大規模言語モデルによる日本語スタイル変換の性能評価

花房 健太郎 [1]　柳本 大輝 [2]　梶原 智之 [2]　二宮 崇 [2]
[1] 愛媛大学工学部　[2] 愛媛大学大学院理工学研究科
hanafusa@ai.cs.ehime-u.ac.jp　yanamoto@ai.cs.ehime-u.ac.jp
kajiwara@cs.ehime-u.ac.jp　ninomiya.takashi.mk@ehime-u.ac.jp

概要

本研究では，日本語における大規模言語モデルのスタイル変換性能を広く調査する．まず，英語において研究されてきた 11 種類のスタイル変換を対象に，それぞれ 120 件の日本語文対を人手で作成し，日本語スタイル変換の評価用データセットを構築した．そして，12 種類の日本語大規模言語モデルを対象に，0-shot および 20-shot の文脈内学習によるスタイル変換の性能を評価した．評価実験の結果，指示チューニング済みのモデルは指示チューニング前のモデルよりもスタイル変換の性能が大幅に改善されること，パラメータ数の多いモデルの方が few-shot 文脈内学習によって性能がより改善されること，factual → romance および offensive → non-offensive のスタイル変換が難しいことが明らかになった．

1　はじめに

大規模言語モデル (LLM: Large Language Model) の文脈内学習 [1] は，多様な言語かつ幅広い自然言語処理タスクで活用され，注目を集めている．日本語に特化した LLM も，日本語・英語・ソースコードなどを用いて事前学習した LLM-jp [2] や英語 LLM に対して日本語データを用いて継続事前学習した Swallow [3] などが開発されている．これらの日本語 LLM の言語生成能力は，機械翻訳や自動要約のタスクでは評価されている [2–4] ものの，スタイル変換に関する日本語 LLM の評価は取り組まれていない．

スタイル変換 [5] は，所与のテキストの内容を保持しながら目的のスタイルへと言い換えるタスクである．このタスクでは通常，系列変換モデルの学習と評価のためにパラレルコーパスを用いる．英語においては，難解な文から平易な文への Simplicity に関するスタイル変換コーパスの Newsela [6]，フォーマルな文からインフォーマルな文への Formality に関するスタイル変換コーパスの GYAFC [7]，攻撃的な文から攻撃的ではない文への Toxicity に関するスタイル変換コーパスの ParaDetox [8] など，多様なスタイル変換のためのパラレルコーパスが利用できる．LLM のスタイル変換性能の評価 [9–11] も，英語においてはこれらのパラレルコーパスを用いて検証が進んでいる．一方で，日本語においてはテキスト平易化とも呼ばれる Simplicity に関するスタイル変換コーパス [12–15] しか存在せず，他のスタイルを対象とするスタイル変換パラレルコーパスは利用できない．そのため，日本語における LLM のスタイル変換性能の評価は充分に検証されていない．

表 1　本研究で扱うスタイルの一覧

Formality	formal ↔ informal
Gender	masculine ↔ feminine
Politeness	impolite → polite
Romance	factual ↔ romance
Sentiment	positive ↔ negative
Simplicity	complex → simple
Toxicity	offensive → non-offensive

本研究では，日本語における多様なスタイル変換の評価のために，11 種類のスタイル変換を対象に，それぞれ 120 文対ずつのパラレルコーパス[1)]を人手で構築する．表 1 に示すこれらの 11 種類は，スタイル変換タスクにおける英語の研究 [5] で主に扱われてきたスタイルである．本コーパスを用いて 12 種類の日本語 LLM のスタイル変換性能を評価した結果，指示チューニングによって性能が大幅に改善されること，パラメータ数の多いモデルの方が few-shot 文脈内学習によって性能がより改善されること，Romance および Toxicity に関するスタイル変換が難しいことが明らかになった．

2　評価用パラレルコーパス

日本語 LLM のスタイル変換性能を調査するために，評価用パラレルコーパス[1)]を人手で構築した．

1)　https://github.com/EhimeNLP/PASTEL-JP

2.1 コーパスの構築

本コーパスは，表 1 に示す 11 種類のスタイル変換について，それぞれ 120 文対で構成される．ただし，Simplicity のスタイル変換については，専門家によって書かれた高品質なパラレルコーパスである MATCHA[2] [15] が存在するため，MATCHA から無作為に選択した 120 文対を使用した．その他の 10 種類のスタイル変換は，日本語母語話者の大学生 5 名が，それぞれ 20 文対から 30 文対ずつを作文した．

2.2 コーパスの評価

構築したコーパスの例を表 2 に示す．本コーパスの品質を，本稿の著者が人手評価した．人手評価の基準には，機械翻訳の人手評価 [16] で用いられる文法性（5 点満点）および同義性（7 点満点）に加えて，スタイルの良さとして以下の 5 段階評価を用いた．

1. 文法性または同義性に著しい問題があるため，スタイルの良さを評価するのに値しない
2. 原文のスタイルがより強化されている
3. 原文とスタイルが全く変わってない
4. 僅かに目的のスタイルになっている
5. 目的のスタイルにふさわしい

文法性　評価の結果，文法性は全タスクにおいて 5 点満点中の 4.9 点以上であり，高品質であった．

同義性　Sentiment のスタイル変換は肯定と否定が入れ替わるため「矛盾する内容」として常に低評価だが，これは特に問題ではない．また，Romance および Toxicity も「情報の過不足」や「若干の齟齬」に該当するため同義性の評価はやや低いが，これらのタスクでは直接的な表現を間接的に言い換えるため，コーパスの品質に起因する原点ではなかった．その他は 7 点満点中の 6.7 点以上の高評価であった．

スタイル　スタイルについては，多くのタスクにおいて 5 点満点中の 4.7 点以上であり大きな問題はなかったが，Toxicity のスタイル変換は 4.2 点とやや低評価であった．これは，部分的には改善が見られたものの，文全体としては攻撃的な雰囲気が隠し切れていないという事例が含まれていたためである．

3 評価実験

本研究で構築した日本語スタイル変換コーパスを用いて，LLM のスタイル変換性能を評価した．

2） https://github.com/EhimeNLP/matcha

表 2　本研究で構築した日本語スタイル変換コーパスの例

Formality: formal ↔ informal
formal: これからますます暑くなりますね
informal: これからもっと暑くなるよね
Romance: factual ↔ romance
factual: 彼は私のことをじっと見つめていた
romance: 彼の瞳に私の姿がしばらく映し出されていた
Toxicity: offensive → non-offensive
offensive: 勝手なことをするな
non-offensive: 行動する前に一度相談してくださいね

3.1 実験設定

評価用コーパスは，スタイルごとに 120 文対で構成されている．このうち，無作為抽出した 20 文対を few-shot 文脈内学習に用いる事例とし，残りの 100 文対を評価に用いた．本実験では，LLM が持つスタイル変換能力を評価するため，LLM のパラメータ更新は行わず，0-shot または 20-shot の文脈内学習によってスタイル変換を実施した．

モデル　日本語 LLM として以下に示す 4 種類の合計 12 モデルを評価した．なお，言語モデリングの事前学習に加えて指示チューニング [17] を行ったモデルが利用可能な場合は，事前学習のみのモデルと指示チューニングモデルの両方を評価した．

日本語データで事前学習　日本語データを中心に，英語やソースコードも含めて事前学習した日本語 LLM として，LLM-jp[3)4)] [2] および Fugaku-LLM[5)6)] を用いた．

英語モデルを継続事前学習　英語 LLM[7)] [18] を日本語データ [19] で継続事前学習した日本語 LLM として，Swallow[8)9)10)11)12)13)] [3] を用いた．

3） https://huggingface.co/llm-jp/llm-jp-13b-v2.0
4） https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-dolly-ichikara_004_001_single-oasst-oasst2-v2.0
5） https://huggingface.co/Fugaku-LLM/Fugaku-LLM-13B
6） https://huggingface.co/Fugaku-LLM/Fugaku-LLM-13B-instruct
7） https://huggingface.co/meta-llama/Llama-2-7b
8） https://huggingface.co/tokyotech-llm/Swallow-7b-hf
9） https://huggingface.co/tokyotech-llm/Swallow-13b-hf
10） https://huggingface.co/tokyotech-llm/Swallow-70b-hf
11） https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-v0.1
12） https://huggingface.co/tokyotech-llm/Swallow-13b-instruct-v0.1
13） https://huggingface.co/tokyotech-llm/Swallow-70b-instruct-v0.1

表 3 BLEU による自動評価の結果（Instruct 列のチェックマークは指示チューニングモデル，0 列は 0-shot 文脈内学習，20 列は 20-shot 文脈内学習，Source 行は入力文のスタイル，Target 行は出力文のスタイルを表す．）

		Source:	formal		informal		masculine		feminine		impolite		factual	
		Target:	informal		formal		feminine		masculine		polite		romance	
	Size	Instruct	0	20	0	20	0	20	0	20	0	20	0	20
LLM-jp	13B	-	2.76	8.48	2.39	7.10	4.65	13.30	7.91	10.15	2.10	6.10	0.03	4.19
LLM-jp	13B	✓	17.10	36.81	21.90	45.96	45.25	76.64	44.39	76.98	18.74	37.87	8.88	12.21
Fugaku-LLM	13B	-	3.43	3.92	3.11	5.72	7.87	8.84	8.26	8.21	2.59	4.35	2.49	2.52
Fugaku-LLM	13B	✓	13.82	33.01	25.27	45.18	27.01	62.69	25.32	69.68	8.50	22.09	9.13	10.97
Swallow	7B	-	3.81	5.51	4.66	6.85	8.41	11.30	8.56	10.89	2.87	5.01	2.68	3.32
Swallow	7B	✓	5.32	4.33	7.91	5.93	10.26	9.17	10.01	8.68	5.49	4.05	4.55	2.85
Swallow	13B	-	4.14	5.58	4.25	6.85	9.28	10.07	9.00	9.74	3.34	5.69	2.47	3.69
Swallow	13B	✓	10.11	38.22	12.11	48.97	13.89	71.29	14.97	64.77	8.00	19.73	5.42	11.70
Swallow	70B	-	5.47	5.87	7.02	6.83	9.64	10.10	9.54	9.70	4.80	5.90	2.68	3.75
Swallow	70B	✓	2.66	45.83	4.16	40.17	4.69	70.87	4.62	76.85	2.54	17.60	2.20	13.87
EvoLLM-JP	7B	✓	19.33	43.87	26.64	46.05	47.29	73.34	54.51	74.23	23.53	29.76	8.42	10.53
EvoLLM-JP	10B	✓	14.54	35.57	30.48	44.35	46.33	71.43	40.63	67.57	21.83	26.99	9.52	9.65

		Source:	romance		positive		negative		complex		offensive	
		Target:	factual		negative		positive		simple		non-offensive	
	Size	Instruct	0	20	0	20	0	20	0	20	0	20
LLM-jp	13B	-	1.97	2.92	0.02	6.39	0.02	6.19	4.48	9.92	0.02	1.45
LLM-jp	13B	✓	11.94	15.94	21.82	42.75	18.79	45.85	13.77	39.68	5.49	4.25
Fugaku-LLM	13B	-	1.91	1.95	3.98	5.36	3.97	4.65	8.57	11.79	0.94	0.53
Fugaku-LLM	13B	✓	11.02	17.40	25.12	34.56	16.71	39.47	8.97	28.40	2.33	4.39
Swallow	7B	-	2.36	2.77	4.52	7.19	4.59	6.89	4.06	10.94	1.31	1.33
Swallow	7B	✓	1.56	29.72	6.92	5.38	4.03	4.83	7.40	10.16	1.66	1.41
Swallow	13B	-	2.19	3.20	5.24	6.95	5.17	6.67	6.81	12.66	1.38	1.75
Swallow	13B	✓	6.02	29.93	11.96	43.20	6.63	27.94	9.78	18.88	2.12	2.95
Swallow	70B	-	2.13	3.23	5.07	7.05	4.80	7.11	8.22	10.69	1.39	1.81
Swallow	70B	✓	4.88	29.42	2.60	46.93	2.25	28.00	5.37	34.29	1.25	5.25
EvoLLM-JP	7B	✓	9.96	24.64	28.47	53.80	13.38	48.84	25.61	43.40	3.80	7.40
EvoLLM-JP	10B	✓	9.16	25.30	26.39	49.48	10.92	47.44	26.20	38.65	2.96	5.11

複数 LLM のモデルマージ　その他の日本語 LLM として，日本語 LLM の Shisa[14] と数学に特化した英語 LLM の WizardMath[15] [20] および Abel[16] を融合した EvoLLM-JP[17][18] [21] を用いた．

評価　スタイル変換の性能は，BLEU [22] による自動評価と人手評価の両方によって評価した．人手評価用のデータは，タスクごとに 3 モデルを選択[19] し，それぞれ 100 文の合計 300 文ずつを用意した．そして，2.2 節の基準で 2.1 節の 5 名が評価した．

14） https://huggingface.co/augmxnt/shisa-gamma-7b-v1
15） https://huggingface.co/WizardLMTeam/WizardMath-7B-V1.1
16） https://huggingface.co/GAIR/Abel-7B-002
17） https://huggingface.co/SakanaAI/EvoLLM-JP-v1-7B
18） https://huggingface.co/SakanaAI/EvoLLM-JP-v1-10B
19） BLEU による自動評価の降順に 1, 3, 5 位のモデルを選択

3.2 実験結果

指示チューニングが有効　BLEU による自動評価の結果を表 3 に示す．Instruct 列に注目すると，多くのモデルとタスクの組み合わせにおいて指示チューニングモデルが高い性能を示し，LLM の指示チューニングがスタイル変換のために有効であることがわかる．ただし，Swallow-7B モデルの few-shot 設定および Swallow-70B モデルの 0-shot 設定では，指示チューニングによって性能が悪化した．

Few-shot 文脈内学習が有効　多くのモデルとタスクの組み合わせにおいて，0-shot よりも 20-shot の方が高い性能を示し，LLM の few-shot 文脈内学習がスタイル変換のために有効であることがわかる．ただし，Swallow-7B モデルは few-shot 文脈内学習に

表 4　人手評価の結果および各スタイルの特徴

	人手評価の平均値			BLEU とのピアソン相関			各スタイルの特徴	
	文法性	同義性	スタイル	文法性	同義性	スタイル	語彙サイズ	編集距離
Formality: formal → informal	4.90	6.32	4.14	0.140	0.352	0.373	658	9.74
Formality: formal ← informal	4.87	6.02	4.37	0.140	0.457	0.421		
Gender: masculine → feminine	4.91	6.79	4.62	0.138	0.378	0.373	542	3.13
Gender: masculine ← feminine	4.92	6.66	4.71	0.161	0.427	0.299		
Politeness: impolite → polite	4.54	5.12	4.07	0.268	0.538	0.418	530	12.31
Romance: factual → romance	4.47	3.66	3.64	0.074	0.199	0.081	845	25.20
Romance: factual ← romance	4.98	4.85	4.61	0.109	0.676	0.417		
Sentiment: positive → negative	4.68	5.07	4.66	0.314	0.642	0.369	629	7.28
Sentiment: positive ← negative	4.59	5.49	4.61	0.373	0.656	0.406		
Simplicity: complex → simple	4.26	4.65	3.13	0.021	0.170	0.041	1,093	17.68
Toxicity: offensive → non-offensive	4.74	3.13	2.80	0.118	0.442	0.284	550	20.86

よって性能が悪化した．また，事前学習のみのベースモデルに比べて，指示チューニングモデルの方が few-shot 文脈内学習の恩恵を大きく受けた．

難しいスタイル変換　Gender に関するスタイル変換は，LLM-jp や EvoLLM-JP が 0-shot でも 40 ポイント以上の BLEU を達成するなど，比較的簡単なタスクであることがわかる．一方で，Romance（factual → romance）や Toxicity に関するスタイル変換は，全てのモデルが 20 ポイント未満の BLEU に留まっており，比較的難しいタスクであると言える．

人手評価　人手評価の平均値を表 4 の左に示す．自動評価と同様に，Gender に関するスタイル変換は高評価であり，Romance（factual → romance）および Toxicity に関するスタイル変換は低評価であった．また，Simplicity に関するスタイル変換も，人手評価においては低評価であった．これは，説明の追加や省略によって原文の意味を完全に保持できない事例や，部分的には平易化できたものの文全体としては不充分という事例が含まれていたためである．

3.3　分析

BLEU と人手評価の相関　表 4 の中列に，BLEU による自動評価と人手評価のピアソン相関を示す．文法性は全体的に高いため，自動評価との相関を得るのは難しかった．同義性およびスタイルについては，Romance（factual → romance）および Simplicity を除いて，弱い相関が見られた．妥当な出力として多様な表現があり得る romance のスタイルと，読み手の知識や言語能力によって評価が変わり得る simple のスタイルは，評価が難しいと考えられる．

表 5　スタイルの特徴と人手評価のピアソン相関

	文法性	同義性	スタイル
語彙サイズ	−0.538	−0.355	−0.376
編集距離	−0.306	−0.829	−0.600

語彙サイズや編集距離と人手評価の相関　表 4 の右に，各スタイルの特徴を示す．語彙サイズの大きい Romance および Simplicity や，編集距離の大きい Romance および Toxicity は，実験結果から難しいスタイル変換であると結論付けたタスクである．また，表 5 に示すように，語彙サイズと文法性，編集距離と同義性およびスタイルは，よく相関する．つまり，各スタイルの統計情報から，LLM によるスタイル変換の性能をある程度見積もることができる．

スタイル変換をしない事例　LLM がスタイル変換の指示に従わず，入力文を発話と捉えて応答を出力することがあった．多くのスタイルにおいてはこのような応答は発生しないが，Simplicity において 2%，Romance において 1%，このような例が見られた．特に多かったのは Toxicity のスタイル変換であり，25% の出力が “なんですぐに連絡してこないの？ → 連絡が遅くなってごめんなさい。” のような応答であった．怒り口調の入力に対して，スタイル変換を行わず応答を出力する事例がよく見られた．

4　おわりに

本研究では，LLM による日本語スタイル変換を評価した．11 種類のスタイル変換を扱う評価用データを構築し，12 種類の LLM を評価した結果，指示チューニングや few-shot 文脈内学習の有効性，Romance や Toxicity の難しさが明らかになった．

謝辞

本研究は JSPS 科研費（若手研究, 課題番号：24K20840）の助成を受けたものです.

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 **Proc. of NeurIPS**, pp. 1877–1901, 2020.

[2] LLM-jp. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv:2407.03963**, 2024.

[3] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **CoLM**, 2024.

[4] 樽本空宙, 畠垣光希, 宮田莉奈, 梶原智之, 二宮崇. ChatGPT の日本語生成能力の評価. 自然言語処理, Vol. 31, No. 2, pp. 349–373, 2024.

[5] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep Learning for Text Style Transfer: A Survey. **CL**, Vol. 48, No. 1, pp. 155–205, 2022.

[6] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. **TACL**, Vol. 3, pp. 283–297, 2015.

[7] Sudha Rao and Joel Tetreault. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In **Proc. of NAACL**, pp. 129–140, 2018.

[8] Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. ParaDetox: Detoxification with Parallel Data. In **Proc. of ACL**, pp. 6804–6818, 2022.

[9] Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. BLESS: Benchmarking Large Language Models on Sentence Simplification. In **Proc. of EMNLP**, pp. 13291–13309, 2023.

[10] Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. Distilling Text Style Transfer With Self-Explanation From LLMs. In **Proc. of NAACL-SRW**, pp. 200–211, 2024.

[11] Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. Are Large Language Models Actually Good at Text Style Transfer? In **Proc. of INLG**, pp. 523–539, 2024.

[12] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In **Proc. of LREC**, pp. 1153–1160, 2018.

[13] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In **Proc. of LREC**, pp. 461–466, 2018.

[14] Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In **Proc. of TSAR**, pp. 179–187, 2022.

[15] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. MATCHA：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. 自然言語処理, Vol. 31, No. 2, pp. 590–609, 2024.

[16] Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. Is This Translation Error Critical?: Classification-Based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors. In **Proc. of HumEval**, pp. 46–55, 2021.

[17] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned Language Models Are Zero-Shot Learners. In **Proc. of ICLR**, 2022.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama2: Open Foundation and Fine-Tuned Chat Models. **arXiv:2307.09288**, 2023.

[19] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a Large Japanese Web Corpus for Large Language Models. In **CoLM**, 2024.

[20] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. **arXiv:2308.09583**, 2023.

[21] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary Optimization of Model Merging Recipes. **arXiv:2403.13187**, 2024.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proc. of ACL**, pp. 311–318, 2002.