

# llm-jp-eval: 日本語大規模 言語モデルの自動評価ツール

Namgi Han, 植田 暢大, 大嶽 匡俊, 勝又 智, 鎌田 啓輔, 清丸 寛一, 児玉 貴志, 菅原 朔, Bowen Chen, 松田 寛, 宮尾 祐介, 村脇 有吾, 劉 弘毅

2024.03.13. 言語処理学会第30回年次大会@神戸

# はじめに：日本語大規模言語モデルとその評価

- 2023年だけで日本語大規模言語モデルが2桁以上発表される中、日本語大規模言語モデルの性能評価の重要性が増している
- 海外では大規模言語モデルに対する評価ベンチマークが充実している
  - Big-Bench (>200タスク)、OpenLLM (>60タスク)、...
- しかし日本語の評価ベンチマークは少ない
  - JGLUE (5タスク)、JP Language Model Evaluation Harness (12タスク)、...



# 大規模言語モデルの評価：海外

- 英語・中国語では40以上の評価ベンチマークが存在<sup>1</sup>
- 評価対象
  - 自然言語推論など、伝統的な自然言語処理のタスク
  - 自動翻訳やコード生成などの生成問題
  - 社会的バイアスや信頼性などの安全性検証
  - 社会科学・理工系のドメイン特化知識
  - 医療・個人化アプリケーションなどの応用能力
  - マルチモーダル能力
- 現在進行形で評価ベンチマークは増え続けている



**Hugging Face**



1) Yupeng Chang et al., A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109, 2023

# 大規模言語モデルの評価：日本（1）

- JGLUE<sup>2</sup>
  - GLUE [3]の日本語バージョン
  - 5つの評価データセットで構成され、評価対象となるタスクが少ない
  - llm-jp-evalはJGLUEの一部を取り込み、13個の評価データセットに対応
- JP Language Model Evaluation Harness<sup>3</sup>
  - JGLUEに加え、海外の多言語データセットも含めた7つのデータセットを追加で対応
  - 一部の評価で、生成結果ではなく出力ラベルの対数尤度を使う
  - llm-jp-evalは全ての評価を生成結果に基づいて行う

2) Kentaro Kurihara et al., JGLUE: Japanese general language understanding evaluation. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

3) <https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable>

# 大規模言語モデルの評価：日本（2）

- Nejumi（Neo）リーダーボード<sup>4</sup>
  - 純粹に生成結果だけで評価を行う
  - 初期はJGLUEだけをサポートしたが、以降、llm-jp-evalを取り込みつつJapanese MT-Benchを追加で対応
- LLM Judge系
  - Japanese MT-Bench<sup>5</sup>、Japanese Vicuna QA<sup>6</sup>、Rakuda<sup>7</sup>など
  - オープンクエスチョンに対するLLMの応答をGPT-4に評価させる
  - 評価を特定の言語モデルに判断させるところで、評価データセットにアノテーションされた正解と比較するllm-jp-evalと違う

4) <https://wandb.me/nejumi>

5) [https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm\\_judge](https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge)

6) <https://github.com/ku-nlp/ja-vicuna-qa-benchmark>

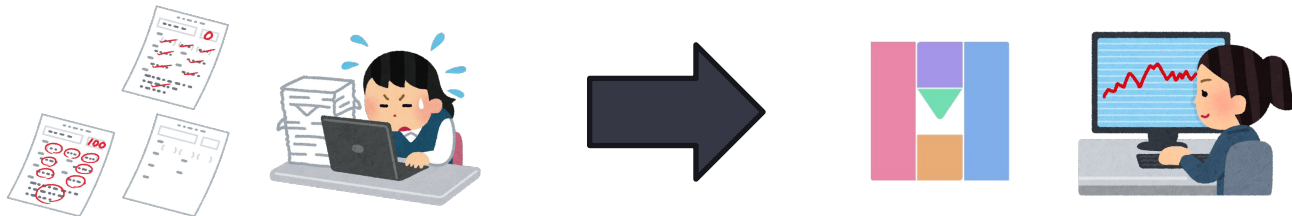
7) <https://yuzuai.jp/benchmark>

# 大規模言語モデルの評価：日本（3）

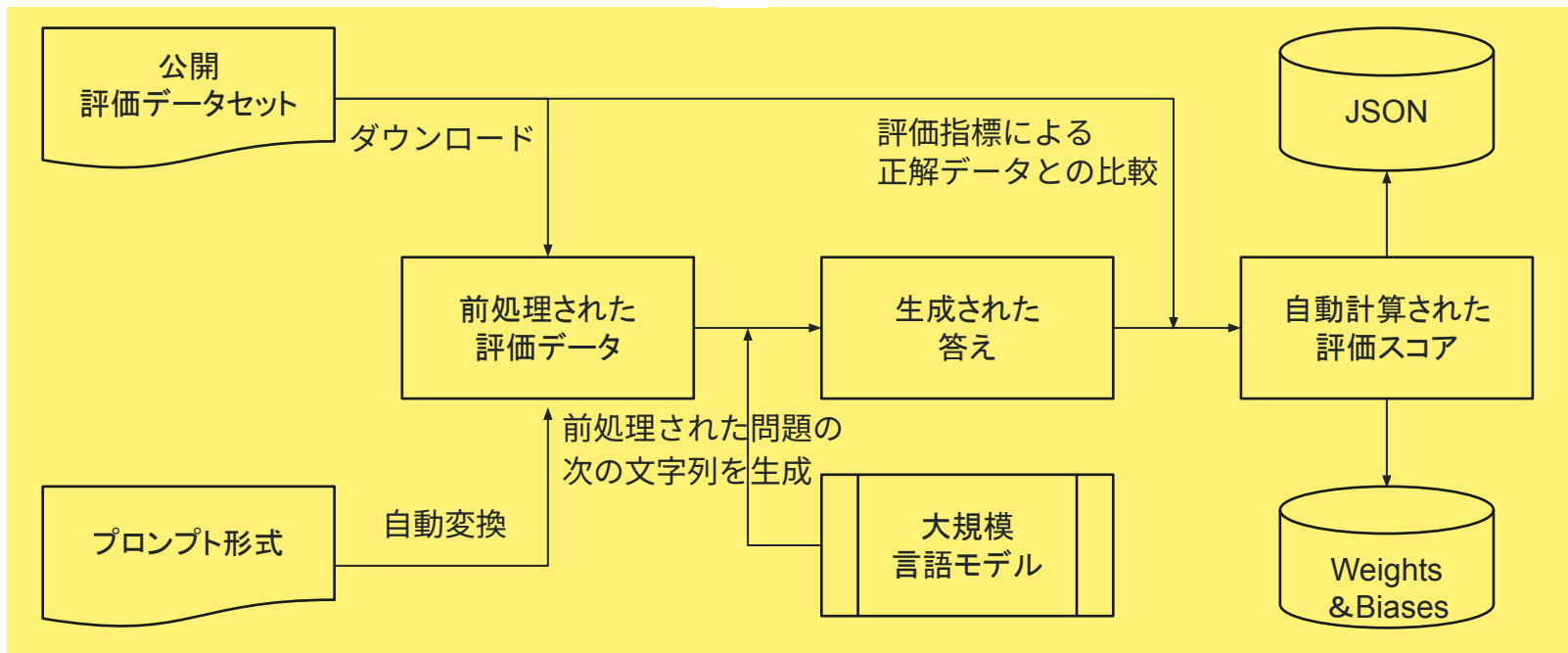
名前	データセット数	問題件数	評価手法
JGLUE	5	322,230	分類器
JP Language Model Evaluation Harness	12	416,764	対数尤度の比較・生成
Nejumi リーダーボード	5	322,230	生成
Nejumi Neo リーダーボード	14	179,036	生成・LLM as a judge
Japanese MT-Bench	8	80	LLM as a judge
Japanese Vicuna QA	10	80	LLM as a judge
Rakuda	4	40	LLM as a judge
<b>llm-jp-eval</b>	<b>13</b>	<b>178,956</b>	<b>生成</b>

# llm-jp-evalの紹介

- Apache License 2.0で公開中：<https://github.com/llm-jp/llm-jp-eval>
- 評価データセットのダウンロード、前処理、評価を自動化
  - 二つのスクリプトを実行するだけで全フレームワークを使用可能
  - 評価結果はJSON・W&Bで管理可能
- 一つの設定ファイルで評価設定を調整可能
- スクリプト・評価データセットは商用利用可能なライセンス

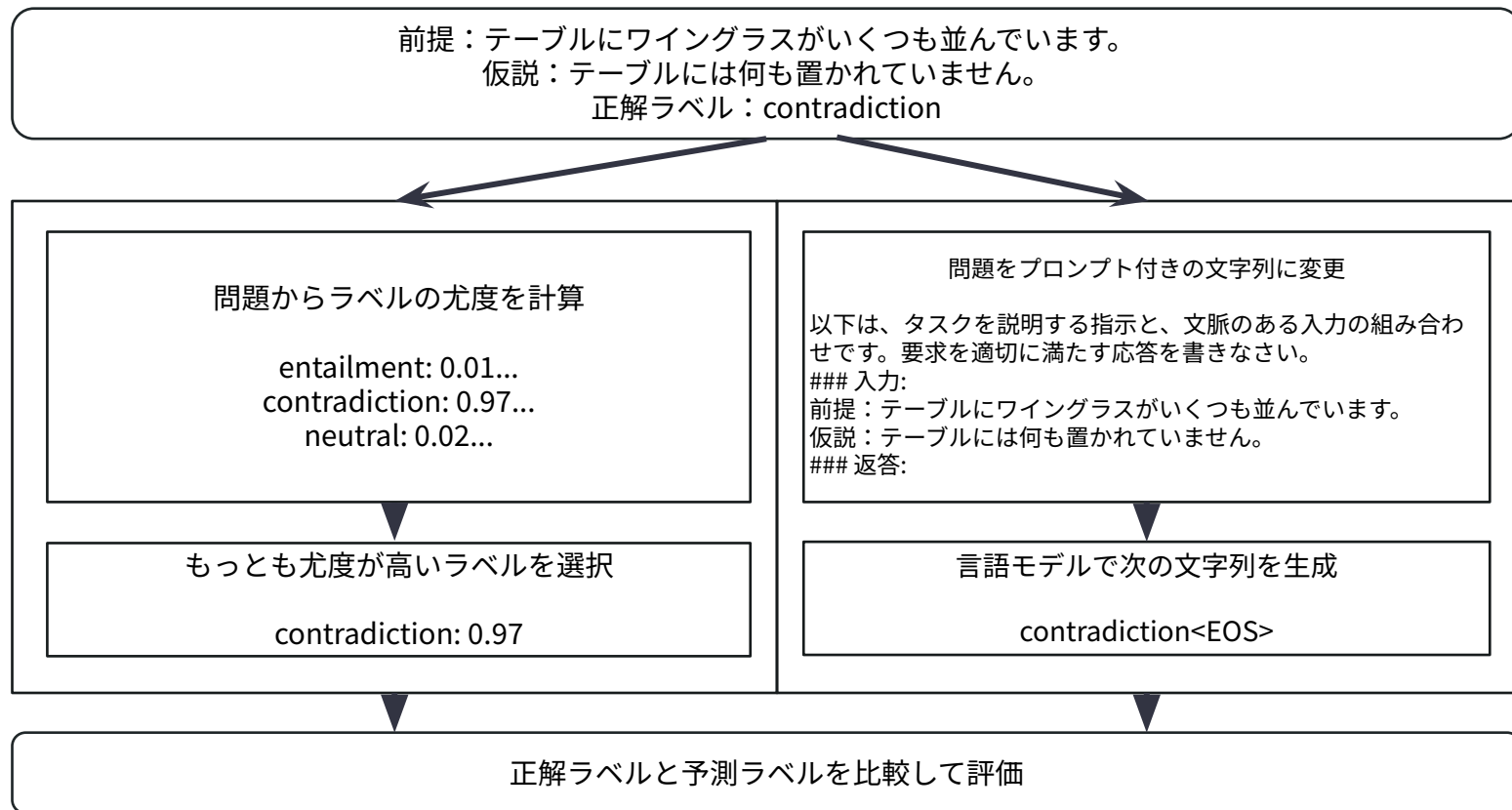


# llm-jp-evalの評価フレームワーク





# 既存の機械学習モデルとllm-jp-evalの評価における比較



# llm-jp-evalの対応データセット（1）

カテゴリー	データセット	ライセンス	評価指標
Natural Language Inference (NLI)	JAMP	CC BY-SA 4.0	Exact Match
	JaNLI	CC BY-SA 4.0	Exact Match
	JNLI	CC BY-SA 4.0	Exact Match
	JSeM	BSD 3-Clause	Exact Match
	JSICK	CC BY-SA 4.0	Exact Match
Question Answering (QA)	JEMHopQA	CC BY-SA 4.0	Char. F1
	NIILC	CC BY-SA 4.0	Char. F1
Reading Comprehension (RC)	JSQuAD	CC BY-SA 4.0	Char. F1
Multiple Choice question answering (MC)	JCommonsenseQA	CC BY-SA 4.0	Exact Match

# llm-jp-evalの対応データセット（2）

- 灰色は今後のアップデートで対応予定

カテゴリ	データセット	ライセンス	評価指標
Entity Linking (EL)	chABSA	CC BY 4.0	Set F1
Fundamental Analysis (FA)	Wikipedia Annotated Corpus	CC BY-SA 4.0	Set F1
Mathematical Reasoning (MR)	MAWPS	Apache-2.0	Exact Match
Semantic Textual Similarity (STS)	JSTS	CC BY-SA 4.0	Pearson/Spearman Coef.
Language Modeling (LM)	JBLiMP	調整中	Exact Match
	JCoLA	調整中	Exact Match
Human Examination (HE)	MMLU (en)	MIT License	Exact Match
	JMMLU	CC BY-SA 4.0	Exact Match

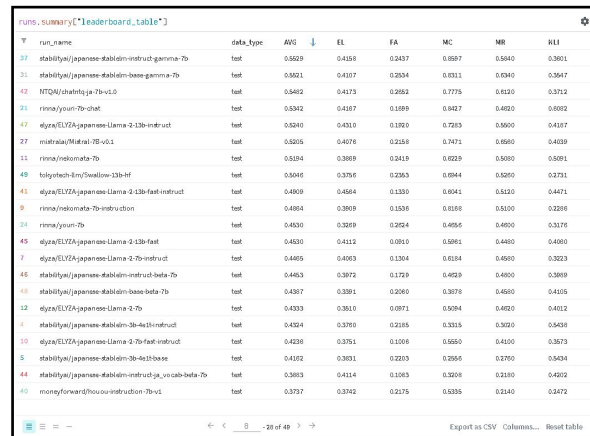
# llm-jp-evalによる評価例

- 評価実験の設定
  - 生成のハイパーパラメータ：HuggigFace Transformersの初期値
  - プロンプト：Alpacaのプロンプト形式
  - 4-shotsでの評価
  - AVGスコアの計算にSTSのスコアは含めない

## - 今回共有する検証内容

- Q1. パラメータの数と言語モデルの性能は正比例するか？
- Q2. 日本語のコーパスを使った継続訓練は有効か？

- 評価結果の詳細：<https://wandb.me/llm-jp-leaderboard>



Y	run_name	data_type	AVG ↓	EL	FA	MC	MR	RLI
37	stability/japanese-#abllm-instruct-gamma-7b	test	0.5029	0.4158	0.2437	0.6897	0.6840	0.3601
31	stability/japanese-#abllm-base-gamma-7b	test	0.5021	0.4107	0.2534	0.6311	0.6340	0.3647
12	NTQA/shahjpa-7b-v1.0	test	0.5482	0.4173	0.2561	0.7775	0.6120	0.3712
21	rinna/joini-7b-chat	test	0.5342	0.4187	0.1889	0.8417	0.4920	0.6052
17	elya/ELYA-japanese-Llama-2-13b-instruct	test	0.5240	0.4310	0.1920	0.7313	0.5600	0.4187
27	instralab/Mistral-7B-v1.0	test	0.5205	0.4078	0.2158	0.7471	0.5660	0.4339
11	rinna/indomata-7b	test	0.5194	0.3939	0.2419	0.6229	0.5980	0.5091
49	holgotech-llm/swallow-13b-hf	test	0.5046	0.3756	0.2263	0.6944	0.5260	0.3731
41	elya/ELYA-japanese-Llama-2-13b-fast-instruct	test	0.4909	0.4664	0.1230	0.6041	0.5120	0.4421
10	rinna/indomata-7b-instruct0m	test	0.4804	0.3809	0.1536	0.6168	0.5100	0.2258
24	rinna/joini-7b	test	0.4830	0.3229	0.2424	0.4956	0.4800	0.5178
45	elya/ELYA-japanese-Llama-2-13b-fast	test	0.4830	0.4112	0.0910	0.5961	0.4480	0.4460
17	elya/ELYA-japanese-Llama-2-7b-instruct	test	0.4465	0.4063	0.1304	0.6164	0.4680	0.3223
16	stability/japanese-#abllm-instruct-beta-7b	test	0.4463	0.3873	0.1729	0.4929	0.4800	0.3889
18	stability/japanese-#abllm-base-beta-7b	test	0.4317	0.3391	0.2060	0.3878	0.4680	0.4105
12	elya/ELYA-japanese-Llama-2-7b	test	0.4333	0.3610	0.0971	0.5004	0.4920	0.4012
11	stability/japanese-#abllm-30-4s1t-instruct	test	0.4324	0.3780	0.2185	0.3315	0.3020	0.5438
10	elya/ELYA-japanese-Llama-2-7b-fast-instruct	test	0.4238	0.3751	0.1006	0.5660	0.4100	0.3673
5	stability/japanese-#abllm-30-4s1t-base	test	0.4182	0.3031	0.2283	0.2666	0.2760	0.5434
14	stability/japanese-#abllm-instruct_4s_1t_cab-beta-7b	test	0.3863	0.4114	0.1063	0.5008	0.3180	0.4202
10	moneyforward/houou-instruction-7b-v1	test	0.3737	0.3742	0.2175	0.5335	0.3140	0.3472

# Q1. パラメータの数と言語モデルの性能は比例するか？

- A. パラメータが大きいほど、評価スコアも上がる傾向
  - 特にQA・RCでその傾向が強い
  - ただ全てのタスクに対して同じ傾向があるわけではない

モデル名	パラメータ	AVG	NLI	QA	RC	MC	EL	FA	MR
cyberagent/open-calm-1b	<b>1B</b>	<b>0.148</b>	0.269	<b>0.213</b>	<b>0.222</b>	0.217	0.087	0.023	0.006
cyberagent/open-calm-3b	<b>3B</b>	<b>0.204</b>	0.368	<b>0.258</b>	<b>0.418</b>	0.203	0.147	0.029	0.008
cyberagent/open-calm-7b	<b>7B</b>	<b>0.224</b>	0.256	<b>0.366</b>	<b>0.564</b>	0.198	0.159	0.015	0.008
llm-jp/llm-jp-1.3b-v1.0	<b>1.3B</b>	<b>0.253</b>	0.310	<b>0.304</b>	<b>0.557</b>	0.205	0.304	0.072	0.018
llm-jp/llm-jp-13b-v1.0	<b>13B</b>	<b>0.343</b>	0.349	<b>0.468</b>	<b>0.721</b>	0.206	0.340	0.189	0.130

## Q2. 日本語のコーパスを使った継続訓練は有効か？

- A. 日本語のコーパスを使う継続訓練は有効
  - Llama-2-7b-hf、mistralai/Mistral-7B-v0.1とそれらで継続訓練を行ったLLMの評価例
  - 前のスライドと同じく、QA・RCで評価スコアが向上する傾向

モデル名	AVG	NLI	QA	RC	MC	EL	FA	MR
meta-llama/Llama-2-7b-hf	<b>0.351</b>	0.363	<b>0.346</b>	<b>0.750</b>	0.246	0.329	0.118	0.304
→ tokyotech-llm/Swallow-7b-hf	<b>0.415</b>	0.318	<b>0.494</b>	<b>0.806</b>	0.368	0.327	0.214	0.374
→ elyza/ELYZA-japanese-Llama-2-7b	<b>0.433</b>	0.401	<b>0.421</b>	<b>0.791</b>	0.509	0.351	0.097	0.462
→ stabilityai/japanese-stablelm-base-beta-7b	<b>0.439</b>	0.411	<b>0.450</b>	<b>0.820</b>	0.388	0.339	0.206	0.458
mistralai/Mistral-7B-v0.1	<b>0.521</b>	0.404	<b>0.355</b>	<b>0.858</b>	0.747	0.408	0.216	0.656
→ stabilityai/japanese-stablelm-base-gamma-7b	<b>0.552</b>	0.355	<b>0.501</b>	<b>0.880</b>	0.831	0.411	0.253	0.634

# おわりに

- 既存の日本語評価データセットを活用し、それらを全て生成問題と見なすことで、日本語大規模言語モデルの性能を評価するフレームワークを提案
- 日本語の評価ベンチマーク構築はまだ課題が多い
  - 評価データセットの数、評価対象の種類が外国の環境に比べて少ない
  - 新たなデータセットの開発も続けつつ、海外のデータセットの翻訳も検討すべき
- llm-jp-evalの今後の課題
  - 現在対応できてないタスク・データセットをサポート：生成問題、安全性検証、…
  - 評価スコアに対する分析：本当に良い言語モデルとは？

