新聞ドメインにおける 大規模言語モデルの継続事前学習と下流タスクデータ量の関係

岸波 洋介 [1]　藤井 諒 [1]　森下 睦 [1]
[1] フューチャー株式会社
{y.kishinami.rh, r.fujii.6d, m.morishita.pi}@future.co.jp

概要

近年の大規模言語モデル (LLM) の発展により，様々なドメインに特化した LLM の研究が進んでいる．ドメイン特化 LLM を下流タスクに用いる場合，タスクのラベル付きデータで教師ありファインチューニングする前に，そのドメインの知識獲得を目指し，特定ドメインの生テキストを用いた継続事前学習が先行して行われることがある．しかしながら下流タスクのデータ量に着目した継続事前学習の有効性については不明な点が多い．本研究では新聞ドメインにおける見出し生成タスクを対象に，ドメイン特化を目的とする継続事前学習と下流タスクのデータ量が性能に与える影響を分析する．分析の結果，下流タスクのデータ量が極めて少ない状況では継続事前学習の効果が大きい可能性が示唆された．

1　はじめに

質問応答や文章校正など，様々な用途で大規模言語モデル (LLM) の活用が進んでいる [1, 2, 3]．特に企業での応用を見据えると，金融，医療などといった業界に特化した LLM のニーズは大きく，様々な業界で活用，研究が進んでいる [4, 5, 6, 7]．LLM の実応用に際しては，OpenAI 社が提供する ChatGPT のように，様々なタスクに汎用的に活用できるチャットボットのニーズも高い一方，解きたいタスクが明確に定まっている場合も一定存在する．

LLM を用いて解きたいタスクが明確に定まっている場合，解きたいタスクのラベル付きデータを用いて教師ありファインチューニング (SFT) を行うことが一般的である [8]．しかしながら，高いタスク性能を達成するためには大量のラベル付きデータが必要となる場合も存在する．タスクによって現実的に用意できるデータ量は異なり，大規模に用意するのは高コストであることも多い．そのため，ベースとなる LLM を特定ドメインの生テキストで追加学習することで，そのドメインの知識獲得を目指す継続事前学習が先行して行われることがある [5, 9, 10]．継続事前学習では，学習に用いるデータに対して正解ラベルを付与する必要がないため，比較的学習データを収集しやすいという特徴がある．しかしながら，実際に継続事前学習することがどのような状況下で有効か，特に下流タスクのデータ量に着目するといまだ不明な点が多い．

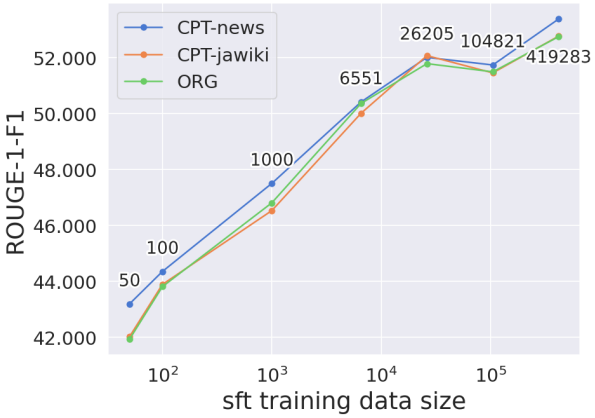

図 1　見出し生成タスクの学習データ量と性能 (ROUGE)．マーカー付近の値は見出し生成タスクの学習データ量を示す．

本研究では新聞ドメインにおける見出し生成タスクを対象に，タスクのデータ量に着目した場合の継続事前学習の有効性を検証した．多くの新聞記事は本文に対して見出しが付与された状態で公開され，本文と見出しの組を集めやすいことから，下流タスクのデータ量による継続事前学習の影響を調査する上で適したタスクであると考えることができる．本研究では，新聞ドメインへ特化させるための継続事前学習を行った上で，SFT に用いるデータ量による性能の差を分析する．分析の結果，SFT の学習データが極めて少ない状況下では継続事前学習の効果が大きい可能性が示唆された．

2 関連研究

2.1 ドメイン特化型 LLM

ChatGPT に代表される汎用的な LLM に対し，特定ドメインに特化した LLM も研究が進められている [4, 5, 6, 7]．金融分野では BloombergGPT [4] や FinGPT [11] など，様々な金融特化 LLM が構築されている [12, 13]．日本語でも，平野らが日本語 LLM に対して金融文書で継続事前学習を行い，金融特化 LLM を構築している [9]．また，医療分野でもドメイン特化の LLM に関する研究，開発が進められている [5, 14, 15, 16]．例えば Llama を医学論文で継続事前学習した PMC-LLaMA [5] や，Med-PaLM [14] などが挙げられる．日本語においても助田らが日本語 LLM に対し医療データで指示チューニングを行い，医療ドメインに特化した LLM を構築している [10]．本研究では新聞記事ドメインを対象に，継続事前学習と下流タスクのデータ量に着目した分析を行う．石原らも BERT や T5 に対し新聞記事で学習を行い，日本語の金融ニュース記事に特化したモデルを構築しているが [17]，本研究は近年発展の著しいより大規模なモデルを対象に，継続事前学習と下流タスクのデータ量に着目した分析を行う．

2.2 ドメイン特化の継続事前学習の分析

最近では，継続事前学習の有効性に関する分析も多数行われている [18, 19, 20]．Yıldız らは LLM の継続事前学習に関して，複数ドメインで構成されるデータセットを用いて，学習するドメインの内容や学習順序に着目した分析を行っている [18]．また，Xie らは金融ドメインを対象に継続事前学習に有用なデータの選択手法を提案し，それらの手法を用いた継続事前学習の有効性を分析している [19]．実験の結果，提案手法によって少ないデータでも効率的に継続事前学習を行えることを示している．Jindal らは継続事前学習の有効性に関して，学習元とするモデルの選択に着目した分析を行っている [20]．実験では継続事前学習のデータ量を変化させた場合の分析も行っており，指示追従データで学習済みのモデルを学習元とした継続事前学習ではデータ量が増えるほど指示追従能力が劣化することを示している．本研究では特に下流タスクのデータ量に着目し，継続事前学習の有効性を分析する．

3 分析 1: 継続事前学習と下流タスクデータ量の関係

本研究では山田ら [21] の定義に倣い，与えられたテキスト全体のトークンを予測する学習方法を継続事前学習，ラベル側のトークンのみ予測する学習方法を教師ありファインチューニング (SFT) と呼ぶ．

新聞ドメインにおける見出し生成タスクを対象に，継続事前学習の有無と下流タスクでの SFT におけるデータ量の関係を分析する．

3.1 実験設定

データセット 継続事前学習，および見出し生成タスクの学習データとして，信濃毎日新聞株式会社が発行した 2013 年 1 月から 2024 年 6 月までの紙面記事の本文，見出しを使用した．連載名の接頭辞などのノイズを除去し，掲載時期に応じてデータを分割した．最終的に学習データとして 2013 年 1 月から 2024 年 4 月までの記事 419283 件，検証データとして 2024 年 5 月の記事 2385 件，評価データとして 2024 年 6 月の記事 2574 件を用意した．このうち継続事前学習には学習データの記事本文，計約 1.85 億トークンを用いた．また，比較対象として新聞以外のドメインの生テキストも用意した．具体的には llm-jp-corpus-v3[1] の日本語 Wikipedia から新聞記事と同量のトークンを抽出した．なお，見出し生成タスクへの SFT では本文，見出しの組を用いた．

継続事前学習 Llama-3.1-Swallow-8B-v0.1[2] をベースモデルとし，新聞記事本文，および日本語 Wikipedia のそれぞれを用いて継続事前学習を行った．なお，学習は 2 エポック行った．

見出し生成タスクへの SFT Llama-3.1-Swallow-8B-v0.1 (ORG)，新聞記事本文で継続事前学習を行ったモデル (CPT-news)，日本語 Wikipedia で継続事前学習を行ったモデル (CPT-jawiki) の 3 つをベースモデルとし，見出し生成タスクへの SFT を行った．SFT 手法として Low Rank Adaptation (LoRA) [22] を用いた．学習に用いるデータ量は複数の設定を用意した．具体的には，全量 (419283)，4 分の 1 (104821)，16 分の 1 (26205)，64 分の 1 (6551)，1000, 100, 50 の計 7 設定を用意し，それぞれで SFT を行った．学習設定の詳細は付録 A に示す．

1) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
2) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-v0.1

表 1　信濃毎日新聞 2024 年 6 月 27 日の記事に対する生成例．\n は改行を示す．

本文	信州大（本部・松本市）は２６日、２０２５年度入学者の選抜要項を発表した。２２年度から実施された高校の新学習指導要領に基づく新教育課程に対応した教科・科目に変更し、一般選抜（前期・後期）では旧教育課程の履修者が不利にならないよう出題する。８学部全体の入学定員は１９６３人で、前年度に引き続き医学部医学科の定員１５人増を国に申請する。\n　医学科の定員増は医師不足に対応する措置で、認められれば入学定員は前年度と同じ１９７８人。定員増の申請は１０月ごろに結果が分かる。\n　農学部農学生命科学科は２４年度までの４コースを「生命・食品科学」「食料生産システム科学」「山岳圏森林・環境共生学」の３コースと「地域協創」の１特別コースに再編。分野横断型教育を強化し、課題解決型学習を推進する。同学科の入学定員は前年度と同じ１７０人。\n　教育学部は５コース（現代教育、国語教育、ものづくり・技術教育、特別支援教育、心理支援教育）で総合型選抜を初めて導入する。募集人員は一般選抜や学校推薦型選抜を削り、１６人分の枠を設けた。卒業後に県内で教職に就く意欲を持った人が対象で、大学入学共通テストは課さない。\n　一般選抜の出願期間は来年１月２７日～２月５日。信大ホームページから出願登録サイトに必要事項を入力し、検定料を支払った上で出願書類などを郵送する。前期試験は２月２５日（一部学部は２６日も）、後期試験は３月１２日。問い合わせは平日に信大入試課（☎０２６３・３７・２１９５）へ。\n
正解見出し	信大　２５年度選抜、新学習指導要領に対応　旧課程にも配慮
ORG	信州大、２５年度入学者の選抜要項発表　医学部医学科の定員１５人増を申請
CPT-jawiki	信州大、２５年度入学者の選抜要項発表　医学部医学科の定員１５人増を申請
CPT-news	信大の２５年度入学者選抜要項発表　医学部医学科の定員１５人増を申請　教育学部は総合型選抜導入

表 2　見出し生成タスクの学習データ量を変化させた際の評価結果 (ROUGE)．横軸は学習データの件数を示す．

	50	100	1000	6551	26205	104821	419283
ORG	41.93	43.81	46.81	50.37	51.79	51.50	52.76
CPT-jawiki	42.02	43.89	46.53	50.02	**52.08**	51.46	52.78
CPT-news	**43.19**	**44.35**	**47.50**	**50.42**	52.02	**51.74**	**53.39**

評価指標　見出し生成タスクの評価指標として ROUGE-1-F1 [23] (ROUGE) と BERTScore-F1 [24] (BERTScore) を用いた．ROUGE の算出には sacrerouge[3] を用い，単語の分割には形態素解析器 MeCab [25] (IPA 辞書) を使用した．BERTScore の算出には bert-score ライブラリ[4] を用い，モデルには tohoku-nlp/bert-large-japanese-v2[5] を使用した．

3.2　実験結果

図 1 および表 2 に，見出し生成タスクの学習データ量を変化させた際の各モデルのスコアの推移を示す[6]．図 1 から，タスクの学習データ量によらず基本的に CPT-news の性能が高くなっていることがわかる．また表 2 から，特に学習データが 50 件の場合に，CPT-news の性能とそれ以外のモデルの性能の差が大きくなっていることがわかる．このことから，見出し生成タスクにおいては，タスクの学習データ量が極めて少ない状況において継続事前学習の効果が大きい可能性が示唆される．この傾向は，ROUGE だけでなく BERTScore の値についても一貫して確認された[7]．

3.3　定性分析

表 1 にタスクの学習データが 50 件の場合の各モデルの見出し生成例を示す．この例では，CPT-news のみ正解見出し同様に「信大」という略称を生成できていた．この事例について 10 回の生成を行ったところ，CPT-news は 10 回全てで「信大」と生成したのに対して，ORG では 2/10 回，CPT-jawiki では 3/10 回の生成にとどまっていた．継続事前学習に用いたデータセットに「信大」という表現がどの程度出現するかを調査したところ，新聞記事では全体の 1.25%，日本語 Wikipedia では 0.14%の文書に出現していることがわかった．この結果から，新聞記事で継続事前学習を行うことで，この表現に代表される対象ドメインらしい表現を生成する傾向が強まったと考えられる．

4　分析 2: 継続事前学習における学習データのドメイン混合

3 節の結果から，見出し生成タスクの学習データが 50 件のように極めて少ない場合に継続事前学習の効果が大きいことが確認できた．継続事前学習に用いる学習データはラベルが不要であるため，比較的データを収集しやすい．しかしながら，ドメインによっては生テキストの収集すら困難なケースも考えられる．例えば企業独自の知識や用語に関する

3）https://github.com/danieldeutsch/sacrerouge
4）https://github.com/Tiiiger/bert_score
5）https://huggingface.co/tohoku-nlp/bert-large-japanese-v2
6）過学習の傾向が見られたため，継続事前学習の 1 エポック目のチェックポイントから SFT を行った．
7）BERTScore の結果は付録 B に示す．

ドメインでは，その企業特有の情報が精緻に文書化されていなければ大規模に収集することは難しい．一方で llm-jp-corpus など，一般公開されているコーパスであれば低コストで扱いやすい．そこで本節では，下流タスクのドメインのデータとそれ以外のドメインのデータの混合データで継続事前学習した場合の性能について分析する．

4.1 実験設定

データセット 新聞ドメイン，他ドメインのデータを混ぜ合わせたデータセットを用意した．新聞ドメインの学習データには 3 節で用いた新聞記事本文約 1.85 億トークン，他ドメインの学習データには 3 節で用いた日本語 Wikipedia を同量使用した．合計トークン数が約 1.85 億トークンとなるように，全量を新聞ドメインにした場合，新聞ドメインを 2/3，1/2，1/3，1/10，1/20 含めた場合，全量を他ドメインにした場合の計 7 設定を用意した．

継続事前学習 7 つの学習データそれぞれを用いて，LLM の継続事前学習を実施した． 3 節と同様に Llama-3.1-Swallow-8B-v0.1 をベースモデルとし，学習は 2 エポック行った．

見出し生成タスクへの SFT 7 つの学習データで継続事前学習を行った各モデルに対し，見出し生成タスクへの SFT を行った．見出し生成の学習データは 3 節の分析において継続事前学習の効果が最も大きくなっていた 50 件とした．

評価指標 3 節と同様に見出し生成の評価指標には ROUGE-1-F1 (ROUGE) と BERTScore-F1 (BERTScore) を用いた．

4.2 実験結果

図 2 および表 3 に継続事前学習の学習データにおける新聞記事の割合を変化させた際のスコアの推移を示す[8]．図 2 から，新聞ドメインを 1/10 含めた場合と 1/3 含めた場合との間にスコアの差が生じていることがわかる．この傾向は ROUGE だけでなく BERTScore の値についても一貫して確認された[9]．このことから，下流タスクと同一のドメインと他のドメインのデータを一定割合以上で混ぜ合わせると性能が急激に向上する可能性が示唆される．本実験で用いた Llama-3.1-Swallow-8B-v0.1 は既に日本語 Wikipedia を用いた学習が行われている．そのため，

8） 3 節と同様に継続事前学習の 1 エポック目のチェックポイントから SFT を行った．
9） BERTScore の結果は付録 B に示す．

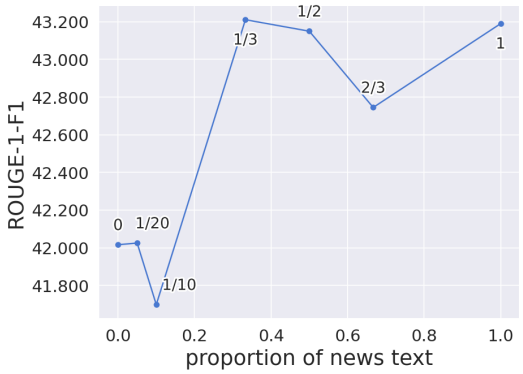

図 2 継続事前学習の学習データにおける新聞記事の割合を変化させた際の性能 (ROUGE)．マーカー付近の値は新聞記事の割合を示す．

表 3 継続事前学習の学習データにおける新聞記事の割合を変化させた際の評価結果．横軸は新聞記事の割合を示す．

	0	1/20	1/10	1/3	1/2	2/3	1
ROUGE	42.02	42.03	41.70	**43.21**	43.15	42.75	43.19
BERTScore	0.679	0.681	0.678	0.684	**0.685**	0.682	0.684

混合データによる学習が過去に学習した言語能力を一定保持しつつ新規ドメインに適用するような効果を発揮している可能性も考えられる．このような過去に学習したドメインとの類似性に着目した混ぜ合わせの分析については今後の課題としたい．

5 おわりに

本研究では新聞ドメインにおける見出し生成タスクを対象に，継続事前学習と下流タスクの学習データ量の関係を調査した．分析の結果，見出し生成タスクにおいては，SFT の学習データが極端に少ない場合に継続事前学習の効果が大きい可能性が示唆された．また，継続事前学習に用いる学習データにおいて見出し生成タスクと同一ドメインのデータの割合を変化させたところ，継続事前学習の学習データに他のドメインのデータが混合している場合にもタスクの性能向上につながるケースが確認された．

本研究はあくまで新聞ドメインの見出し生成タスクに着目した分析となっており，他のドメイン，タスクに対しても同様の傾向があることまでは示すことができない．したがって今後は金融や特許などの他ドメインや，それらの下流タスクでも分析し，ドメインやタスクによらず同様の傾向がみられるのかを調査したい．また，学習元とするモデルの種類やモデルサイズによる影響も考えられるため，これらを変更した際の分析も行いたいと考えている．

謝辞

本研究の実施にあたり，データを提供いただいた信濃毎日新聞株式会社の皆様に感謝申し上げます．また，本研究成果の一部は，データ活用社会創成プラットフォーム mdx および東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して得られたものです．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, pp. 4171–4186, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In arXiv:2005.14165, 2020.

[3] OpenAI. GPT-4 technical report. In arXiv:2303.08774, 2024.

[4] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. In arXiv:2303.17564, 2023.

[5] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards building open-source language models for medicine. In arXiv:2304.14454, 2023.

[6] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. In arXiv:2211.09085, 2022.

[7] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. ChemLLM: A chemical large language model. In arXiv:2402.06852, 2024.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Proceedings of NeurIPS, pp. 27730–27744, 2022.

[9] Masanori Hirano and Kentaro Imajo. Construction of domain-specified Japanese large language model for finance through continual pre-training. In arXiv:2404.10555, 2024.

[10] 助田一晟, 鈴木雅弘, 坂地泰紀, 小寺聡. JMedLoRA：instruction-tuning による日本語大規模モデルの医療ドメイン適用. 言語処理学会年次大会発表論文集, pp. 2548–2553, 2024.

[11] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. In arXiv:2306.06031, 2023.

[12] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. In arXiv:2306.12659, 2023.

[13] Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G. Constantinides, and Danilo Mandic. FinLlama: Financial sentiment classification for algorithmic trading applications. In arXiv:2403.12285, 2024.

[14] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. In arXiv:2212.13138, 2022.

[15] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. In arXiv:2305.09617, 2023.

[16] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. In arXiv:2308.09442, 2023.

[17] 石原祥太郎, 村田栄樹, 中間康文, 高橋寛武. 日本語ニュース記事要約支援に向けたドメイン特化事前学習済みモデルの構築と活用. 自然言語処理, Vol. 31, No. 4, pp. 1717–1745, 2024.

[18] Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pre-training in large language models: Insights and implications. In arXiv:2402.17400, 2024.

[19] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models. In Proceedings of ACL, pp. 10184–10201, 2024.

[20] Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in LLMs. In arXiv:2410.10739, 2024.

[21] 山田正嗣, 井本稔也. 金融ドメイン特化のための大規模言語モデルのインストラクションチューニング評価. 人工知能学会全国大会論文集, 2024.

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In Proceedings of ICLR, 2022.

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pp. 74–81, 2004.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In Proceedings of ICLR, 2020.

[25] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Proceedings of EMNLP, pp. 230–237, 2004.

表 4　見出し生成タスクの学習データ量を変化させた際の評価結果 (BERTScore)．横軸は学習データの件数を示す．

	50	100	1000	6551	26205	104821	419283
ORG	0.679	0.689	0.700	**0.720**	0.726	0.726	0.732
CPT-jawiki	0.679	0.690	0.698	0.718	0.726	0.726	0.731
CPT-news	**0.684**	**0.691**	**0.704**	**0.720**	**0.727**	**0.727**	**0.734**

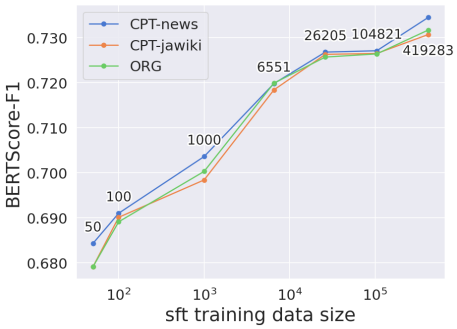

図 3　見出し生成タスクの学習データ量と性能 (BERTScore)．マーカー付近の値は見出し生成タスクの学習データ量を示す．

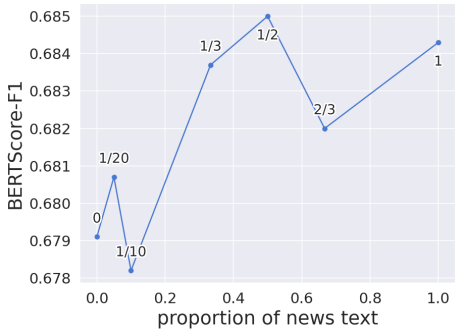

図 4　継続事前学習の学習データにおける新聞記事の割合を変化させた際の性能 (BERTScore)．マーカー付近の値は新聞記事の割合を示す．

表 5　信濃毎日新聞 2024 年 6 月 12 日の記事に対する生成例．\n は改行を示す．

本文	県内でツキノワグマによる人的被害が増加しているのを受け、県は１１日、「野生鳥獣被害対策本部会議」を県庁で開いた。７月１２日までの約１カ月間、熊が出没する恐れがある地域での「集中点検」を実施すると決めた。\n　集中点検は、県や市町村職員、県クマ対策員らが熊の目撃があった地域などを見回る。地域の土地所有者らには、熊の餌になる物の除去や生ごみなどの適切な処理、熊が身を潜められるやぶの刈り払いを呼びかける。\n　県内でのツキノワグマの目撃件数は今年５月、平年の１・８倍の１０６件。県は今月５日に「ツキノワグマ出没注意報」を出したが、８日に上高井郡高山村で４０代女性が右手をかまれるなどして重傷を負うなど、６月に入って人的被害は４件発生している。\n　会議では、県環境保全研究所の研究員が「体長１メートル未満の熊の目撃が多く、人里への恐怖心が薄い親離れしたばかりの若い熊が餌を探して動き回っている可能性がある」と説明。県森林づくり推進課の塚平賢治・鳥獣対策担当課長は会議後の取材に「熊の出没を繰り返さないためには地域ぐるみでの対策が大切。集中点検で必要な取り組みを助言していきたい」と話した。\n
正解見出し	熊警戒　来月１２日まで「集中点検」　県、人的被害の拡大受け
ORG	ツキノワグマ被害増加 県、集中点検実施へ 県庁で対策本部会議
CPT-jawiki	県内でツキノワグマによる人的被害増加 県、集中点検実施へ
CPT-news	ツキノワグマの人的被害増加　県が「集中点検」実施へ

A　実験設定の詳細

継続事前学習　3 節および 4 節で同一の学習設定を用いた．バッチサイズを 16，weight decay を 0.1，gradient clipping を 1.0 とした．また，初期学習率は 5e-6 とし，5e-7 まで線形に減衰させた．学習には Megatron-LM[10] を使用し，NVIDIA H100 94GB 4GPU で実施した．

SFT　LoRA-R を 64，LoRA-alpha を 16，LoRA-dropout を 0.15，学習対象のパラメータは all-linear とした．また，学習率は 0.0001，最大系列長は 3200 とした．バッチサイズは学習データ 1000 件以下の場合は 1，6551 件では 2，26205 件では 8，104821 件では 32，419283 件では 128 とした．各モデルは 1 エポック学習し，最終チェックポイントを評価に用いた．学習には trl[11] を使用した．実験は学習データ 104821 件以上の場合は NVIDIA A100 40GB 8GPU で，それ以外の設定では NVIDIA A100 40GB 1GPU で実施した．

見出し生成タスクのプロンプト　SFT および推論の際に与えるプロンプトは Llama-3.1-Swallow-8B-Instruct-v0.1[12] のテンプレートに従った．具体的には，システムプロンプトに「あなたは誠実で優秀な日本人のアシスタントです。」を，ユーザプロンプトに「次の新聞記事に見出しをつけてください。」，改行，記事本文を，アシスタントプロンプトに記事見出しを埋め込んだ．学習の際には記事見出しの先頭トークン以降に損失を流すようにした．

B　実験結果の詳細

定量評価　図 3 および表 4 に 3 節の実験における各モデルの BERTScore を示す．また，図 4 に 4 節の実験における BERTScore の推移を示す．

定性分析　表 5 は 3 節の実験における学習データ 50 件の場合の見出し生成例である．表 5 を見ると，CPT-news のみ正解見出し同様に集中点検というフレーズを鍵括弧で囲むことができていた．この事例について 10 回の生成を行ったところ，CPT-news は 9 回「集中点検」のように鍵括弧で囲まれた表現を生成したのに対して，ORG では 0/10 回，CPT-jawiki では 3/10 回の生成にとどまっていた．継続事前学習に用いたデータセットに鍵括弧で囲まれた表現がどの程度出現するかを調査したところ，新聞記事では全体の 50.0%，日本語 Wikipedia では 10.5%の文書に出現していることがわかった．この結果からも，新聞記事で継続事前学習を行うことで対象ドメインらしい表現を生成する傾向が強まったことが示唆される．

10）https://github.com/NVIDIA/Megatron-LM
11）https://github.com/huggingface/trl
12）https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1