新聞記事からつくる 時事と社会に強い日本語 LLM

服部 翔 [1,2]　水木 栄 [1,2]　藤井 一喜 [1,2]　中村 泰士 [1,2]　塩谷 泰平 [1]　植木 快 [3]
新妻 巧朗 [3]　川畑 輝 [3]　田森 秀明 [3]　Youmi Ma[1]　前田 航希 [1]　大井 聖也 [1,2]
齋藤 幸史郎 [1]　岡本 拓己 [1]　石田 茂樹 [1]　横田 理央 [1,2]　高村 大也 [2]　岡崎 直観 [1,2]
[1] 東京科学大学　[2] 産業技術総合研究所　[3] 株式会社朝日新聞社
{kakeru.hattori@nlp., okazaki@, swallow@nlp.}comp.isct.ac.jp,
{ueki-k1, niitsuma-t, kawabata-a, tamori-h}@asahi.com

概要

大規模言語モデル（LLM）の事前学習において新聞記事はどのような恩恵をもたらすのか？本研究では，LLM の日本語継続事前学習における新聞記事データの有用性，およびその効果を引き出すための手法について報告する．はじめに，新聞記事のみを用いて LLM の継続事前学習を行ったが，テキスト量と多様性の不足のためか，十分な効果を得ることができなかった．そこで，ドメイン適応の既存研究を参考に，新聞記事をシードとして LLM で合成データを生成し，継続事前学習のデータに追加した．実験の結果，合成データを併用することにより前述の問題を解消し，新聞記事に関連する分野を中心に，LLM の日本語能力が向上した．

1　はじめに

大規模言語モデル（LLM）の事前学習において，高品質なデータを用いることの重要性が知られている [1, 2]．日本語は英語などに比べてデータ量が少ないため，ウェブページ以外の有用な言語資源を開拓することも重要となる．新聞記事は，日本語テキストとしての品質が担保されている，日本の社会・文化・時事に関する情報が豊富に含まれているため，LLM の事前学習に有益と考えられているが，その効果は未だ実証されていない．

そこで，本研究ではまず，**朝日新聞社が所有する直近 41 年間の新聞記事テキスト**を用いて日本語 LLM に継続事前学習を行い，その効果を調査した．我々の知る限り，LLM の事前学習における日本語の新聞記事の効果を検証した研究はこれが初めてである．ところが，生の記事のみを用いた場合は期待通りの効果が得られず，日英ほぼ全てのタスクで性能が下落した．学習量を増やすために記事を複数エポック学習すると，この傾向はより顕著になり，特に英語の生成を伴うタスクで性能が悪化した．

この現象は，生の記事のみではテキストの量や多様性が不足していたことが原因であると推測した．また，LLM が特定の知識を獲得するにはその事実に多数回触れる必要がある [3] と言われており，ウェブテキストで補えない知識を少量の新聞記事から獲得する取り組みにおいて，悩ましい課題である．ドメイン適応の分野における既存研究では，読解問題 [4, 5] や QA [6]，エンティティに焦点を当てたテキスト [7] などの合成データを LLM で生成して学習に加え，より多様な知識表現を与えることで，ドメイン知識に関するタスクの性能を改善できたことが報告されている．良質な合成データを構築するにあたっては，生成の情報元となるシード（種）テキストの品質担保が重要 [8] とされているが，元々内容の正確性や有用性が確保されている新聞記事であれば，シードとしての適性が高いと予想できる．

これらを踏まえ，本研究ではさらに，新聞記事を情報元として LLM で**合成データ**を構築し，学習データに加えることで，データの量や多様性に起因する問題の解消と，新聞記事に含まれる有用な知識の獲得を促し，モデルの日本語性能の向上を目指した．実験の結果，合成データを併用した学習では前述の問題を解消し，既に大規模・高品質な日本語コーパスで学習済みであったベースモデル [9] の日本語の知識や推論能力をさらに改善できた．より具体的には，時事的知識，社会科学，人文科学など新聞記事に関連する分野で改善幅が大きいことを確認し，新聞記事と合成データの効果を実証した．

LLM に新聞記事のような高品質なデータを与えても，知識をうまく定着させることができなかったが，合成データで言い換えながら繰り返し教えることで学習効果が高まったという知見は，興味深い．

2 データセットの構築

2.1 新聞記事データ

まず，朝日新聞社が所有する 1984 年から 2024 年までの記事データ（8,141,148 件・4,631,978,404 文字）に対して，必要最小限の前処理とフィルタリングを行い，最終的に 5,918,638 件・4,175,763,559 文字のデータを得た．処理の概要は以下の通りである．

- 全ての記事の全角スペースを空文字列に置換
- 本文の日本語文字数が 200 字未満の記事を除外
- 文の平均文字数が 10 文字未満の記事を除外
- 最頻出の {2, 3, 4}-gram の出現率が {0.20, 0.18, 0.16} より高い記事を除外
- ひらがな文字の割合が 0.1 未満の記事を除外

2.2 合成データの生成

続いて，新聞記事に含まれる知識を多様な表現に変換し，LLM への定着を促すことを主眼に置き，LLM を用いた合成データの構築を行った．具体的には，Cosmopedia [10] などを参考にしてプロンプト（付録 A）を作成し，**QA 形式**および**教科書形式**のデータを生成した．合成データを生成する LLM には Gemma 2 27B IT[1)]，計算機には TSUBAME 4.0 を使用し，vLLM[2)] によるバッチ推論を行った．構築した全てのデータ（表 1）の合計規模は，元の新聞記事のみの 5.74 倍（3.28BT → 18.82BT）である．

QA 形式の合成データ　既存研究 [4, 5, 6] の実績を踏まえ，LLM の推論や知識引き出しの性能向上を目指し，テキストに含まれる知識事項を，実際の応用タスクに近い問題と解答の形式に変換した．QA は短答式，論述式，多肢選択式の 3 形式で生成し，問題ごとの解説文も付与するように指示することで，全体として元の知識を再構成し，LLM が多様な形式で知識を学習できるように工夫した．なお本形式は，新聞記事に加え，後述する教科書形式の合成データをシードとする生成も行った．

教科書形式の合成データ　新聞記事は有用な知識や情報を含む一方，特定の人物や日時に発生した事件など，限定的・一時的な事象の記述も多く，これら全てを丸暗記することが必ずしも LLM の改善に有益とは限らないと考えた．そこで，記事が提供する教育的な内容や，取り扱う事象の周辺知識に焦点を置き，教科書のように一般的な教養を順序立てて説明する文書を生成した．元の記事では簡単に触れられるだけの用語や概念についても，教育的価値が高い内容であれば，LLM（Gemma 2）が独自に情報を補完して生成することを許容した．この手法は単なる形式変換に留まらず，強力な LLM（Gemma 2）からの知識転移を活用する意味合いが強い．新聞記事は，LLM から有用な知識を引き出すトリガーとして機能することが期待されるが，記事に明確に記載されていない内容では，誤った情報を生成するリスクも増加する．今回は生成されたテキストをそのまま採用したが，生成後の品質確認や修正処理の導入については，今後の課題としたい．

1） https://huggingface.co/google/gemma-2-27b-it
2） https://docs.vllm.ai/en/stable/

表 1　構築したデータセットの一覧

データセット名	トークン数
朝日新聞記事	3.28BT
QA 形式データ（記事由来）	4.95BT
教科書形式データ	5.45BT
QA 形式データ（教科書形式データ由来）	5.14BT

2.3 構築した合成データの事例分析

2.2 節の合成データが，生の新聞記事の情報をどのように変換したか，「ユニバーサルデザインタクシー」（2016 年 1 月 5 日付夕刊）の記事を事例として分析した．各テキストの抜粋は付録 B に示す．

記事を元に生成した QA 形式データは「東京都が新規購入する事業者に 1 台あたり 60 万円を補助する事業」など，記事の具体的な記述内容を忠実に問題形式へと変換している．一方，教科書形式データは，「ユニバーサルデザイン（UD）」の目的や「ユニバーサルデザインタクシー」の特徴など，より一般化された教養の提供に重点を置いた構成となっている．また，記事内で簡単に触れられていた用語や概念について，LLM（Gemma 2）が独自に詳細な解説を補完して生成した箇所も多く見られる．

3 実験

新聞記事データや合成データの有効性を検証するため，LLM の継続事前学習を行った．ベースモデルには Llama 3.1 Swallow 8B v0.1[3)] を使用した．同モデルは Meta 社が公開する Llama 3.1 8B[4)] に対して教育的な日本語コーパス [9] で継続事前学習を行うこと

3） https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-v0.1
4） https://huggingface.co/meta-llama/Llama-3.1-8B

表 2　継続事前学習前後の日英ベンチマークスコア比較

実験設定	QA JCom.	QA JEMHopQA	QA NIILC	教養科目 JMMLU	英日翻訳 WMT20	日英翻訳 WMT20	要約 XL-Sum	機械読解 JSQuAD	数学 MGSM	コード生成 JHumanEval	swallow-evalation 日本語	swallow-evalation 英語	時事 QA ニュース Q	教養科目 pfgen-bench
Llama 3.1 8B	0.844	0.446	0.405	0.477	0.221	0.208	0.179	0.896	0.356	**0.327**	0.436	**0.564**	0.460	0.409
Llama 3.1 Swallow 8B	0.912	0.509	0.601	0.518	**0.291**	**0.231**	0.202	0.899	**0.460**	0.281	**0.491**	0.558	0.633	0.671
生の新聞記事のみを用いた学習														
記事（1 エポック）	0.908	0.479	0.550	0.501	0.270	0.163	**0.230**	0.896	0.388	0.251	0.464	0.526	0.621	0.609
+再学習（10%）	**0.913**	0.491	0.566	0.502	0.277	0.222	0.217	0.900	0.396	0.265	0.475	0.542	0.632	0.597
記事（20BT）	0.863	0.489	0.552	0.469	0.256	0.057	0.213	0.863	0.360	0.095	0.422	0.379	0.613	0.573
合成データを併用した学習														
記事+**QA**	0.908	0.519	0.587	**0.541**	0.267	0.204	0.223	0.911	0.424	0.261	0.484	0.519	**0.663**	**0.687**
記事+**教科書**	0.904	0.532	0.584	0.487	0.271	0.207	0.224	0.902	0.428	0.282	0.482	0.520	0.649	0.671
記事+**合成データ全て**	**0.913**	**0.547**	**0.612**	0.537	0.270	0.212	0.217	**0.913**	0.420	0.265	**0.491**	0.532	0.655	0.671

で日本語能力を強化したものであり，本実験ではこのモデルに対してさらなる継続事前学習を行った．

評価には llm-jp-eval [11] などの複数のツールを改変・統合した swallow-evaluation[5]（Ver. 202407）を用い，日英の幅広いタスクで LLM の性能を網羅的に評価した．評価タスクは Swallow の開発で共通に用いられている[6]10 件の日本語理解・生成タスク [12, 13, 14, 15, 16, 17, 18, 19] と 9 件の英語理解・生成タスク [20, 21, 22, 23, 24, 25, 26, 27, 28] である．これらに加えて，朝日新聞社が作成したニュース Q [29] および，今城らが作成した日本語生成ベンチマークである pfgen-bench [30] を追加で用いた．

ニュース Q は 2022〜2023 年度の朝日新聞記事を元に作成された 3〜4 択の QA 集であり，多くの問題は記事の内容から解答を導くことができる．本研究では学習を通して，モデルが記事の知識を獲得できたかどうかを確認する指標として採用した．

3.1　生の新聞記事のみを用いた学習

まず，生の新聞記事のみを用いて学習を行った．学習トークン数は 3.28BT（1 エポック），20BT（6.1 エポック）の 2 パターンで実験を行った．

表 2 中部に各タスクの評価結果を示した（日本語タスクに関してはタスク毎のスコアも示した）．生の記事のみを用いて学習したモデルでは，要約（XL-Sum）を除いた日英全てのタスクでスコアが低下した．記事を 20BT まで繰り返し学習した場合はこの傾向がより顕著になり，特に日英翻訳（WMT20），コード生成（JHumanEval）といった英語の生成を伴うタスクで性能が悪化した．

これらの現象は，既存研究 [4, 7] の報告を踏まえると，生の新聞記事のみではテキストの量や多様性が不足していたことが原因であると推測される．例えば，生の新聞記事における英小文字，英大文字の割合がそれぞれ 0.19%，0.40%と低いことは，英文字の生成頻度を過度に低下させた可能性がある．

また上記に加え，Yang ら [7] の実験設定を参考に，ベースモデルの学習に用いた日本語コーパス [9] を全体の 10%の割合で再学習として追加し，新聞記事 1 エポックと合わせて 3.64BT を学習する設定も試行した．この結果，日英翻訳（WMT20）などを中心にスコアの下落幅が軽減したものの，全体的なスコアの低下傾向を解消することはできなかった．

3.2　合成データを併用した学習

3.1 節の実験で生じた問題を解消するため，2.2 節で構築した合成データを併用した学習を行った．QA 形式（記事由来），教科書形式の合成データのみをそれぞれ追加した場合と，表 1 の全ての合成データを追加した場合の 3 パターンで実験を行った．いずれも 20BT まで学習を行った．

評価結果を表 2 下部に示した．全体的に，合成データを併用して学習したモデルは，ベースモデルの日本語タスク平均スコアを維持しつつ，新聞記事と関連度の高い複数のタスクでスコアが向上した．

まず，**ニュース Q** ではベースモデルと比較して 1.6 から 3.0 pt のスコア改善を達成した．同ベンチマークでの正答率向上は，モデルが新聞記事に含まれる知識を追加で獲得したことを反映しており，学習で期待した成果が実際に表れている．また，その他の一般的な日本語ベンチマークにおいても，**教養科目**（JMMLU，pfgen-bench），**QA**（JEMHopQA），**機械読解**（JSQuAD），**要約**（XLSum）といった日本語の知識や推論に関する幅広いタスクで，スコアを維持もしくは向上させることができた．これらの結果は，新聞記事を合成データのシードとして活用する手法が，同テキストの学習効果を引き出し，日本語 LLM の性能向上に寄与したことを示唆している．

一方，英語全般，翻訳（WMT20），数学（MGSM），コード生成（JHumanEval）では依然としてスコアが

5）https://github.com/swallow-llm/swallow-evaluation
6）https://swallow-llm.github.io/evaluation/about.ja

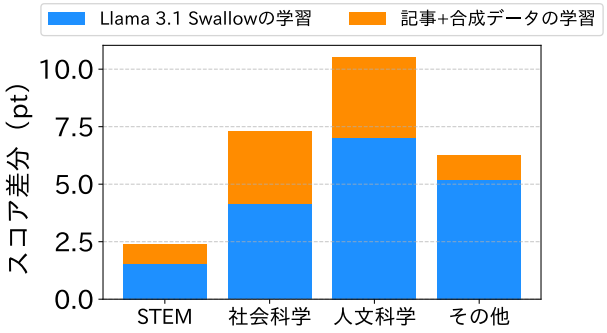

図 1　JMMLU のカテゴリ別平均正答率変化

低下した．これは，英語コーパスを学習に一切含めなかったこと，新聞記事には数学やコード生成に特化した情報が少ないことを考慮すると，自然な結果である．合成データの追加のみで，3.1 節で述べた英語の生成を伴うタスクのスコア低下を大幅に軽減できた理由としては，合成データの英文字割合が生の新聞記事と比べて高かったことが考えられる．

合成データの種類別効果比較　記事由来の QA 形式データのみ，教科書形式データのみ，を追加した場合を比較すると，日本語タスクの平均スコアは概ね同水準であるが，本実験で改善が見られたタスク群に着目すると，JEMHopQA を除いて QA 形式データの方が改善が大きい傾向が見られ，特に教養科目（JMMLU）では差が明確である．また QA 形式データは，知識量よりも日本語の基礎能力が重要な機械読解（JSQuAD）[31] でも改善が見られるなど，多様なタスクに対して効果を示した．QA は既存の学習データと大きく異なる新鮮な形式であるため，応用タスクへの効果がより大きかった可能性がある．また，シード（新聞記事）の情報に忠実な形式変換のみを行った同データが，単体でも高い効果を示したことは，良質なシードの選定と，多様な知識表現に変換可能な優れた指示文の調整が，効果的な合成データの構築に重要であることを示唆している．

3.3　JMMLU のカテゴリ別正答率変化

新聞記事がどのような分野の知識獲得に寄与したか分析するため，Llama 3.1 8B を起点として，Llama 3.1 Swallow 8B の学習 [9]，および本研究での新聞記事+合成データ全てを用いた学習前後で，JMMLU [19] の各カテゴリ別平均スコアがどのように変化したかを調査した．カテゴリは JMMLU の翻訳元である MMLU [25] での分類に準拠した．

結果を図 1 に示した．いずれの学習でも全カテゴリでスコアが向上したが，記事+合成データによる学習では社会科学（+3.2 pt）や人文科学（+3.5 pt）により特化してスコアが向上した傾向が見られる．これらの結果は，新聞記事が同分野に関連する内容を多く含むことを反映していると考えられる．また，全体的には Llama 3.1 Swallow 8B の学習と比較してスコア上昇幅が小さいが，学習トークン数の圧倒的な差（230BT vs 20BT）を考慮すれば，記事+合成データの学習効果は十分に大きいと言える．

表 3　ニュース Q の時事問題における正答率変化

実験設定	時事問題		非時事問題	
Llama 3.1 Swallow 8B	0.581		**0.791**	
記事**+QA**	0.623	+4.2 pt	0.785	-0.6 pt
記事**+教科書**	0.611	+3.0 pt	0.767	-2.4 pt
記事**+合成データ全て**	**0.627**	+4.6 pt	0.743	-4.8 pt
Gemma 2 9B IT	*0.583*		*0.682*	
Gemma 2 27B IT	*0.630*		*0.727*	

3.4　ニュース Q の時事問題への効果

新聞記事が時事的な知識の定着に寄与したかどうかを分析するため，ニュース Q [29] の「時間依存あり」ラベルが付与された問題群を時事問題とみなし，正答率の変化を調査した．

結果を表 3 に示した．記事と合成データを用いて学習したモデルは，主に時事問題で正答率が向上していることがわかった．合成データの生成に用いた LLM（Gemma 2 27B IT）自身も時事問題を高い精度で解答できているため，合成データの効能が情報元の新聞記事と Gemma 2 からの知識転移のどちらに由来するのか，正確には断定できない．ただし，Gemma 2 からの知識転移が少ない QA 形式データのみを用いた場合でも，十分なスコア向上が達成されていることを踏まえると，少なくとも，新聞記事が時事的な知識の獲得に有用な合成データのシードである，ということは確かだろう．

4　おわりに

本研究では LLM の事前学習における新聞記事データの有用性について検証した．実験では，記事をシードとする合成データを生成し，学習に追加することで，LLM の日本語知識や推論に関する能力を有効に改善できることを示した．今後，新聞をはじめとして，様々なドメインの高品質なデータから特化型 LLM を構築・活用していくにあたり，本研究が参考となる知見を提供すると期待している．

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP18002）の結果得られたものです．また，LLM の継続事前学習の実験では，国立研究開発法人産業技術総合研究所が構築・運用する AI 橋渡しクラウド（ABCI: AI Bridging Cloud Infrastructure）の「大規模言語モデル構築支援プログラム」の支援を受けました．この成果は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の結果得られたものです．本研究は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

参考文献

[1] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, et al. The FineWeb datasets: Decanting the Web for the finest text data at scale. arXiv:2406.17557, 2024.

[2] Jeffrey Li, Alex Fang, Georgios Smyrnis, et al. DataComp-LM: In search of the next generation of training sets for language models. arXiv:2406.11794, 2024.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. arXiv:2404.05405, 2024.

[4] Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models to domains via reading comprehension. arXiv:2309.09530, 2023.

[5] Jie Chen, Zhipeng Chen, Jiapeng Wang, et al. Towards effective and efficient continual pre-training of large language models. arXiv:2407.18743, 2024.

[6] Zhengbao Jiang, Zhiqing Sun, Weijia Shi, et al. Instruction-tuned language models are better knowledge learners. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5421–5434. Association for Computational Linguistics, August 2024.

[7] Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. arXiv:2409.07431, 2024.

[8] Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 technical report. arXiv:2412.08905, 2024.

[9] 服部翔, 岡崎直観, 水木栄ほか. Swallow コーパス v2: 教育的な日本語ウェブコーパスの構築. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[10] Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, et al. Cosmopedia. Hugging Face, 2024.

[11] Namgi Han, 植田暢大, 大嶽匡俊ほか. 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.

[12] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proc. of LREC**, pp. 2957–2966, 2022.

[13] 石井愛, 井之上直也, 関根聡. 根拠を説明可能な質問応答システムのための日本語マルチホップ QA データセット構築. 言語処理学会第 29 回年次大会 (NLP2023), 2023.

[14] 関根聡. 百科事典を対象とした質問応答システムの開発. 言語処理学会第 9 回年次大会 (NLP2003), 2003.

[15] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, et al. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In **Findings of ACL-IJCNLP 2021**, 2021.

[16] Freda Shi, Mirac Suzgun, Markus Freitag, et al. Language models are multilingual chain-of-thought reasoners. In **Proc. of ICLR**, 2023.

[17] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, and *others*. Findings of the 2020 conference on machine translation (WMT20). In **Proceedings of the Fifth Conference on Machine Translation**, Online, 2020. Association for Computational Linguistics.

[18] 佐藤美唯, 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.

[19] 尹子旗, 王昊, 堀尾海斗ほか. プロンプトの丁寧さと大規模言語モデルの性能の関係検証. 言語処理学会第 30 回年次大会 (NLP2024), 2024.

[20] Todor Mihaylov, Peter Clark, Tushar Khot, et al. Can a suit of armor conduct electricity? a new dataset for open book question answering. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, Brussels, Belgium, 2018. Association for Computational Linguistics.

[21] Mandar Joshi, Eunsol Choi, Daniel Weld, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, Vancouver, Canada, 2017. Association for Computational Linguistics.

[22] Rowan Zellers, Ari Holtzman, Yonatan Bisk, et al. HellaSwag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, Florence, Italy, July 2019. Association for Computational Linguistics.

[23] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers**. Association for Computational Linguistics, 2018.

[24] Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In **Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021**, Vol. ACL/IJCNLP 2021 of **Findings of ACL**. Association for Computational Linguistics, 2021.

[25] Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. arXiv:2009.03300, 2021.

[26] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

[27] Mirac Suzgun, Nathan Scales, Nathanael Schärli, et al. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In **Findings of the Association for Computational Linguistics: ACL 2023**. Association for Computational Linguistics, July 2023.

[28] Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

[29] 植木快, 川畑輝, 田口雄哉ほか. 時事情報に関する日本語 QA ベンチマーク『ニュース Q』. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[30] 今城健太郎, 平野正徳, 鈴木脩司, 三上裕明. pfgen-bench: 日本語事前学習モデルのための文章生成性能評価ベンチマーク, 2024.

[31] Koshiro Saito, Sakae Mizuki, Masanari Ohi, et al. Why we build local large language models: An observational analysis from 35 Japanese and multilingual LLMs. arXiv:2412.14471, 2024.

A　合成データ生成のプロンプト

2.2 節で述べたように，本研究では新聞記事から QA 形式と教科書形式の 2 種類の合成データを生成した．以下に，それぞれの生成にあたって実際に用いたプロンプトを示す．

QA 形式の合成データ生成のプロンプト

あなたは優秀な塾講師です。これから 1 つのテキストが与えられます。以下の指示に従い、与えられたテキストから、教育的価値が高く、社会や学問において幅広い活用機会がある知識を抽出し、生徒が反復的に演習形式で学習できる教材として、問題集とその解答解説を作成してください。

 # 指示

全ての問題は、問題文に続けて解答と解説を作成してください。
1. まず、テキスト全体を通して重要な用語が正答となるような、短答式の問題を 3 問作成してください。
2. 続いて、テキスト全体を通して重要なトピックに関する理解を問う、多肢選択式の問題を 2 問作成してください。
3. 最後に、テキスト全体を通して重要なトピックに関して 2〜3 文程度で説明させるような、論述式の問題を 2 問作成してください。例えば「〇〇 とは何か詳しく説明してください。」「〇〇 が 〇〇 なのは何故でしょうか？」といった、特定のトピックの詳細や因果関係について論理的に説明を求める問題が想定されます。

 # テキスト

{TEXT}

 # 出力

教科書形式の合成データ生成のプロンプト

あなたは優秀な塾講師です。これから 1 つのテキストが与えられます。以下の指示に従い、与えられたテキストに含まれる教育的な内容を解説し、教科書の形式でまとめてください。

 # 指示

1. 各章の冒頭に、学習目標や目的を明示します。
2. 知識は順序立てて説明し、基礎的な概念から応用的な内容へと進行します。各章やセクションは、特定のテーマに焦点を当て、関連する情報をまとめて提示します。
3. 専門的な概念や理論を、できるだけ簡潔で分かりやすい言葉で説明します。新しい用語や定義には、特別な注意を払い、繰り返し説明することもあります。
4. 全ての前提や背景情報は教科書内で提供します。学習者は教科書のみを読んで、内容を十分に理解することができます。
5. 章末に「まとめ」を提供することで、学習者が内容を再確認し、理解を深められるようにします。また、特に教育的価値が高く、社会や学問において幅広い活用機会がある専門的な用語や概念は抽出して一般化し、補足的な解説を加えます。

 # テキスト

{TEXT}

 # 出力

B　構築したデータセットの事例

2 節で構築した新聞記事および合成データの一例を以下に示す（「.....」は省略した部分である）．**橙色**は新聞記事の情報が忠実に引用された部分，**水色**は新聞記事に含まれる用語に関連する情報を LLM（Gemma 2）が独自に補完し，生成したとみなせる部分である．生の新聞記事から生成した QA 形式の合成データは，新聞記事の情報をそのまま問題と解答の形式に変換しているのに対し，教科書形式の合成データや教科書形式の合成データから生成した QA 形式の合成データは，LLM（Gemma 2）が独自に生成した周辺知識も多く含まれていることが確認できる．

生の新聞記事

2020 年東京五輪・パラリンピックに向けてユニバーサルデザイン (UD) のタクシーを増やそうと、東京都が新規購入する事業者に 1 台あたり 60 万円を補助する事業を始める。新年度当初予算案に 61 億円を計上し、5 年間で都内のタクシー約 5 万台の 2 割にあたる 1 万台の普及をめざす。
UD タクシーは、スロープやリフトを備えたワゴンタイプの車両で、足腰の弱いお年寄りや障害者が利用しやすい。
補助の対象は、車いすのまま乗車できるハイブリッド車、プラグインハイブリッド車、電気自動車の 3 種類。国土交通省も UD タクシーに同額を補助しているが、都は環境性能の高い車両に限定。国の補助制度との併用も可能という。都によると、ハイブリッド車の UD タクシーは約 330 万円。国交省や都などの説明では、UD タクシーは昨年 3 月末現在、全国で 690 台走っている。うち都内は約 50 台。(伊藤あずさ)
（引用：2016 年 1 月 5 日付　朝日新聞　夕刊）

QA 形式の合成データ（生の新聞記事由来）

.....

(1) 東京都が新規購入する事業者に補助する UD タクシー 1 台あたりの金額はいくらですか？
解答: 60 万円
解説: テキストによると、東京都は新規購入する事業者に 1 台あたり 60 万円を補助する事業を始めると述べています。

(2) UD タクシーに求められる環境性能の高い車両の種類を 3 つ挙げなさい。
解答: ハイブリッド車、プラグインハイブリッド車、電気自動車
.....

教科書形式の合成データ

第 1 章 ユニバーサルデザインタクシーの普及
.....

1.2 ユニバーサルデザインとは

ユニバーサルデザイン (UD) とは、年齢、性別、身体能力、知識、経験にかかわらず、すべての人が使いやすい製品や環境を設計することです。誰もが快適に、そして平等に利用できる社会の実現を目指します。
.....

1.3 ユニバーサルデザインタクシー

ユニバーサルデザインタクシーは、**UD の理念に基づき、誰もが利用しやすいように設計されたタクシー**です。従来のタクシーでは、車椅子を利用する人や高齢者などが利用しづらい場合がありました。UD タクシーは、

* **スロープやリフトの設置**: 車椅子でそのまま乗車できるよう、スロープやリフトが備えられています。
* **車内の空間確保**: 車椅子がスムーズに移動できるよう、車内スペースが広く確保されています。
* **使いやすい座席**: 高齢者や身体に不自由がある人が座りやすく、乗り降りしやすいように設計されています。
.....

QA 形式の合成データ（教科書形式データ由来）

.....

問題 1
ユニバーサルデザイン (UD) とは、どのような目的で設計されたものですか？

解答
年齢、性別、身体能力、知識、経験にかかわらず、すべての人が使いやすい製品や環境を設計すること。誰もが快適に、そして平等に利用できる社会の実現を目指す。
.....