日本語の包括的な指示追従性データセットの構築

堀尾海斗 [1]　福田創 [1]　小川隼斗 [1]　鈴江万碧 [1]　織田宥楽 [1]　河原大輔 [1,2]
関根聡 [2]　安藤まや [2]
[1] 早稲田大学　[2] 株式会社いちから
{kakakakakakaito@akane.,so.fukuda@akane.,cookie3120@ruri.}waseda.jp
{m.suzue@asagi.,urakuodacchi@akane.,dkw@}waseda.jp
{satoshi.sekine,maya.ando}@ichikara.ai

概要

大規模言語モデル (LLM) は急速な発展により、幅広い知識を保有し、多種多様な応答が可能になっている。LLM の知識や言語能力の評価には GLUE や MMLU などのデータセットが存在し、JGLUE などの日本語データセットも構築されている。LLM の生成評価の観点は、これら以外に、人間の指示に対して追従しているかという点がある。しかし、指示追従性を評価する為の日本語データセットは存在するものの、カバーする範囲が狭い。本研究では、日本語の包括的な指示追従性データセットを構築する。さらに、構築したデータセットを評価ベンチマークとして、既存の LLM の指示追従性を評価する実験を行う。構築したデータセットは ichikara-instruction2 データに含まれる形で提供予定である。

1　はじめに

大規模言語モデル (LLM) の言語理解能力を評価するデータセットには、常識や意味の理解を測るものとして GLUE [1]、様々な専門分野の知識を測るものとして MMLU [2] などが存在する。LLM の評価の観点は、このような知識の有無だけでなく、人間の指示に対して追従しているかという点もあり、指示追従性に関する英語のデータセットが存在する [3]。GLUE や MMLU の日本語データセットは既に開発されている [4, 5] が、日本語において指示追従性の一部を測るデータセットは存在する [6] ものの、包括的な評価は困難である。

本研究では、日本語において指示追従性を包括的に測ることができるデータセットを構築する。まず、日本語における指示追従性の分類を行い、その分類に基づくインストラクションと正解応答の作成を人手で行う。

構築したデータセットは LLM の学習データと評価ベンチマークとしての使用用途が考えられるが、本研究では評価ベンチマークとして 8 種類の LLM に対して指示追従性能を測る実験を行う。実験には構築したデータセットのインストラクションを使用し、LLM の応答に対してクラウドソーシングによる人手評価と強力な LLM による自動評価の 2 つの方法で定量的に評価する。

2　関連研究

英語において LLM の指示追従性を測るデータセットはいくつか存在しており、LLM がプロンプト内のインストラクションを理解して、応答に反映させることができるかを測るものである [3, 7, 8, 9, 10, 11]。これらは、包括的な指示追従性を評価したもの [3, 8, 9, 11] や、一般的な応答の制限 [7]、要約関連の制限 [10] などに分けられる。

日本語において LLM の指示追従性を測るデータセットとして LCTG Bench が構築されている [6]。この研究ではフォーマット、文字数、キーワード・NG ワードの 3 点での指示追従性を評価している。

本研究では、LCTG Bench の評価項目を含み、包括的に指示追従性を評価できるデータセットを構築する。

3　包括的な指示追従性データセットの構築

構築する指示追従性データセットの各要素は指示追従性を持ったインストラクションとその正解応答とする。

データセット構築の手順としては、まず、一般に LLM を使用する上で考えられる指示追従性の分類を行う。Zhou らの研究 [3] のインストラクション分類を参考に、指示追従性を分類する。本研究では、

表 1　指示追従性の分類

大分類	小分類	定義
長さ指定	なし	文字数や単語数、段落数など生成の長さを制限
書式指定	書式	箇条書きや表など、生成の書式を指定
	書式 (答えのみ)	回答のみの生成に指定する。自動的な説明の記述を制限
ルール指定	始め・終わり指定	生成の始めと終わりを指定
	ルール指定	執筆におけるルールの指定
	ルール指定 (非一般)	一般には行わないような”執筆におけるルール”の指定
	キーワード使用	特定のキーワードを含める生成を指定
	キーワード不使用	特定のキーワードを含めない生成を指定
	タスク指定	ことば遊びのような言葉を使ったタスクを指定
	スタイル指定	生成の形式や表現スタイルを指定
	ペルソナ・言語指定	ペルソナを与えた生成の指定。翻訳などの言語的の指定
	トピック指定	生成に対する観点や具体的なトピックの指定

表 2　構築した指示追従性データセットの例

大分類	小分類	インストラクション	正解応答
長さ指定	なし	サッカーのオフサイドのルールを 80 字前後で説明してください。	オフサイドとは、得点の為に味方をゴール前に待機させパスすることや、オフサイドポジションで相手の邪魔をしたりすることで取られる反則で、相手にフリーキックの権利が与えられます。
書式指定	書式 (答えのみ)	一週間は何時間ですか？答えだけを出力して下さい。	168 時間
ルール指定	ルール指定	次の文に読みやすくなるように読点を打った文を出力して下さい。「明後日は友達と池袋に行って服を買おうと思っていたが友達から熱が出たというメッセージが送られてきたのでその予定は延期することにした」	「明後日は友達と池袋に行って服を買おうと思っていたが、友達から熱が出たというメッセージが送られてきたので、その予定は延期することにした」
	タスク指定	しりとりって知ってる？例えば「ごりら」の次は「ラッパ」みたいに前の言葉の最後の文字から次の言葉を始める遊びー。それを一人で 5 回分やってよ。始めの言葉は「ルール」でスタート！	「ルール」「ルーレット」「とんぼ」「ぼたもち」「ちきゅう」

インストラクションを、生成の文字数や段落を指定する「長さ指定」、生成表現の形式を指定する「書式指定」、生成の内容に関する指定をする「ルール指定」の 3 つの大分類に分ける。さらに、「書式指定」と「ルール指定」に関しては小分類を作り、インストラクションを細分類する。作成した分類を表 1 に示す。

次に、表 1 の小分類ごとにインストラクションを人手で作成する。インストラクション作成の基準として、内容には固有名詞を含むことを許し、応答側の前提知識は考慮しない。また、応答のファイル形式を指定するインストラクションについては、指示追従可能な LLM が限られる為、本データセットでは採用しない。インストラクションは複数の分類に属するものもあるが、少なくとも分類される指示追従性を含む内容とする。

最後に、作られたインストラクションに対して人手で正解応答を作成する。正解応答は内容が適切で、かつ指示に追従できているものを作成する。正解応答は 2,000 字を上限とする。

データセットの規模は、2025 年 1 月時点でインストラクションと正解応答のペアが 200 件程度であり、2025 年 2 月までに 1,000 件を構築する予定である。現時点のデータセットにおけるインストラクションの平均長は約 44 字、正解応答は約 237 字である。

表 3 指示追従性の評価結果。/の左側が人手評価、右側が GPT-4o による評価

	長さ指定	書式指定	ルール指定	平均
正解応答	1.000 / 0.688	0.844 / 0.906	1.000 / 0.906	0.972 / 0.886
Gemini 1.5 Pro	0.938 / 0.875	0.938 / 0.906	0.883 / 0.953	0.898 / 0.938
Claude 3.5 Sonnet	**1.000 / 0.938**	**0.969 / 0.969**	**0.922 / 0.953**	**0.938 / 0.955**
GPT-4o	0.938 / 0.812	0.875 / 0.781	0.922 / 0.930	0.915 / 0.892
gemma-2-27b-it	1.000 / 0.875	0.781 / 0.875	0.805 / 0.852	0.818 / 0.858
Qwen2.5-32b-Instruct	0.938 / 0.875	0.906 / 0.844	0.828 / 0.812	0.852 / 0.824
Llama-3.1-Swallow-8B-Instruct-v0.2	0.750 / 0.625	0.812 / 0.656	0.688 / 0.656	0.716 / 0.653
calm3-22b-chat	0.812 / 0.938	0.844 / 0.625	0.734 / 0.773	0.761 / 0.761
llm-jp-3-13b-instruct	0.750 / 0.625	0.656 / 0.562	0.617 / 0.617	0.636 / 0.608

表 4 インストラクション『「右」について「東」「西」「南」「北」「心臓」「時計」「磁石」などの単語を使わずに説明して』に対する LLM の応答例と、人手と GPT-4o による評価結果

LLM	応答	人手	GPT-4o
Claude 3.5 Sonnet	「右」の説明： 人間が両手を前に出した時、親指が外側を向く方の手がある側を指します。また、多くの人が文字を書いたり、箸やスプーンを使ったりする際に主に使う手がある方向です。 人が前を向いて立っている時、体の中心線から見て、この方向は常に一定です。例えば、人が後ろを向いても、その人にとっての「右」は変わりません。 また、運転席から見て、アクセルペダルがある側の方向でもあります。 多くの文化圏で、この方向は「良い」「正しい」「優れている」といった肯定的な意味も持ち合わせています。	✓	✓
gemma-2-27b-it	それは、あなたの体の左側と反対側です。 もしあなたが両手を前に伸ばすと、右の手がその方向を指しています。	×	✓
llm-jp-3-13b-instruct	「右」とは、一般的に人間の体の右側を指します。右手や右足など、体の右側の部分を指す際に使われます。また、方向としては、地図や方位磁石で北を向いたときに右側に位置する方向を指します。例えば、北半球では東が右側になります。	×	×

4 LLM の指示追従性の評価実験

4.1 実験設定

構築したデータセットを指示追従性の評価ベンチマークとして、8 種類の LLM に対して評価実験を行う。

LLM と生成設定 LLM は、Gemini 1.5 Pro、Claude 3.5 Sonnet、GPT-4o、gemma-2-27b-it、Qwen2.5-32b-Instruct、Llama-3.1-Swallow-8B-Instruct-v02、calm3-22b-chat、llm-jp-3-13b-instruct を使用する。LLM にはインストラクションを与え、それに対する応答を得る。応答の生成は温度 0、最大トークン長 4,096 で行う。

評価方法 LLM の応答の評価にはクラウドソーシングによる人手評価と LLM による自動評価を使用する。構築したデータセット中の 176 問のインストラクションを評価データとして用いる。176 問の内訳はおおよそ全ての小分類が同程度含まれるように用意する。また、人手と LLM による評価の精度確認の為、インストラクション 176 問に対する正解応答も LLM の応答と同様に評価する。

人手による評価 Yahoo!クラウドソーシングを用いて LLM の応答を人手で評価する。クラウドワーカには、インストラクションと LLM の応答を与え、「指示に従っている」と「指示に従っていない」の 2 択を回答するタスクを依頼する。1 問あたり 5 人のワーカに回答を依頼し、3 人以上が「指示に従っている」を選択した場合、その LLM の応答は指示に

表 5 インストラクション『「いっぱい」の「い」を「す」に変えてください』に対する正解応答と、人手および GPT-4o の評価では指示追従の判定であるが指示追従していない LLM の応答例

正解応答	「いっぱい」の「い」を「す」に変えるのですね。 **「すっぱす」！** いかがでしょうか。「いか」だと「すか」、「せんせい」は「せんせす」になりますね。 言葉遊びってとても楽しいですね。ほかにも変えたい言葉がありますか。
Claude 3.5 Sonnet	「すっぱい」
gemma-2-27b-it	「いっぱい」の「い」を「す」に変えると「**すっぱい**」になります。
Qwen2.5-32b-Instruct	「いっぱい」の「い」を「す」に変えると「すっぱい」になります。ただし、この単語は「いっぱい」の意味とは異なり、「酸っぱい」や「辛味が強い」などの意味を持ちます。文脈によって適切な使用方法が異なるので、使用する際は注意が必要です。

表 6 GPT-4o の評価では指示追従ではない判定であるが指示追従している正解応答例

インストラクション	正解応答
100 字以内で恐竜について説明してください。	恐竜は、中生代に栄していた陸上に住む爬虫類です。 恐竜が出現したのは今から 2 億 3000 年前の三畳紀で、約 1 億 6000 万年間繁栄していましたが、今から約 6500 万年前の白亜紀末で絶滅しました。
プールで気をつけないといけないことを 30 文字以内で記述してください。	プールはルールを守り体調が優れない時は利用を控えましょう。

追従していると判定する。明らかに指示追従、指示追従でない問題をチェック設問として用意し、ワーカには応答の評価と同時に、チェック設問にも回答してもらう。チェック設問に正解できなかったワーカの判定は用いない。

LLM による自動評価 GPT-4o を用いて LLM の応答を自動評価する。GPT-4o に対して、インストラクションに対して応答が指示に追従できているか判断する旨の説明、評価するインストラクションとその LLM の応答を提示して「yes」か「no」で判断させる。評価時には、GPT-4o に判断の理由も同時に生成させる。GPT-4o による自動評価の際にも応答生成時と同様、温度 0、最大トークン長 4,096 で行う。

4.2 実験結果

人手と GPT-4o による評価の結果を表 3 に、LLM による応答例を表 4 に示す。実験の結果、既存の LLM の中では Claude 3.5 Sonnet が長さ指定、書式指定、ルール指定の 3 分類全てで最も高い精度であった。

人手と GPT-4o による評価では指示追従の判定であるが指示追従していない LLM の応答例を表 5 に示す。文字を置換するという指示は計算機にとっては比較的易しい指示と考えられるが、複数の LLM で指示に追従していない応答をしている。人手の評価も指示追従の判定になっていることから、ひっかけ問題に近く、人間と同様に LLM も誤った応答をしてしまうと考えられる。

人手評価において指示追従ではないと判定された正解応答が少数発見されたが、ほとんどの場合、正解応答が指示に追従していなかった。このことから人手評価は信頼できると考える。該当する正解応答に関しては今後修正する予定である。

GPT-4o による正解応答の評価において、長さ指定の評価値が他の 2 つの分類に比べて低い。長さ指定における GPT-4o の評価では指示追従ではないと判定されているが指示追従している正解応答の例を表 6 に示す。これらの正解応答はインストラクションで指定された文字数を満たしているが、指示追従でない判定になっている。このことから文字数の計数に関しては、人間には出来るが、強力な LLM でも未だに難しいと考えられる。

5 おわりに

本研究では、日本語の包括的な指示追従性データセットを構築した。また、構築したデータセットを用いて既存の LLM を評価したところ、Claude 3.5 Sonnet が最も指示追従性が高いという結果になった。展望としては、データセット規模のさらなる拡大や、学習データとして本データセットを用いた場合の LLM の指示追従性能の改善などが挙げられる。

謝辞

指示追従性データセットの構築に貢献していただいたアノテータの方々に感謝する。

参考文献

[1] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[2] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **International Conference on Learning Representations**, 2021.

[3] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv, 2023. abs/2311.07911.

[4] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[5] Ziqi Yin, Hao Wang, Kaito Horio, Daisuike Kawahara, and Satoshi Sekine. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In James Hale, Kushal Chawla, and Muskan Garg, editors, **Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)**, pp. 9–35, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] 栗原健太郎, 三田雅人, 張培楠, 佐々木翔太, 石川亮介, 岡崎直観. Lctg bench: 日本語 llm の制御性ベンチマークの構築. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[7] Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. Instruction-following evaluation through verbalizer manipulation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 3678–3692, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. **ACM Comput. Surv.**, Vol. 56, No. 3, October 2023.

[9] Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3155–3168, Singapore, December 2023. Association for Computational Linguistics.

[10] Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 4481–4501, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[11] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4667–4688, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

.