英語中心の大規模言語モデル開発からの脱却に向けて

高瀬 翔　本多 右京
サイバーエージェント
{takase_sho,honda_ukyo}@cyberagent.co.jp

概要

大規模言語モデルの生成した系列の分析を通じ，大規模言語モデルの事前学習データは大規模 Web コーパスである CommonCrawl における言語分布と比較して，はるかに英語に偏っている可能性を示す．また，学習済みの大規模言語モデルを特定の言語に適応させる手法として継続事前学習が知られているが，日本文化理解のタスクにおける性能は日本語の学習に費やした推定総コストに対数比例していることを示し，継続事前学習が学習効率の観点では事前学習と同等であることを示す．本稿で得られた知見を通じて，大規模言語モデル開発が英語中心の状況から脱却し，各言語に寄り添ったベンチマーク構築やモデル開発が促進されることを期待する．

1　はじめに

GPT-3 [1] の優れた性能報告依頼，Llama や Qwen，Gemma シリーズなど，様々な組織が大規模言語モデルを発表してきた [2, 3, 4]．新たな大規模言語モデルが次々と発表され，ベンチマークタスクでの性能向上が日々報告されている．しかしながら，既存の大規模言語モデルの大半は英語に主眼を置いており，主要なベンチマークデータも英語で構築されている．このため，英語以外の言語については議論が十分になされているとは言い難い．

実際，いくつかのテクニカルレポートにおいて，大規模言語モデルの学習に使用したデータ中で英語が占める割合は有意に多いと報告されている [5, 3]．例えば，Llama 2 の事前学習データはおおよそ 90% が英語文書で構成されている．大規模に Web をクロールしたコーパスとして知られる CommonCrawl においては，英語文書の割合はおおよそ 40%であるため[1)]，Llama 2 の事前学習データは Web コーパスにおける言語分布と比較して英語に大きく偏っていると言える．本研究では，各大規模言語モデルについて，生成した系列を元に学習データの言語分布を推定し，多言語対応の大規模言語モデルと謳っているモデルであっても学習データは英語に大きく偏っている可能性を示す（図 1）．

特定の言語に対応した高品質な大規模言語モデルを効率的に構築する手法として継続事前学習が知られているが，本研究では継続事前学習されたモデルの分析を通じて，継続事前学習は特定の言語の大規模言語モデルを「極めて効率的に構築する手法」ではない可能性を示す．継続事前学習は確かに対象言語の性能を向上させるが，追加した学習コスト相応の性能向上であること，特に，文化に関連した知識を問う問題については，学習の総コストではなく，その問題の言語の学習に費やしたコストに性能が対数比例することを示す（図 2）．すなわち，文化に関連した知識を問う問題については継続事前学習の効率はゼロから事前学習する場合と同等であることを示す．これらの結果を元に大規模言語モデル開発の方向性について議論する．

2　大規模言語モデルの言語分布

公開されている大規模言語モデルの中にはテクニカルレポートを通じて構築過程が説明されているモデルも存在するが [5, 6, 7]，事前学習データに含まれる言語の分布はあまり報告されていない．本研究では，大規模言語モデルの生成する系列は学習データと強い相関があるという仮定の元，事前学習データの言語分布を生成した系列から推定する．実際に，大規模言語モデルの使用した学習データをモデルに生成させることは可能であると報告されており [8, 9, 10]，生成した系列から学習データを推定する手続きは妥当であると考えられる．

分布の推定　大規模言語モデルに文頭を表す特殊記号である BOS トークンを与え，1,000 トークンさせるという手続きを 100,000 系列分行う．その後，fastText [11, 12] を用いて各系列の言語を特定する．

1）https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

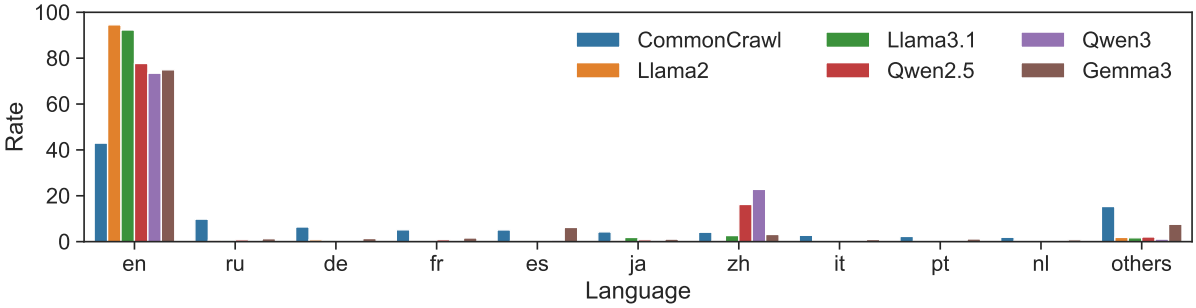


図 1　各大規模言語モデルについて生成した系列から推測した事前学習データの言語分布．CommonCrawl コーパスでの上位 10 言語に限定している．

表 1　MGSM データセットにおける各大規模言語モデルの 6-shot での性能．モデル名は HuggingFace Hub に記載されている名前を使用している．

モデル	En	Ru	De	Fr	Es	Ja	Zh	Th	Te	Bn	Sw	平均スコア
Qwen2.5-32B-Instruct	87.2	72.8	71.2	64.4	86.4	76.0	80.8	69.2	58.0	82.4	56.8	73.2
Qwen3-32B	95.6	86.4	88.0	82.4	91.6	77.6	89.6	88.8	84.0	83.2	74.4	85.6
Llama3.1-8B-Instruct	78.8	62.4	67.2	64.0	73.6	53.2	60.0	60.8	53.2	54.0	54.8	62.0
Gemma-3-27B-it	96.0	86.0	86.8	80.4	92.4	75.6	84.8	88.8	82.8	84.8	84.4	85.7

得られた言語の分布を事前学習データの言語分布の推定値として扱う．この手続きの妥当性を検証するために，学習データの言語分布を報告している数少ないモデルである Llama 2 の，テクニカルレポートで報告されている言語分布と推定した言語分布とのケンドールの順位相関係数を計算する．結果として，Llama 2 7B，13B のそれぞれについて，0.77，0.80 という値が得られた．これらの値は十分高いことから本手続きで言語分布を推定することは妥当であると考えている[2)]．

分布の比較　図 1 に，広く使われている大規模言語モデルおよび CommonCrawl コーパスについて推定した言語分布を示した．CommonCrawl コーパスについては 1,000 文字以上からなる文書を 100,000 件サンプリングし，大規模言語モデルの生成した系列の言語推定と同様に fastText を用いて言語の推定を行った．大規模言語モデルについては各組織がゼロから学習した事前学習モデルを調査の対象とした．なお，Llama 2 7B，13B について，チューニングモデルについても調査したところ，生成した系列から推定した言語分布と報告されている事前学習データの分布とのケンドールの順位相関係数がそれぞれ 0.62，0.60 と事前学習モデルよりも低かった．このため，事前学習データの言語分布の推定には事前学習モデルを用いる方が適切と考えられる．

2)　事前学習データの言語分布は公開されていないが，データ自体が公開されている LLM-jp のモデルについてもケンドールの順位相関係数を計算したところ，およそ 0.60 程度と，こちらも強い相関を示した．データのサンプリング手続きなど，詳細は付録 B に記す．

図 1 より，事前学習データの英語の割合は CommonCrawl コーパスの英語の割合よりもはるかに多いことが分かる．すなわち，各大規模言語モデルの学習データは英語に大きく偏っていると言える．このため，英語では高い性能を達成していたとしても，他の言語では十分な性能に達していない可能性がある．実際，これらの大規模言語モデルは数学タスクのような本質的には言語非依存の能力を検証するタスクであっても，英語以外の言語では性能が大きく低下する．表 1 に多言語の数学能力を検証するベンチマークである Multilingual Grade School Math（MGSM）における各大規模言語モデルの Few-shot 性能を記した．なお，Few-shot における事例数は先行研究に従い 6 としている [13]．表 1 に示されているように，ロシア語は CommonCrawl において 2 番目に多い言語であるにも関わらず，MGSM における性能はどのモデルも英語と比べて 10 ポイント程度低下している．

3　継続事前学習の効果

大規模言語モデルを対象の言語に適応させる手法として継続事前学習が知られている [14, 15, 16, 17, 18]．既存研究では継続事前学習により，ゼロから学習するよりも大幅に低いコストで対象言語の大規模言語モデルを構築可能であると報告されてきた．本節では継続事前学習の効果について，言語非依存タスクと対象言語の文化理解タスクの 2 種類を用いて再検証を行う．

表 2　言語非依存タスクと日本文化理解のタスクにおける各大規模言語モデルの性能．MGSM については 6-shot，MMLU ProX については 5-shot，NIILCQA および JMMLU については 4-shot，JAQKET については 0-shot で評価を行った．

モデル	言語非依存タスク				日本の文化理解タスク		
	MGSM		MMLU ProX				
	En	Ja	En	Ja	NIILCQA	Part of JMMLU	JAQKET
日本語を対象に継続事前学習した大規模言語モデルとその元モデル							
Llama-3.1-8B-Instruct	78.8	53.2	43.5	28.7	30.7	71.6	27.5
Llama-3.1-Swallow-8B-Instruct-v0.5	76.8	60.0	43.0	36.2	**58.3**	85.7	65.5
Qwen2.5-32B	92.0	82.4	61.1	51.0	35.4	84.3	48.4
Qwen2.5-bakeneko-32B-Instruct	**94.4**	**84.0**	**71.9**	**64.9**	32.3	83.5	51.2
Qwen2.5-32B-Instruct	87.2	76.0	68.7	61.2	26.8	84.3	49.9
ABEJA-Qwen2.5-32b-Japanese-v1.0	90.8	32.8	66.3	58.1	42.5	**90.9**	**70.5**
ELYZA-Shortcut-1.0-Qwen-32B	61.6	69.2	61.7	62.9	36.2	87.3	61.7
日本語を多く含むデータで事前学習を行った大規模言語モデル							
Sarashina2.2-3B-Instruct-v0.1	76.0	60.4	36.1	30.2	52.0	81.9	68.2
LLM-jp-3.1-13B-Instruct4	57.2	66.0	29.8	28.5	51.2	80.8	69.4

モデル　継続事前学習を行ったモデルの種類数の豊富さから，日本語大規模言語モデルを対象とする．具体的には，Llama-3.1-8B-Instruct [19] を元に学習した Llama-3.1-Swallow-8B-Instruct-v0.5，Qwen2.5-32B [6] を元に学習した Qwen2.5-bakeneko-32B-Instruct，Qwen2.5-32B-Instruct [6] を元に学習した ABEJA-Qwen2.5-32b-Japanese-v1.0 と ELYZA-Shortcut-1.0-Qwen-32B を評価に用いる．なお，これらのモデルは継続事前学習後にインストラクションチューニングが行われている．

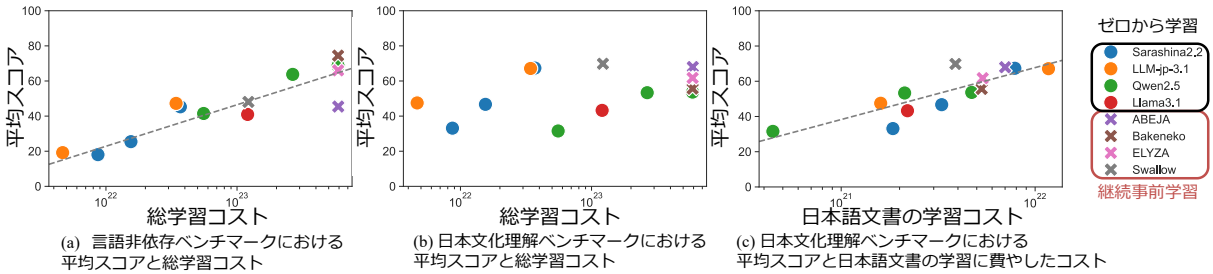
評価データセット　言語非依存タスクとして，数学タスクおよび大規模な多言語理解データセット（Massive Multilingual Language Understanding，MMLU）を用いる．小学生を対象とした英語の数学問題データセットである GSM8K [20] および MMLU を元に洗練したデータセットである MMLU Pro [21] について，英語から複数の言語に翻訳したデータセットである MGSM [13] および MMLU ProX [22] の日本語部分を用いる．また，参考として英語部分での性能も評価する．加えて，日本文化理解タスクのベンチマークデータとして，NIILC 質問応答データセット（NIILCQA） [23]，JMMLU における日本文化について問う問題部分 [24]，および JAQKET[3] を用いる．言語非依存タスクの評価には LM Evalution Harness[4] を用い，日本文化理解タスクの評価には FlexEval[5] を用いた．

結果 1: 性能向上　表 2 に言語非依存タスクおよび日本文化理解タスクの性能を示した．また，表には日本語大規模言語モデルとしてゼロから学習されたモデルの性能も参考として記している．言語非依存ベンチマークについて，Llama-3.1-Swallow-8B-Instruct-v0.5 は元のモデルと比べ日本語部分での性能は向上している．Qwen2.5-bakeneko-32B-Instruct も同様に元のモデルから性能が向上している．従って，継続事前学習は言語非依存タスクにおける，対象言語での性能を向上させることが可能である．しかしながら，ABEJA-Qwen2.5-32b-Japanese-v1.0 や ELYZA-Shortcut-1.0-Qwen-32B の結果にあるように，継続事前学習によって対象言語の性能が低下する場合もある．この結果から，質の高いモデルを構築するためには学習戦略を慎重に設計する必要があると推測される．なお，日本文化理解のタスクについては，Qwen2.5-bakeneko-32B-Instruct を除いたすべてのモデルが元のモデルから性能向上している．

結果 2: 学習効率　多くの場合，継続事前学習によって対象言語の性能は上がる場合が多いことを表 2 は示している．しかしながら，継続事前学習はどの程度効率的なのだろうか．より詳細には，継続事前学習はゼロからの事前学習と比べて，学習に費やしたコストに対しての性能向上幅が大きいのだろうか，という疑問がある．これを調査するため，学習コストと性能との関係を描画する．具体的には，言語非依存ベンチマークの日本語部分および日本文化理解ベンチマークの平均スコアと，事前学習部分も含めて学習に費やした総コスト，日本語文書の学習に費やしたコストとの関係を描画する．なお，学習コストとしてはパラメータ数と学習トークン数の積を用いる．事前学習における日本語文書の学習トークン数としては 2 節で推定した量を用いる．

図 2 に学習コストと性能との関係を示す．なお，ゼロから事前学習を行ったモデルとの比較のため，継続事前学習の元モデルである Qwen2.5 およ

3）github.com/kumapo/JAQKET-dataset
4）github.com/EleutherAI/lm-evaluation-harness
5）github.com/sbintuitions/flexeval/



(a) 言語非依存ベンチマークにおける平均スコアと総学習コスト
(b) 日本文化理解ベンチマークにおける平均スコアと総学習コスト
(c) 日本文化理解ベンチマークにおける平均スコアと日本語文書の学習に費やしたコスト

図 2 平均スコアと各学習コスト．

び，日本語大規模言語モデルとして Sarashina2.2 と LLM-jp-3.1 の各サイズも性能を評価し，プロットしている．加えて，図 2 (a) と (c) については各点を元に回帰した直線を引いている．図 2 (a) に示されているように，言語非依存ベンチマークの日本語部分の平均スコアは学習の総コストに対数比例している．この結果は，継続事前学習による言語非依存ベンチマークでの対象言語の性能向上は，学習コーパスの多くが英語で占められている，ゼロからの事前学習による性能向上と比べて特段効率的ではないことを示唆している．

一方で，図 2 (b) は継続事前学習は日本文化理解ベンチマークにおいて，学習元のモデルと比較すると，総学習コストに対して高いスコアを達成している．この図からは総学習コストの観点では，継続事前学習は学習コストに対して効率的な手法であると考えられる．しかしながら，図 2 (c) によると，日本文化理解ベンチマークのスコアは日本語文書の学習に費やしたコストに対数比例している．すなわち，継続事前学習による性能向上は，日本語文書に対して新たに投入した学習コスト相当であると言える．言い換えれば，継続事前学習は対象言語に対するゼロからの事前学習を行った場合と比べて，特段効率的な手法ではないと言える．

4 結果を受けての議論

2 節と 3 節で得られた知見を元にすると，言語非依存タスクでの性能は総学習コストに依存するため，現状のような英語中心の大規模言語モデル開発が続いても継続的に上がっていくと考えられる．さらに，言語非依存タスクにおける各言語の性能は，英語のような学習データ中での主要な言語で大規模言語モデルに思考させることで，対象言語での学習を行うことなく性能向上できると報告されている [13]．従って，言語非依存タスクでの性能を向上させることのみが目的の場合には，英語以外のデータを大規模言語モデルの学習に取り入れる積極的な理由は薄い．この状況が続けば，事前学習コーパス中で主要でない言語に関連した文化理解について，大規模言語モデルは弱いままである可能性がある．

加えて，言語非依存ベンチマーク内の事例は真に言語や特定の文化に依存しない問題となっているか，という疑問もある．言語非依存ベンチマークを構築する際は，まず英語で論理的思考を問うデータセットを構築し，それを複数の言語に翻訳するという手続きが主要である．このため，データセットが暗黙的に英語や西欧文化の影響を受けている可能性はある．実際，MMLU データセットには *US History* や *US Law* のようなタスクが含まれており，アメリカ文化への偏りが報告されている [25]．

大規模言語モデルは様々なアプリケーションにおける根幹のツールとなりつつあり，文化的に強い偏りを含むことは望ましくない．そのような偏りを認識し，軽減するために，言語非依存ベンチマークだけではなく，多様な文化に依存したベンチマークを構築することが重要であろう．少なくとも，言語非依存ベンチマーク上での性能向上だけを議論する状況は脱していくべきであると考えている．

5 おわりに

本研究では，公開されている大規模言語モデルの事前学習コーパスは，大規模な Web コーパスである CommonCrawl と比べ，英語を多分に含む可能性を示した．加えて，継続事前学習されたモデルも含めていくつかのモデルを日本文化理解ベンチマークで評価することによって，継続事前学習はゼロからの事前学習と比較して特段効率的な手法ではなく，対象言語について投入した学習コスト相当の性能向上となっていることを示した．本研究で得られた知見が，英語中心の大規模言語モデル開発の次のステップとして，個別の言語をどのように扱うか考えていく一助になることを期待する．

参考文献

[1] Tom Brown, et al. Language models are few-shot learners. In **Advances in Neural Information Processing Systems 33 (NeurIPS)**, pp. 1877–1901, 2020.

[2] Hugo Touvron, et al. Llama: Open and efficient foundation language models. 2023.

[3] Jinze Bai, et al. Qwen technical report. 2023.

[4] Gemma Team, et al. Gemma: Open models based on gemini research and technology. 2024.

[5] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models. 2023.

[6] An Yang, et al. Qwen2.5 technical report. 2025.

[7] Gemma Team, et al. Gemma 3 technical report. 2025.

[8] Nicholas Carlini, et al. Extracting training data from large language models. In **30th USENIX Security Symposium (USENIX Security 21)**, pp. 2633–2650, 2021.

[9] Milad Nasr, et al. Scalable extraction of training data from aligned, production language models. In **The Thirteenth International Conference on Learning Representations (ICLR)**, 2025.

[10] Zhangchen Xu, et al. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In **The Thirteenth International Conference on Learning Representations (ICLR)**, 2025.

[11] Armand Joulin, et al. Fasttext.zip: Compressing text classification models. 2016.

[12] Edouard Grave, et al. Learning word vectors for 157 languages. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)**, 2018.

[13] Freda Shi, et al. Language models are multilingual chain-of-thought reasoners. In **The Eleventh International Conference on Learning Representations (ICLR)**, 2023.

[14] Wen Yang, et al. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. 2023.

[15] Ramon Pires, et al. Sabiá: Portuguese large language models. In **Intelligent Systems**, pp. 226–240, 2023.

[16] Jun Zhao, et al. Llama beyond english: An empirical study on language capability transfer. 2024.

[17] Yiming Cui, et al. Efficient and effective text encoding for chinese llama and alpaca. 2024.

[18] Kazuki Fujii, et al. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In **First Conference on Language Modeling (COLM)**, 2024.

[19] Aaron Grattafiori, et al. The llama 3 herd of models. 2024.

[20] Karl Cobbe, et al. Training verifiers to solve math word problems. 2021.

[21] Yubo Wang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 95266–95290, 2024.

[22] Weihao Xuan, et al. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1513–1532, 2025.

[23] Satoshi Sekine. Development of a question answering system focused on an encyclopedia (in japanese). In **9th Annual Meeting of the Association for Natural Language Processing**, 2003.

[24] Ziqi Yin, et al. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In **Proceedings of the Second Workshop on Social Influence in Conversations (SICon)**, pp. 9–35, 2024.

[25] Shivalika Singh, et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 18761–18799, 2025.

[26] Woosuk Kwon, et al. Efficient memory management for large language model serving with pagedattention. In **Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles**, 2023.

[27] LLM-jp. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. 2024.

表 3　CommonCrawl と各大規模言語モデルの生成した系列から推定した言語分布について，各言語のパーセンテージ．

Model	En	Ru	De	Fr	Es	Ja	Zh	It	Pt	Nl	others
CommonCrawl	43.0	9.8	6.4	5.1	5.1	4.2	4.1	2.8	2.3	1.9	15.3
Llama 2	89.7	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	9.1
Llama 3.1	92.3	0.4	0.3	0.2	0.3	1.8	2.6	0.1	0.1	0.1	1.7
Qwen 2.5	77.7	0.8	0.6	0.8	0.5	0.8	16.2	0.3	0.3	0.1	1.9
Qwen 3	73.4	0.6	0.3	0.3	0.6	0.4	22.8	0.1	0.5	0.0	1.0
Gemma 3	74.9	1.3	1.4	1.6	6.2	1.1	3.1	0.9	1.2	0.8	7.5
LLM-jp-3	63.4	0.0	0.0	0.0	0.0	36.4	0.1	0.0	0.0	0.0	0.0
Sarashina2.2	73.7	0.5	0.3	0.4	0.4	21.3	2.0	0.1	0.2	0.0	1.1

表 4　系列生成のハイパーパラメータ．

ハイパーパラメータ	値
Temperature	1.0
Top_p	1.0
Top_k	-1
max_tokens	1,000
stop	None
ignore_eos	True
repetition_penalty	1.0

A　系列生成のハイパーパラメータ

2 節における系列生成のためには vLLM [26] を使用し，デフォルトのハイパーパラメータを用いた．詳細な値は表 4 に示す．

B　LLM-jp モデルにおける推定した言語分布の相関

本研究で用いた，大規模言語モデルの生成系列を元に学習データの言語分布を推定する手法の妥当性を検証するために，Llama 2 モデルに加えて，LLM-jp の配布しているモデル [27] についてもケンドールの順位相関係数を計算する．LLM-jp は Llama 2 のように事前学習データの言語分布を公開してはいないが，学習データそのものは公開している．このため，公開されているデータの一部をサンプリングし，サンプリングしたデータの言語分布と LLM-jp の各モデルの生成した系列の言語分布との相関係数を計算する．

本文の実験で用いた LLM-jp-3.1-1.8B-Instruct4 と LLM-jp-3.1-13B-Instruct4 の事前学習モデルである，LLM-jp-3-1.8B と LLM-jp-3-13B に着目し，この学習に用いられた LLM-jp Corpus v3[6] を対象のデータとして用いる．事前学習データは極めて巨大であり，全体をまとめた後にサンプリングを行うことは難しい．事前学習データは各ドメインに分割されているため，まず，報告されている各ドメインのトークン数の割合に従って各ドメインからサンプリングを行い，それらのデータを結合した．さらに，言語分布を推定するために，この結合したデータから 100,000 文書をサンプリングした．

学習データと生成した系列の言語分布について，ケンドールの順位相関係数を計算したところ，LLM-jp-3-1.8B と LLM-jp-3-13B でそれぞれ 0.61 と 0.63 であった．これらの値は言語分布間に強い相関があることを示すため，生成した系列から言語分布を推定する，本研究で用いた手法は妥当であると考えられる．

C　推定した言語分布の詳細な値

サンプリングした CommonCrawl および，Llama 2 を除く各大規模言語モデルの生成した系列から推定した言語分布の詳細な値を表 3 に示す．表 3 における Llama 2 の行については論文内で報告されている値である [5]．系列の生成については，Llama-3.1-8B[7]，Qwen2.5-3B[8]，Qwen3-4B-Base[9]，Gemma-3-4B-pt[10]，LLM-jp-3-1.8B[11]，Sarashina2.2-3B[12] を用いた．なお，先述のとおり，LLM-jp-3-1.8B は LLM-jp-3.1-1.8B-Instruct4 の事前学習モデルである．

6）https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

7）huggingface.co/meta-llama/Llama-3.1-8B
8）huggingface.co/Qwen/Qwen2.5-3B
9）huggingface.co/Qwen/Qwen3-4B-Base
10）huggingface.co/google/gemma-3-4b-pt
11）huggingface.co/llm-jp/llm-jp-3-1.8b
12）huggingface.co/sbintuitions/sarashina2.2-3b