

地域毎の教育課程に基づく多言語指示学習データの自動生成

† 金山 博 * Gabriele Prato † 趙 陽

† 日本アイ・ビー・エム株式会社 東京基礎研究所

* Chandar Research Lab, Mila – Quebec AI Institute, Université de Montréal

hkana@jp.ibm.com, pratobag@mila.quebec, yangzhao@ibm.com

概要

本稿では、LLM の問題解決の能力を向上させるために、多言語の高品質な訓練データを自動的に生成する方法について述べる。様々な分野を網羅するべく、LLM によって地域に特化した教育課程の内容を列挙させて、それに基づいた各単元の宿題や試験問題とその解答を生成させるという手順で、質問応答のペアを大量に獲得した。生成されたデータを LLM の指示学習に適用することにより、とりわけ MGSM の正解率が大幅に向上することが確認できたほか、複数の言語に対するモデルの振る舞いを検証する土台を作ることができた。

1 はじめに

LLM の高度化により、推論を伴うタスク解決の能力が向上している。しかし、英語以外の言語の性能が相対的に劣っていること [1, 2] や、英語圏の文化への偏重 [3, 4] が指摘されているため、多言語の学習データの拡充を図りたい。事前学習に用いるデータとして、数学に関する文書の効果が高いことが知られている [5, 6] ため、特に教育に関連するデータを増強したいと考えた。

学習資源の多言語化の手段として、既存の英語のデータを翻訳する方法 [7, 8, 9, 10] があるが、元となるデータの品質や自動翻訳の性能に依存することや、英語圏の知識や文化に偏るという欠点がある。人手でデータを収集または作成する方法 [11, 12, 13, 14] もあるが、多くの文化圏のデータを十分に確保するためには多大なコストを要する。

近年着目されているのが、強い LLM にプロンプトを与えてデータを自動生成する方法であり、パラメータ数が小さいモデルの学習に利用できる場面が多い。既存の質問応答のペアを与えてより複雑なものに拡張する手法が試みられている [15, 16, 17] が、多様なデータを大量に生成することには限界があ

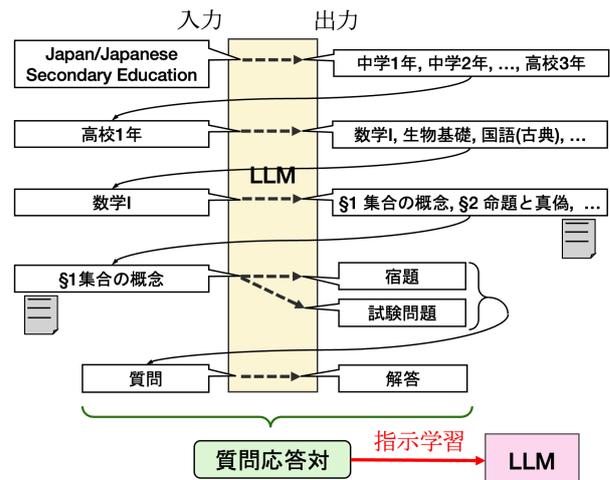


図 1: 指示学習データの階層的な生成の概念図。

る。そこで我々は、教育課程に関するデータを階層的に自動生成して、指示学習 (instruction tuning) に活用する方法を提案する。具体的には、指定した地域と言語に基づいた中等教育・高等教育の過程、すなわち学年・分野・科目・単元・学習内容を順に生成し、各科目の宿題と試験問題、そしてそれらの解答を LLM に生成させることによって、質問応答の対を獲得する。図 1 にその概念を示す。

これにより、多言語の高品質なデータを大量に獲得することが可能となる。特に、幅広い分野を網羅したり、難易度を制御したりすることもできる。また、指定した国や地域で学ぶ内容を対象言語で直接生成するため、対象言語の流暢さを保ちつつ、英語圏の文化への偏りを防げるという利点がある。

本稿ではまず、階層的なデータ生成の手順を述べる (2.1 節)。それにより、言語が異なる 7 地域の教育課程に基づく、各々約 70 万件の質問応答対を獲得した (2.2 節)。それを LLM の指示学習に適用したところ、web から収集した Aya の指示学習データと同等ないし大幅に上回る性能が得られた (3 節)。

```

{
  "course": "数学 I",
  "units": [
    {
      "unit_id": "ユニット 1",
      "name": "集合と論理",
      "overview": "集合の概念、記号、要素、部分集合、和・積・差、ベン図、命題と真偽、論理記号と真理値表の基礎を学ぶ。",
      "learning_objectives": "集合の表記ができ、集合演算を用いて問題を解ける。命題の真偽を判断し、論理記号を正しく使用する。" },
    {
      "unit_id": "ユニット 2",
      "name": "実数体系",
      "overview": "自然数、整数、有理数、無理数、実数の順序性と絶対値、数の性質（閉性、結合法則、分配法則）を学習する。",
      ...
    }
  ]
}

```

表 1: Japan/Japanese の「高校 1 年」「数学 I」の単元の記述を JSON 形式で自動生成させた例。

2 質問応答ペアの自動生成

2.1 生成の手順

1 節で述べた通り、広範な分野と多くの文化的な側面を網羅する高品質な学習データを獲得するために、既存のデータの収集やその翻訳に頼るのではなく、地域毎の教育課程に基づいたデータをスクラッチから生成させる。その際にパラメータ数が大きい既存の LLM を用いる。

図 1 に示したように階層的な生成を行う。最初に与える情報は地域 (loc) と言語 (lang)、そして教育のレベル¹⁾であり、以下のプロンプトを用いて学年のリストを出力させる²⁾。

Generate a list of all the grades in secondary education offered by {lang}-speaking schools in {loc}. Specifically, produce this list in JSON, using the following format:

これにより、地域/言語が United States/English であれば「6th grade」から「12th grade」が、Japan/Japanese なら「中学 1 年」から「高校 3 年」といったリストが JSON 形式で得られる。高等教育なら「Bachelor's Degree」「Master's Degree」などの学位を得る。

得られた学年ごとに、以下のプロンプトで科目 (course) のリストを得る³⁾。但し、高等教育については科目が多岐にわたるので、科目を生成する前に各学位に対応する discipline、program を順に生成し、それらを用いて科目のリストを得る。

Generate a list of courses taught in {grade} offered by {lang}-speaking schools in {loc}. Be as exhaustive as you can.

- 1) 中等教育 (Secondary education) と高等教育 (Tertiary education) を対象とする。初等教育は得られる知識の複雑性が低いのでデータの獲得は不要と考えた。
- 2) このプロンプトは中等教育の例。対象の言語によらずプロンプトは英語を用いた。
- 3) {grade} に学年の情報を代入する。代入される文字列だけは英語でなく各言語の表記となるが、生成する際に支障はなかった。以下、フォーマットに関する指示は省略する。

科目 (course) として、Japan/Japanese の「高校 1 年」であれば「数学 I」「生物基礎」など、Canada/French の「Secondaire II (中学 2 年)」なら「Géographie du Québec (ケベックの地理)」などが得られる。各科目に対して、単元 (unit) を生成する。

Generate a curriculum for the entire duration of a {course} course at the {grade} level, taught in {loc}, in {lang}.

これにより、表 1 の例のような、各科目につき概ね 6~10 程度の単元の学習の内容が生成される。

JSON で記述された科目のメタデータや単元の内容に続けて以下のプロンプトを与え、各単元に基づいた「homework (宿題)」を生成させる。

Generate a homework for {unit} using the following JSON format. The homework should include an appropriate number of tasks depending on the unit's scope. Feel free to use Markdown and line breaks for the task descriptions, as well as LaTeX for equations.

また、科目全体にわたった試験問題を生成する。

Generate an end-of-year exam for the course using the following JSON format. This is meant to be an hour-long, in-class written exam. The number of questions, their style, and difficulty should be appropriate for such an exam.

上記の 2 つのプロンプトによって、各科目の内容に沿った問題が生成されるが、その中には外部ツールを使わないといけないもの、画像や音声の出力を求めるもの、他の問題に依存するなど自己完結しないものなど、学習データに含める問題として不適格なものが含まれるため、それらを除外するためのフィルタリングを行う⁴⁾。

適格と判断された宿題や試験問題について、問題文そのものをプロンプトとして与えて解答を生成させ、質問応答のペアを得る。表 2 は、US/English の高等教育・Bachelor's Degree の科目「Calculus I (Differential Calculus)」の問題と解答の例である。

- 4) その検査の方法は補足資料 A.1 に記す。

Applying Limit Laws																						
問題	Compute the following limits using the appropriate limit laws. Show each step and indicate which law you are using. 1. $\lim_{x \rightarrow 3}(2x^2 - 5x + 1)$ 2. $\lim_{x \rightarrow 0} \frac{x^2 - 4x}{x}$ 3. $\lim_{x \rightarrow -1} \sqrt{x + 5}$ 4. $\lim_{x \rightarrow 2} \frac{x^3 - 8}{x - 2}$																					
解答	<p>Below each limit is evaluated step-by-step using the standard limit laws (constant-multiple, sum/difference, product, quotient, power, and the law for continuous functions). For each algebraic manipulation we state the law that justifies it.</p> <p>1. $\lim_{x \rightarrow 3}(2x^2 - 5x + 1)$</p> <table border="1"> <thead> <tr> <th>Step</th> <th>Expression</th> <th>Why this step is allowed</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>$\lim_{x \rightarrow 3}(2x^2 - 5x + 1)$</td> <td>-</td> </tr> <tr> <td>2</td> <td>$\lim_{x \rightarrow 3}(2x^2) - \lim_{x \rightarrow 3}(5x) + \lim_{x \rightarrow 3} 1$</td> <td>Sum/Difference Law: $\lim(f \pm g) = \lim f \pm \lim g$ (provided the individual limits exist).</td> </tr> <tr> <td>3</td> <td>$2 \lim_{x \rightarrow 3} x^2 - 5 \lim_{x \rightarrow 3} x + \lim_{x \rightarrow 3} 1$</td> <td>Constant-Multiple Law: $\lim c \cdot f = c \lim f$.</td> </tr> <tr> <td>4</td> <td>$2[\lim_{x \rightarrow 3} x]^2 - 5[\lim_{x \rightarrow 3} x] + 1$</td> <td>Power Law: $\lim(f)^n = (\lim f)^n$. The limit of a constant is the constant itself.</td> </tr> <tr> <td>5</td> <td>$2(3)^2 - 5(3) + 1$</td> <td>Evaluate the simple limits: $\lim_{x \rightarrow 3} x = 3$ and $\lim_{x \rightarrow 3} 1 = 1$.</td> </tr> <tr> <td>6</td> <td>$2 \cdot 9 - 15 + 1 = 18 - 15 + 1 = 4$</td> <td>Arithmetic.</td> </tr> </tbody> </table> <div style="text-align: center; border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> $\lim_{x \rightarrow 3}(2x^2 - 5x + 1) = 4$ </div> <p>(以下略)</p>	Step	Expression	Why this step is allowed	1	$\lim_{x \rightarrow 3}(2x^2 - 5x + 1)$	-	2	$\lim_{x \rightarrow 3}(2x^2) - \lim_{x \rightarrow 3}(5x) + \lim_{x \rightarrow 3} 1$	Sum/Difference Law: $\lim(f \pm g) = \lim f \pm \lim g$ (provided the individual limits exist).	3	$2 \lim_{x \rightarrow 3} x^2 - 5 \lim_{x \rightarrow 3} x + \lim_{x \rightarrow 3} 1$	Constant-Multiple Law: $\lim c \cdot f = c \lim f$.	4	$2[\lim_{x \rightarrow 3} x]^2 - 5[\lim_{x \rightarrow 3} x] + 1$	Power Law: $\lim(f)^n = (\lim f)^n$. The limit of a constant is the constant itself.	5	$2(3)^2 - 5(3) + 1$	Evaluate the simple limits: $\lim_{x \rightarrow 3} x = 3$ and $\lim_{x \rightarrow 3} 1 = 1$.	6	$2 \cdot 9 - 15 + 1 = 18 - 15 + 1 = 4$	Arithmetic.
Step	Expression	Why this step is allowed																				
1	$\lim_{x \rightarrow 3}(2x^2 - 5x + 1)$	-																				
2	$\lim_{x \rightarrow 3}(2x^2) - \lim_{x \rightarrow 3}(5x) + \lim_{x \rightarrow 3} 1$	Sum/Difference Law: $\lim(f \pm g) = \lim f \pm \lim g$ (provided the individual limits exist).																				
3	$2 \lim_{x \rightarrow 3} x^2 - 5 \lim_{x \rightarrow 3} x + \lim_{x \rightarrow 3} 1$	Constant-Multiple Law: $\lim c \cdot f = c \lim f$.																				
4	$2[\lim_{x \rightarrow 3} x]^2 - 5[\lim_{x \rightarrow 3} x] + 1$	Power Law: $\lim(f)^n = (\lim f)^n$. The limit of a constant is the constant itself.																				
5	$2(3)^2 - 5(3) + 1$	Evaluate the simple limits: $\lim_{x \rightarrow 3} x = 3$ and $\lim_{x \rightarrow 3} 1 = 1$.																				
6	$2 \cdot 9 - 15 + 1 = 18 - 15 + 1 = 4$	Arithmetic.																				

表 2: 生成された問題（しばしば複数の問題が含まれる）と解答（紙面の都合により一問目の解答のみ）の例。LaTeX・マークダウンを含む JSON 形式の元データを、MathJax でレンタリングしたもの。

地域	言語	宿題	試験	合計
Germany	German	482,243	277,212	759,455
United States	English	681,546	541,844	1,223,390
Spain	Spanish	465,757	359,057	824,814
France	French	350,546	212,828	563,374
Japan	Japanese	329,852	362,325	692,177
Russia	Russian	577,281	298,380	875,661
China	Chinese	334,877	358,217	693,094
7 地域合計 (うち理工系 (STEM)) (うち数学 (Math))		3,222,102	2,409,863	5,631,965 2,362,703 180,319

表 3: 獲得できた質問応答ペアの数。地域・言語はプロンプトに与えた表記のまま。

2.2 質問応答ペアの生成

2.1 節の手順に従って、gpt-oss-120b[18] を用いて言語が異なる 7 地域のデータを生成した。表 3 はそれぞれの中等教育・高等教育の課程由来で獲得できた質問応答のペアの数を示す。US/English は生成された科目の数が大きくデータ量が最大であるが、他の地域向けにも遜色ないデータ量が得られている。各単元に関連した「宿題」は生成された質問が多いものの、フィルタリングを通過する割合は僅か 2~3 割で、特に人文・芸術系の分野で低かった。「試験問題」は生成される数は少ないが 9 割程度が適格と

みなされ、最終的に両者は同程度の量となり、2 種類の生成の方法が相互補完的といえる。

また、3 節の実験のために、全データのうち、理工系 (STEM) の分野と、数学 (Math) に限定した部分集合を作成した。これは、各科目の名称を与え、それらが STEM, Math に属するか否かを LLM に問うて判断したものである。表 3 にそれらのデータ数を合わせて示してある。

3 実験

2 節の方法で獲得したデータを、LLM の指示学習に適用し、数学やその他の問題で評価した。

3.1 実験設定

以下の 2 つの設定で実験を行った。

指示学習：事前学習のみが完了させてあるベースの言語モデル google/gemma-3-4b-pt [19] および ibm-granite/granite-4.0-micro-base [20] に対して、質問応答ペアを用いて指示学習を行う。

継続指示学習：上記の 2 モデルに対応する指示学習済のモデル google/gemma-3-4b-it [19] および ibm-granite/granite-4.0-micro⁵[20] に対して、継続指示学習を行う。

5) 上述の -base モデルとともに、パラメータ数は 3 billion。

モデル	学習	de	en	es	fr	ja	ru	zh	平均
gemma	Aya	14.8	27.2	28.8	16.4	14.0	20.0	15.2	19.5
	全分野	36.8	41.2	36.0	30.8	25.6	38.8	30.0	34.2
	理工系	37.6	44.8	41.6	36.8	30.8	38.8	37.6	38.3
	数学	42.0	46.8	44.4	39.6	31.6	41.2	35.6	40.2
	Aya+数学	31.2	45.6	41.2	34.4	22.0	37.2	26.8	34.1
granite	学習無	16.0	24.8	13.6	14.4	11.2	13.2	14.8	15.4
	Aya	39.6	43.6	30.8	38.8	32.0	36.4	25.6	35.3
	全分野	62.8	75.6	67.2	34.8	52.4	65.2	63.6	60.2
	理工系	62.0	75.6	66.8	42.0	52.0	62.8	63.2	60.6
	数学	64.8	83.6	68.8	42.4	54.8	70.8	62.0	63.9
Aya+数学	59.2	81.2	62.8	60.4	55.6	67.2	60.8	63.9	

表 4: gemma と granite のベースモデルに対して指示学習をした場合の MGSM の正解率。提案手法で生成したデータを太字で示す。gemma のベースモデルは指示に追従しないため「学習無」は割愛した。

モデル	学習	de	en	es	fr	ja	ru	zh	平均
gemma	学習無	16.8	66.0	34.4	40.0	14.8	24.8	23.2	31.4
	Aya	25.2	66.8	42.8	49.2	27.6	36.0	35.2	40.4
	全分野	31.2	72.4	52.4	57.2	21.6	42.8	36.4	44.9
	理工系	32.8	73.2	51.6	59.2	24.4	44.4	39.6	46.5
	数学	35.2	78.8	53.6	63.2	26.0	47.6	43.6	49.7
Aya+数学	48.8	71.2	56.4	58.8	38.0	52.0	48.4	53.4	
granite	学習無	38.4	31.6	37.6	30.4	24.4	34.0	27.2	31.9
	Aya	40.4	52.8	32.4	37.2	28.0	30.4	28.4	35.7
	全分野	37.6	32.4	34.8	29.6	24.8	34.0	27.6	31.5
	理工系	38.8	35.6	32.4	27.2	32.0	35.2	38.4	34.2
	数学	46.0	40.4	34.8	30.4	38.0	40.4	38.4	38.3
Aya+数学	36.0	30.8	38.0	31.2	26.4	34.0	28.4	32.1	

表 5: 指示学習済のモデルに対して継続指示学習をした場合の MGSM の正解率。

評価データとして以下の 2 つを用いた。いずれもモデルの出力が解答と合致するか否かを gpt-oss-120b に判定させて正解率を測る。

MGSM: 数学の問題の多言語のデータセット [21] で、常に単一の数が解答となる。今回データを生成した 7 言語を用いて評価する。

MMMLU: 多くの分野の推論能力を測る 4 択問題のデータセット MMLU の多言語版 [22]。今回の対象と共通する 6 言語を用いて評価する。

提案手法により生成したデータの比較対象として、多言語の指示学習のデータ Aya Dataset [13] を用いた。これは 70 言語の指示と応答のペアを web 等から収集したもので、202,352 件のデータを含む。

3.2 実験結果

表 4 に、ベースモデルに対して指示学習⁶⁾を施して MGSM の各言語 250 件で評価した結果を示す。

6) 設定したパラメータは補足資料 A.2 に記載した。

モデル	学習	de	en	es	fr	ja	zh	平均
gemma	Aya	47.7	55.6	50.9	48.4	42.9	48.6	49.0
	全分野	49.1	55.9	51.3	48.8	43.9	46.6	49.3
	数学	46.3	48.3	49.8	48.0	43.1	46.0	46.9
granite	Aya	51.1	59.2	54.0	53.4	47.4	49.6	52.4
	全分野	52.7	55.6	54.7	53.9	47.0	48.3	52.0
	数学	49.1	45.7	52.1	49.4	44.4	46.9	47.9

表 6: gemma, granite のベースモデルに対する指示学習の MMMLU での評価。

google/gemma-3-4b-pt の場合、全言語において、提案手法によるデータ (全分野・理工系・数学) での指示学習が、Aya よりも高い性能を示した。ほとんどの言語において、全分野よりも数学に限定したデータによる指示学習の性能が高かった。

granite では指示学習により大幅な性能の向上が見られている。gemma の場合と同様に数学に限定した場合の性能が高く、またフランス語では Aya データと併用した場合に特に高い正解率が得られた。

表 5 は指示学習済のモデルに対して、継続指示学習をした場合の結果である。google/gemma-3-4b-it においてはベースモデルからの学習よりも高い性能が得られており、もともとの指示学習を補強できていることがわかる。また、Aya と数学のデータを結合するとさらに良い結果が得られた。一方、ibm-granite/granite-4.0-micro では正解率上昇の幅が小さく、表 4 の結果を超えなかった。元の指示学習の特性や整合性など、原因を調査中である。

表 6 は MMMLU の 6 言語で評価した結果である。両モデルで、提案手法により自動生成されたデータは Aya データと概ね同等の効果が得られている。数学以外の多様な問題が含まれるため、提案手法によるデータの効用が期待されたが、MGSM のような顕著な性能の向上は見られていない。回答の形式の違いの影響なども考えられるが、分野別の比較などの詳細な分析は今後の課題とする。

4 まとめ

指示学習向けの多言語のデータを効率よく獲得する方法を提案し、ベンチマーク上での性能を確認した。多言語で均質、かつ広範な分野を扱うデータセットが得られたので、LLM の推論能力の向上に必要な言語をまたがる知識の検討、そして言語のみならず、米国と英国の英語などに見られる文化の違いなどの考察に活用していきたい。

参考文献

- [1] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: A metric for quantifying LLM performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, pp. 28186–28194, 2025.
- [2] Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. Multilingual Performance Biases of Large Language Models in Education. *arXiv e-prints*, p. arXiv:2504.17720, April 2025.
- [3] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, Vol. 3, No. 9, p. pgae346, 2024.
- [4] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. CultureLLM: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, Vol. 37, pp. 84799–84838, 2024.
- [5] Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text, 2023.
- [6] Jiazhen Li, Lu Yu, Qing Cui, Zhiqiang Zhang, Jun Zhou, Yanfang Ye, and Chuxu Zhang. Mass: Mathematical data selection via skill graphs for pretraining large language models, 2025.
- [7] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023.
- [8] Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Extrapolating Large Language Models to Non-English by Aligning Languages. *arXiv e-prints*, p. arXiv:2308.04948, August 2023.
- [9] Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-X: Multilingual Replicable Instruction-Following Models with Low-Rank Adaptation. *arXiv e-prints*, p. arXiv:2305.15011, May 2023.
- [10] Haoyu Wang, Shuo Wang, Yukun Yan, Xujia Wang, Zhiyu Yang, Yuzhuang Xu, Zhenghao Liu, Liner Yang, Ning Ding, Xu Han, Zhiyuan Liu, and Maosong Sun. UltraLink: An Open-Source Knowledge-Enhanced Multilingual Supervised Fine-tuning Dataset. *arXiv e-prints*, p. arXiv:2402.04588, February 2024.
- [11] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *EMNLP*, 2022.
- [12] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. OpenAssistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, Vol. 36, pp. 47669–47681, 2023.
- [13] Shivalika Singh, Freddie Vargus, DanielD’ souza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividatas Mataciunas, LauraO’ Mahony ほか. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, 2024.
- [14] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: Im ChatGPT interaction logs in the wild. *CoRR*, 2024.
- [15] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Kingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. PolyLM: An Open Source Polyglot Large Language Model. *arXiv e-prints*, p. arXiv:2307.06018, July 2023.
- [16] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.
- [17] Rishabh Maheshwary, Vikas Yadav, Hoang H Nguyen, Khyati Mahajan, and Sathwik Tejaswi Madhusudhan. M2Lingual: Enhancing multilingual, multi-turn instruction alignment in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9676–9713, 2025.
- [18] OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Korylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025.
- [19] Gemma Team. Gemma 3. 2025.
- [20] IBM Research. Granite 4.0 language models. <https://github.com/ibm-granite/granite-4.0-language-models>, 2025. Accessed: 2025-10-01.
- [21] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. *arXiv e-prints*, p. arXiv:2210.03057, 2022.
- [22] Yi Zhou, Yusuke Sakai, Yongxin Zhou, Haonan Li, Jiahui Geng, Qing Li, Wenxi Li, Yuanyu Lin, Andy Way, Zhuang Li, Zhongwei Wan, Di Wu, Wen Lai, and Bo Zeng. Multilingual MMLU benchmark leaderboard. <https://huggingface.co/spaces/StarscreamDeceptions/Multilingual-MMLU-Benchmark-Leaderboard>, 2024.
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, p. arXiv:1412.6980, December 2014.

A 補足資料

A.1 質問のフィルタリング

問題文を生成した後のフィルタリングに使うプロンプト。出力された JSON 形式を見て、1 が True の場合、2, 3 が空リストでない場合、4 が False のものは回答可能な自己完結した問題でないため、データから除外した。

You are an LLM tasked with classifying the properties of each individual task in an assignment. For every task, produce exactly one JSON object that captures the properties described below.

- 1 'required_preceding_tasks' (array of strings)
- 2 'modalities' (array of strings)
- 3 'tools' (array of strings): An array of strings that lists every external service or tool
- 4 'input_completeness' (boolean): True if the task description contains all content needed to solve the task or if it is a follow-up task that can obtain missing content from a prior task. False otherwise.

A.2 指示学習のパラメータ

3 節の実験では、Adam optimizer [23] で、バッチサイズ 256、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ 、 $\epsilon = 1e-8$ で instruction tuning を行った。学習率は 0 から以下の表の値まで漸増させた。

モデル	学習	MGSM	MMMLU
gemma-3-4b-pt	Aya	2e-6	5e-6
	全分野	4e-6	1e-5
	理工系	8e-6	–
	数学	5e-6	6e-7
gemma-3-4b-it	Aya+数学	1e-6	–
	Aya	1e-6	–
	全分野	2e-6	–
	理工系	1e-6	–
granite-4.0-micro-base	数学	1e-6	–
	Aya+数学	1e-6	–
	Aya	4e-5	4e-6
	全分野	2e-6	4e-6
granite-4.0-micro	理工系	2e-6	–
	数学	2e-6	4e-6
	Aya+数学	2e-6	–
	Aya	4e-5	–
granite-4.0-micro	全分野	1e-6	–
	理工系	1e-6	–
	数学	1e-6	–
	Aya+数学	2e-6	–

A.3 その他の例

France/French の「Histoire – Le Moyen Âge et les débuts de la modernité (中世と近代の歴史)」の宿題と解答の例。特に人文系においては作文など正解が一意に定まらないものが多いが、それぞれの分野における有用な知識が得られる可能性がある。

問題	Rôle-play écrit
	Imagine que tu es un marchand français en 1540. Rédige une lettre (150-200 mots) à ton confrère de Bordeaux décrivant: - L'impact de l'arrivée du sucre et du tabac sur tes ventes. - Les difficultés rencontrées à cause de l'inflation du blé. - Une proposition de coopération pour profiter des nouvelles routes commerciales. Utilise un ton réaliste et n'hésite pas à insérer des termes du vocabulaire étudié (ex.: mercantilisme, monnaie, inflation).
解答	Mon cher ami, J'espère que la santé de votre maison se porte bien en ces temps de grande agitation commerciale. Depuis l'entrée du sucre de Canarie et du tabac d'Amérique dans nos entrepôts, nos ventes ont connu un essor inattendu: les nobles de la cour se livrent désormais à des banquets où le sucre remplace le miel, et le tabac se vend au comptoir des tavernes comme le nouveau luxe. Cette demande croissante double nos recettes, mais elle impose aussi un besoin pressant de monnaie d'échange, car les paiements en argent frappé tardent à suivre le volume des échanges. Hélas, l'inflation du blé, provoquée par les mauvaises récoltes du Nord, pèse lourdement sur nos marges. Les prix du grain s'envolent, rendant difficile la subsistance de nos ouvriers et la fixation d'un prix stable pour nos marchandises. Je vous propose de nouer une coopération: combinons nos cargaisons de vin et de draps avec mes cargaisons de sucre et de tabac, utilisant les nouvelles routes vers l'Atlantique et la Méditerranée. Ainsi, sous le feu du mercantilisme royal, nous pourrions mutualiser les frais de transport, stabiliser notre monnaie et profiter des profits grandissants. Dans l'attente de votre réponse, recevez, cher confrère, mes salutations distinguées.

Japan/Japanese の「物理基礎」の試験問題と解答の例。本件を含めて各言語の数例については解答の正しさを確認した。全数の検査はしていないが、3 節の実験を通して指示学習の効果が見られたことから、一定の品質は保たれていることが見込まれる。

問題	温度と熱量
	次の条件で加熱された水の熱量 Q を求めよ。 - 質量 $m = 250\text{g}$ 、初温度 20°C 、最終温度 80°C 、水の比熱 $c = 4.186\text{J}\cdot\text{g}^{-1}\cdot\text{K}^{-1}$ 式 $Q = mc\Delta T$ を使い、単位は kJ で示すこと。
解答	温度変化量 ΔT $\Delta T = T_f - T_i = 80^\circ\text{C} - 20^\circ\text{C} = 60\text{K}$ 熱量 Q の計算 $Q = mc\Delta T$ $Q = (250\text{g})(4.186\text{J}\cdot\text{g}^{-1}\cdot\text{K}^{-1})(60\text{K}) = 250 \times 4.186 \times 60\text{J} = 250 \times 251.16\text{J} = 62790\text{J}$ kJ に変換 $62790\text{J} = \frac{62790}{1000}\text{kJ} = 62.79\text{kJ}$ 結果 $Q \approx 6.28 \times 10^1\text{kJ} (= 62.8\text{kJ})$ この熱量が、250g の水を 20°C から 80°C に加熱するのに必要な熱エネルギーです。