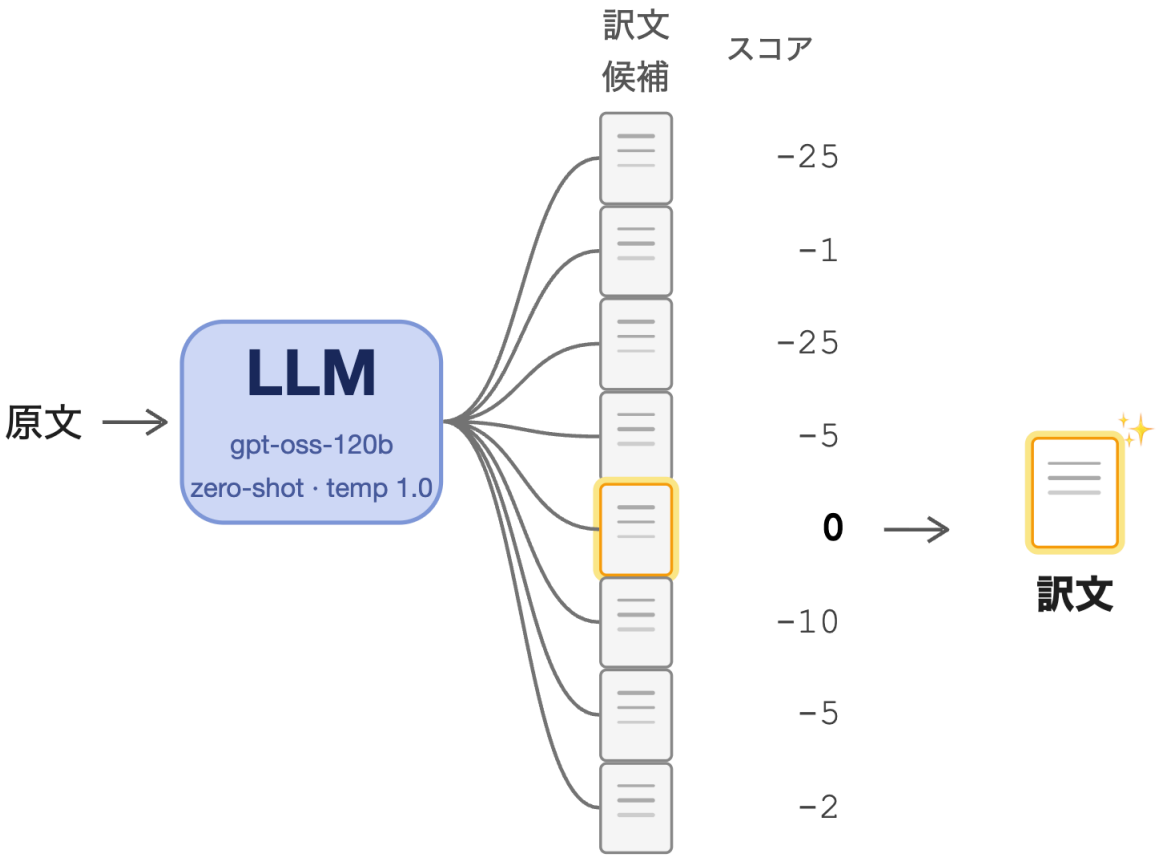
日英RLVRデータセット構築

- s1データセットを基に、RLVRに適した日英の設問・解答データセットを構築
  - (1) 設問の邦訳
  - (2) 検証可能な解答の抽出
  - (3) 設問・解答ペアがRLVRに適しているかどうか判定



設問の邦訳

- Best-of-N翻訳
  - 8個の訳文候補を生成
    - gpt-oss-120b [OpenAI, 2025]
    - 推論設定 `medium`
    - 温度1.0
  - 翻訳品質のスコア付け
    - GEMBA-MQM [Kocmi+, 2023]
  - 最も高いスコアを持つ訳文を採用



Best-of-N翻訳: GEMBA-MQMにより訳文候補をスコア付けし、最も高いスコアの訳文を採用

GEMBA-MQM による翻訳品質評価

- GEMBA-MQM [Kocmi+, 2023] によるスコア付け
  - `accuracy` , `fluency` , `style` , `terminology` , `other errors` カテゴリそれぞれに対して、 `minor` (-1) / `major` (-5) / `critical` (-25) の3段階で減点し、合計スコアを算出 $[-25, 0]$
- スコア例
  - id 1 は翻訳誤りなしのため 0
  - id 2 は `Find the value of ...` の訳が抜けているため -25
  - id 3 は 遊び場を作るリチャード。 という文が不自然なため -5

id	score	原文	訳文
1	0	Simplify $2\cos^2(\ln(2009)i) + i\sin(\ln(4036081)i)$.	$2\cos^2(\ln(2009)i) + i\sin(\ln(4036081)i)$ を簡単化せよ。
2	-25	Let a be a positive constant. Find the value of $\ln a$ such that $\frac{\int_1^e \ln(ax)\,dx}{\int_1^e x\,dx} = \int_1^e \frac{\ln(ax)}{x}\,dx.$	a を正の定数とする。 $\frac{\int_1^e \ln(ax)\,dx}{\int_1^e x\,dx} = \int_1^e \frac{\ln(ax)}{x}\,dx.$
3	-5	Richard is building a rectangular playground from 200 feet of fencing. The fencing must entirely enclose the playground. What is the maximum area of this playground?	200フィートのフェンスを使って、長方形の遊び場を作るリチャード。フェンスは遊び場を完全に囲まなければならない。遊び場の最大面積は何平方フィートか。

検証可能な解答の抽出

- s1データセットは16の情報源から収集されたSFT用のデータセット
  - 問題点: RLVRに適した報酬関数(ルールベース)で検証可能な**解答**が提供されているわけではない
- 検証可能な形式の解答フィールドがない場合、以下の方法で**自由記述の応答から解答を抽出**
  - 正規表現 (例: `answer is\s*([^\s]+)`)による抽出
  - Math-Verify [Kydlíček+, 2025] による LaTeX 形式の数式部分の抽出
  - 抽出結果を人手によりフィルタリング
- 抽出例

自由記述の応答	抽出した解答
... $S = 1 \times 3 + 2 \times 5 + 3 \times 7 + 4 \times 1 = 3 + 10 + 21 + 4 = 38.$ Final Answer: The final answer is $\boxed{38}$	\boxed{38}

設問・解答ペアがRLVRに適しているかどうか判定

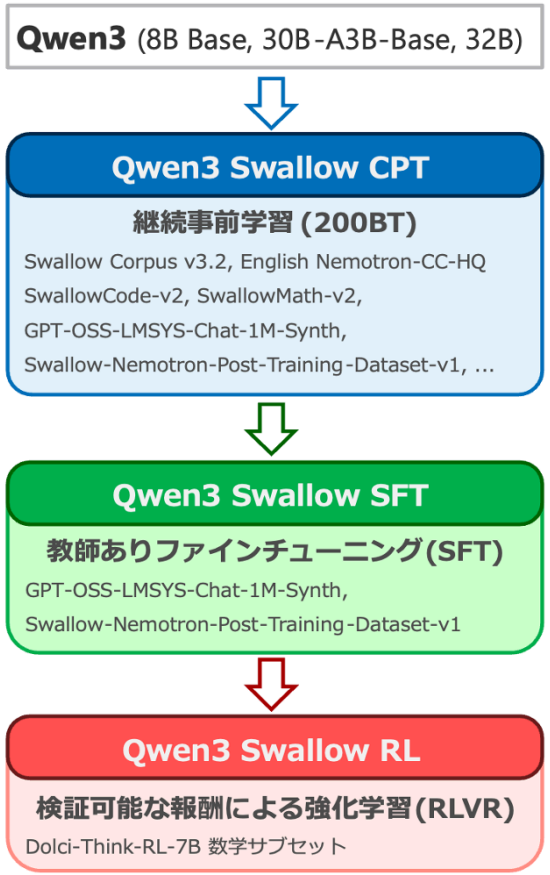
- 設問の邦訳と解答抽出の結果を基に、RLVRの報酬関数で検証可能かどうか判定
- 判定基準
  - (1) 設問がルールベースの自動検証が困難な**証明問題**である場合、$False$
  - (2) 設問の邦訳の**翻訳品質が低い**(スコアが-10以下)場合、$False$
  - (3) 検証可能な**解答抽出に失敗**した場合、$False$
  - (4) それ以外の場合、$True$
- 上記の基準で設問・解答ペアをフィルタリングし、RLVRに適した日英それぞれ**43,351件**の設問・解答データセットを構築
  - 日本語設問のサブセットを `s1-Ja` 、英語設問のサブセットを `s1-En` とする
- データセット公開URL
  - https://huggingface.co/datasets/tokyotech-llm/s1-test-time-scaling-synth-public

実験

- 構築した日・英RLVRデータセットを用いてそれぞれRLVRを実施
- 日本語の推論能力を評価するためにベンチマークを測定

RLVR実験設定

- 学習対象モデル: Qwen3-Swallow-8B-SFT-v0.2 [水木+, 2026]
- データセット: 日本語(s1-Ja)、英語(s1-En)
- フレームワーク: slime [Zhu+, 2025]
- アルゴリズム: GRPO [Shao+, 2024] を基に、Dynamic Sampling, Clip-Higher [Yu+, 2025] を導入
- 学習率: 1.0×10^{-6}, グローバルバッチサイズ: 512, 学習ステップ数: 220, 回答のサンプリング数: 16, コンテキスト長: $24,576$
- 報酬関数: ルールベース
  - 応答文字列から区切りトークンで推論過程と回答を抽出
  - 回答が正解と数式的に同値である場合に 1、それ以外は 0



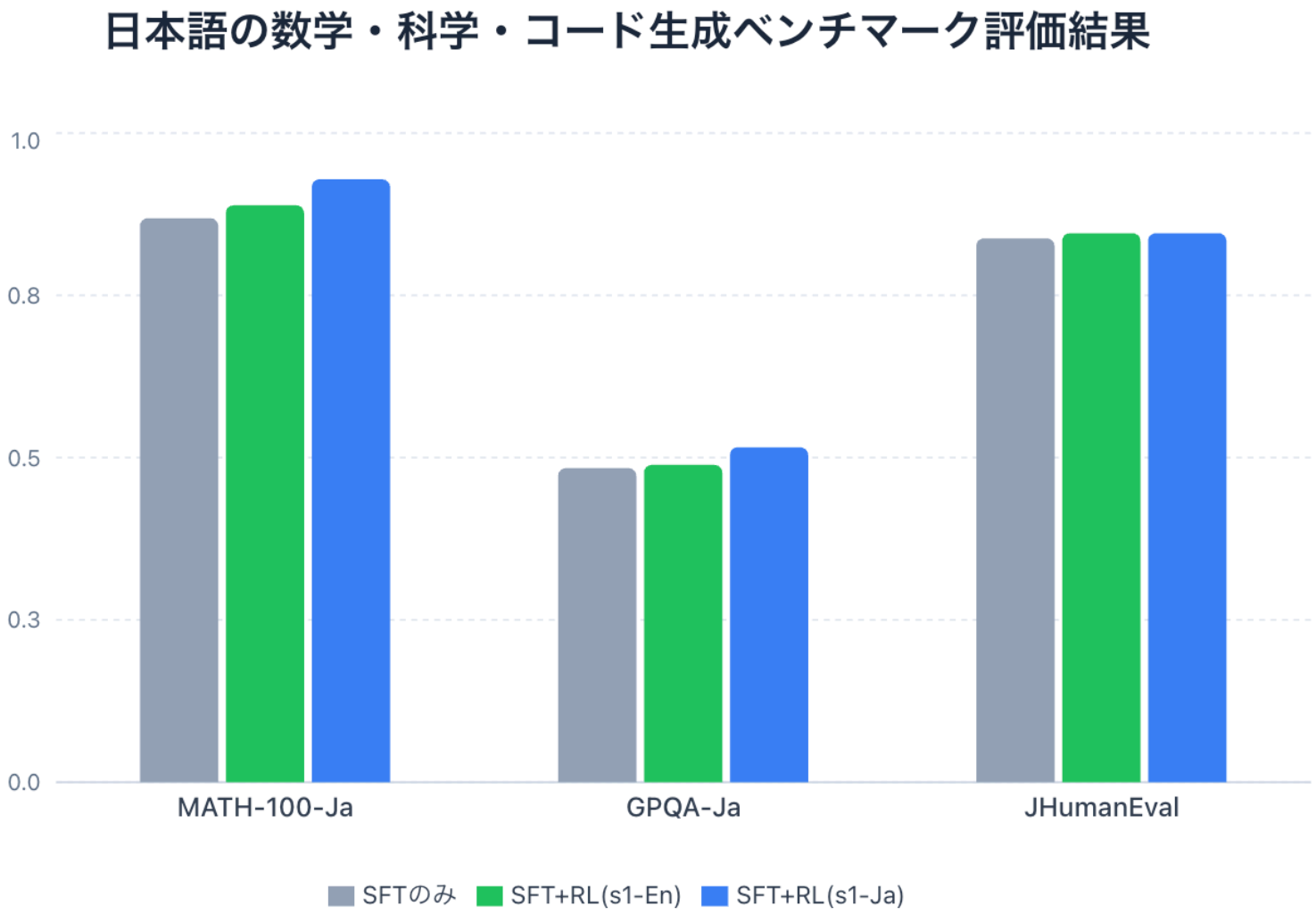
参考: Qwen3-Swallowの学習パイプライン

評価

- 日本語の推論能力を評価するために以下の3つのベンチマークを選択
  - MATH-100-Ja [Son+, 2025] (数学)
  - GPQA-Ja [Huang+, 2025] (科学)
  - JHumanEval [佐藤+, 2024] (コード生成)
- 評価フレームワーク
  - swallow-evaluation-instruct [Mizuki+, 2025]

実験結果

- SFT+RL(s1-Ja) (青色)
  - SFTのみ(灰色)と比べて
  - MATH-100-Jaで+6.0pt
  - GPQA-Jaで+3.2pt
- SFT+RL(s1-En) (緑色)
  - 改善幅は限定的
- JHumanEval
  - 言語間の差はほとんど見られず



考察

- 日本語RLVRデータセットで学習したモデルは、日本語ベンチマークにおいて一貫して性能向上
  - 日本語LLMの推論能力向上において、英語RLVRデータセットのみでは十分でなく、**日本語RLVRデータセットを併用した方がより効果的な可能性**
- JHumanEvalにおいては言語間の差はほとんど見られなかった
  - コード生成タスクでは自然言語よりもプログラミング言語の構文やアルゴリズムの理解が必要であり、設問言語の影響が相対的に小さい可能性
  - **タスク特性に応じて言語依存性の程度が異なる可能性**

まとめ

- RLVRに適した日英の設問・解答データセットを構築
  - 設問の邦訳
  - 検証可能な解答の抽出
  - 設問・解答ペアが検証可能かどうか判定
- RLVRを実施し日本語の推論能力をベンチマークで評価
  - 日本語RLVRデータセットで学習したモデルは、日本語の数学・科学・コード生成ベンチマークにおいて一貫して性能向上
  - 日本語LLMの推論能力向上において、英語RLVRデータセットのみでは十分でなく、日本語RLVRデータセットを併用した方がより効果的な可能性
- 今後の課題
  - 本研究の実験は単一のベースモデルに基づいており、モデルアーキテクチャや規模の違うモデルに対する検証が必要
  - より多様な推論タスクへの適用に加え、タスク間および言語間における推論能力の転移について検証を進める予定

参考文献

- [Guo+, 2025] Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning, Nature, Vol. 645, No. 8081, pp. 633–638 (2025)
- [Huang+, 2025] Huang, X., Zhu, W., Hu, H., He, C., Li, L., Huang, S., and Yuan, F.: BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models, in Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 16751–16774 (2025)
- [Kocmi+, 2023] Kocmi, T. and Federmann, C.: GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4, in Proceedings of the Eighth Conference on Machine Translation, pp. 768–775 (2023)
- [Kydlíček+, 2025] Kydlíček, H.: Math-Verify: Math Verification Library, https://github.com/huggingface/math-verify (2025)
- [Mizuki+, 2025] Mizuki, S., Saito, K., Oi, M., Ichinose, T., Matsushita, N., Miyamoto, S., Nguyen, T. D., and Moon, S.: 大規模言語モデル評価フレームワークswallow-evaluation-instruct v202510, https://github.com/swallow-llm/swallow-evaluation-instruct (2025)
- [Muennighoff+, 2025] Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candes, E., and Hashimoto, T.: s1: Simple test-time scaling, in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 20275–20321 (2025)
- [Lambert+, 2025] Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L.J.V., Liu, A., Dziri, N., Lyu, X., et al.: Tulu 3: Pushing Frontiers in Open Language Model Post-Training, in Proceedings of the Second Conference on Language Modeling (2025)
- [NVIDIA, 2025] NVIDIA: Llama-Nemotron: Efficient Reasoning Models, arXiv:2505.00949 (2025)
- [OpenAI, 2024] OpenAI: Learning to reason with LLMs, https://openai.com/index/learning-to-reason-with-llms/ (2024)
- [OpenAI, 2025] OpenAI: gpt-oss-120b & gpt-oss-20b Model Card, arXiv:2508.10925 (2025)
- [Qwen Team, 2025] Qwen Team: Qwen3 Technical Report, arXiv:2505.09388 (2025)
- [Son+, 2025] Son, G., Hong, J., Ko, H., and Thorne, J.: Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning, in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 14333–14368 (2025)
- [Team Olmo, 2025] Team Olmo: OLMo 3, arXiv:2512.13961 (2025)
- [Shao+, 2024] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., and Guo, D.: DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, arXiv:2402.03300 (2024)
- [Yu+, 2025] Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., et al.: DAPO: An Open-Source LLM Reinforcement Learning System at Scale, in Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (2025)
- [Zhu+, 2025] Zhu, Z., Xie, C., Lv, X., and Contributors, slime: slime: An LLM post-training framework for RL Scaling, https://github.com/THUDM/slime (2025), GitHub repository. Corresponding author: Xin Lv.
- [佐藤+, 2024] 佐藤美唯, 高野志歩, 梶浦照乃, 倉光君郎: LLM は日本語追加学習により言語間知識転移を起こすのか？, 言語処理学会第30 回年次大会(NLP2024) (2024)
- [水木+, 2026] 水木栄, 藤井一喜, 川村政貴, Dung, N. T., 片山結太, 齋藤幸史郎, 一瀬達矢, 宮本空, 松下直矢, 大井聖也, Ma, Y., 太田晋, 大葉大輔, 高村大也, 横田理央, 岡崎直観: 蒸留による日英推論型大規模言語モデル構築戦略の探索, 言語処理学会第32 回年次大会(NLP2026) (2026), to appear