

# 評価・チューニングWG 研究成果報告

評価・チューニングWG 幹事

宮尾 祐介 (東京大学)

村脇 有吾 (京都大学)

○菅原 朔 (国立情報学研究所)

2025/11/26 Japanese Symposium on Open LLMs

## LLM の評価・チューニングにかかわる研究（8件）

【評価】 ①～④⑦⑧

- LLM はどのような能力を持つか？
- どのように評価するべきか？

【チューニング】 ⑤⑥

- LLM は訓練によりどのような振る舞いを見せるか？
- どのように訓練すべきか？



# 日本語を対象とした LLM の大規模人手評価（井之上+）

- **背景:** LLMの品質自動評価には研究の余地あり
  - 人手評価との乖離 (Shen+2023)、評価プロンプトの僅かな違いに敏感 (Wang+2024)
- **本研究:** LLMを大規模人手評価。結果も公開。自動評価の研究にもぜひ！
  - 評価対象: 484件の日本語プロンプト (既存ベンチマークより抽出)、10のLLM
  - 評価方法: 5つの評価軸、5段階評価
  - 作業者: 53名 (プロ、学部生、大学院生) + 外注、1事例あたり2人以上

**表 2** 評価対象とした LLM の一覧.

名前	サイズ
<b>日本語非特化モデル:</b>	
Llama 3	8B
GPT-3.5 (turbo-1106)	-
GPT-4 (0613)	-
Claude 3.5 Sonnet (20240620-v1:0)	-
<b>日本語特化モデル:</b>	
Swallow v0.1 (Llama 3 ベース)	8B
Swallow v0.1 (Llama 3 ベース)	70B
ELYZA LLM for JP (Llama 3 ベース)	8B
llm-jp v2.0 (16x)	13B
llm-jp v2.0	13B
CyberAgentLM3 (CALM3)	22B

**表 3** 評価項目の一覧.

項目	基準
関連性	プロンプトに対する応答として成立しているか。応答がプロンプトの要求（文字数制限、箇条書き等）を満たしているか。
流暢性	文章が日本語として正しいか（文法、ことばの使い方等）、構成などが整っているか、自然で読みやすいか。
正確性	情報が事実として正しいか。
詳細性	情報量が多いか。
総合評価	上記 4 つの評価を踏まえた総合的な評価。

1

# アノテーションの結果と分析



プロンプト	Llama 3 8B	CALM3 22B
富士山は日本で最も高い山ですが、それに次ぐ高さを誇る山二つを挙げ、それぞれの所在地を記述してください。	富士山は日本で最も高い山で、標高は3,776メートルです。次いで高い山は、北海道の得能岳（標高2,291メートル）と、長野県の三頭山（標高2,893メートル）です。 → 関連性:5 / 流暢性:5 / 詳細性:3 正確性:1 / 総合評価:3	日本で富士山に次ぐ高さを誇る山は、北岳と穂高岳です。北岳は山梨県に位置し、南アルプス山脈の一部を成しています。一方、穂高岳は長野県と岐阜県の境にある北アルプス山脈に位置しています。北岳は標高3,193メートルで、日本第二の高さを誇り、その美しい山容と豊かな自然で知られています。穂高岳は複数の峰から成り、中でも奥穂高岳が標高3,190メートルで北岳に次いで日本で三番目に高い山です。これらの山々は、… → 関連性:5 / 流暢性:5 / 詳細性:4 正確性:5 / 総合評価:4
あめが18個あります。6人で分けます。1人分は18 ÷ 6 = 3個になります！1人分は何個になりますか？	簡単な問題ですが、解く楽しみです！18個を6人で分ける場合、1人分は18 ÷ 6 = 3個になります！ → 関連性:4 / 流暢性:2 / 詳細性:4 正確性:5 / 総合評価:4  [Swallow 8B v0.1] 18個のあめを6人で分ける場合、1人分は18個を6人で割った数になります。18 ÷ 6 = 3 ですから、1人分は3個のあめになります。 → 関連性:5 / 流暢性:4 / 詳細性:5 正確性:5 / 総合評価:5	18個のあめを6人で分けると、1人あたり何個になるかを計算します。まず、18を6で割ります。18 ÷ 6 = 3 となります。したがって、1人あたり3個のあめをもらうことになります。 → 関連性:5 / 流暢性:5 / 詳細性:5 正確性:5 / 総合評価:5

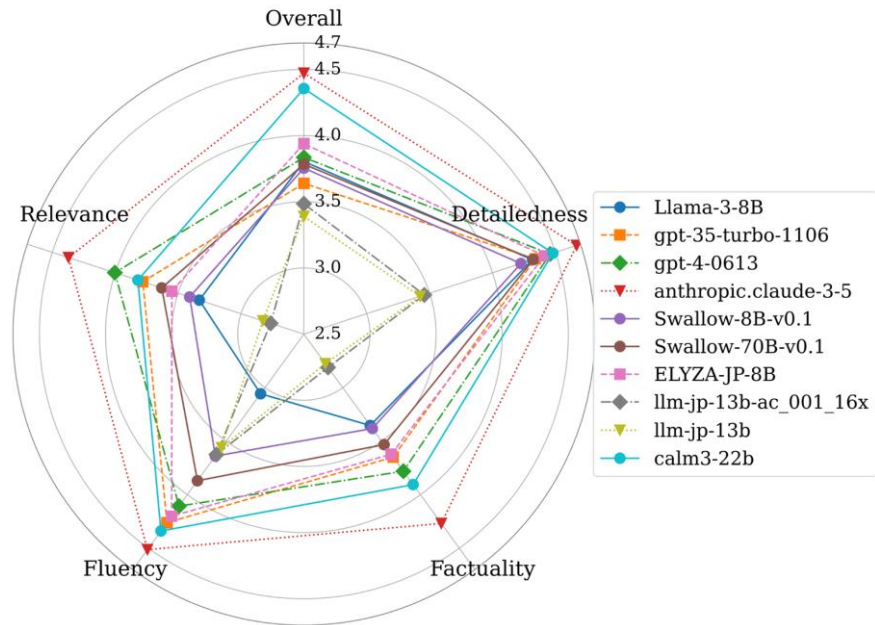


図1 484件の日本語プロンプトについて、10種類のLLMの応答を5つの評価項目で人手により評価した。

# ②

## LLMの日本語におけるコンテキスト追従能力の評価 (滝沢+)

YANS 2025 奨励賞

- **概要:** 質問者に関する情報やスタイルの指定といったコンテキストをLLMが理解し回答に反映できるかを評価
- **データ作成:** llm-jp-instructionsデータセット [1]から主観的な問題191問を抽出しコンテキストを追加
  - Audience制約 (質問者に関する情報) はgpt-4oで問題ごとに作成
  - Style制約では定型の一文を追加
- **評価:** 「既存の問題に対する回答」と「コンテキスト付き問題に対する回答」を比較し、どちらがよりコンテキストを反映した回答か LLM-as-a-Judge(gpt-4o)で4段階評価
  - (1)コンテキスト回答がより反映、(2)Default回答がより反映、(3)どちらも同じ程度に反映している、(4)どちらも反映していない
  - Simplicity (簡潔・詳細) については出力長も比較

カテゴリ	タイプ	例
Default	-	銭湯に持っていくと便利なものを教えてください。
Situation	Audience	手荒れしやすい体質です。 銭湯に持っていくと便利なものを教えてください。
Style	Simplicity	Simple 簡潔に回答してください。 銭湯に持っていくと便利なものを教えてください。
	Detailed	詳細に回答してください。 銭湯に持っていくと便利なものを教えてください。

他 7 Styleタイプ: Formality(Formal, Informal), Politeness(Polite, Impolite), Gender(Feminine, Masculine), Romance, Engagingness, Humor, Toxicity

図1 問題・コンテキスト例

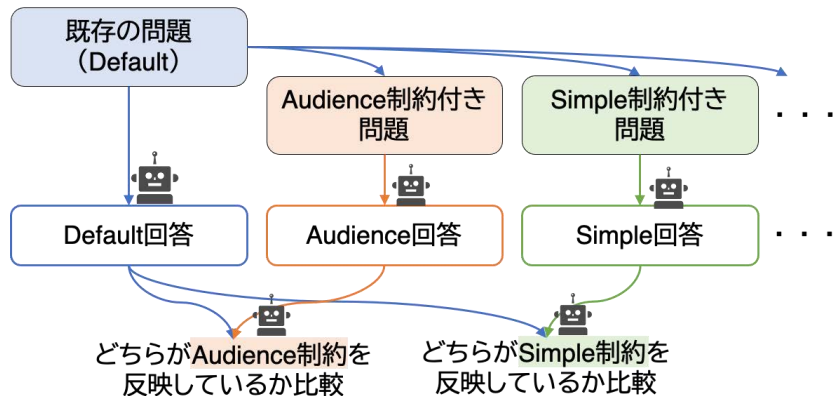


図2 データ作成・評価フロー

[1] <https://huggingface.co/datasets/llm-jp/llm-jp-instructions>

# 2

## 結果と分析

- 「コンテキスト回答がより反映」の割合を比べると、Audience制約に対する回答能力は他のStyleと比べて中程度。LLMが指示されなくても持つようなStyle (Formal, Polite, Detailed) は低かった。Offensiveはllama-3.1が比較的高くチューニングの影響の可能性がある。
- 簡潔・詳細の指示に応じてgpt-4oは比較的回答長が変化している。OLMo-2は出力長の変化によらず「どちらも反映していない」が多い。LLM-as-a-judgeにおいて回答長以外の要因が影響している可能性がある。

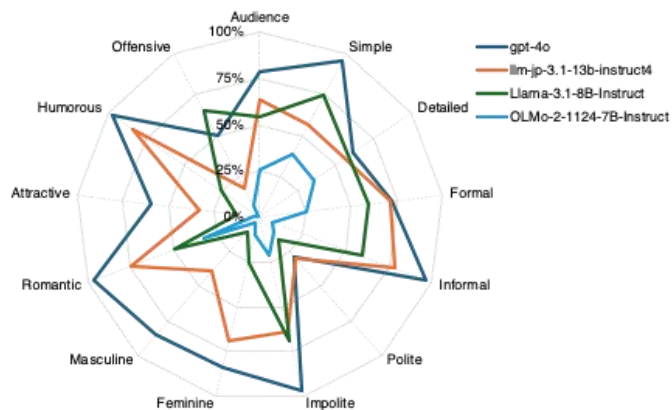


図3 「コンテキスト回答がより反映」の割合

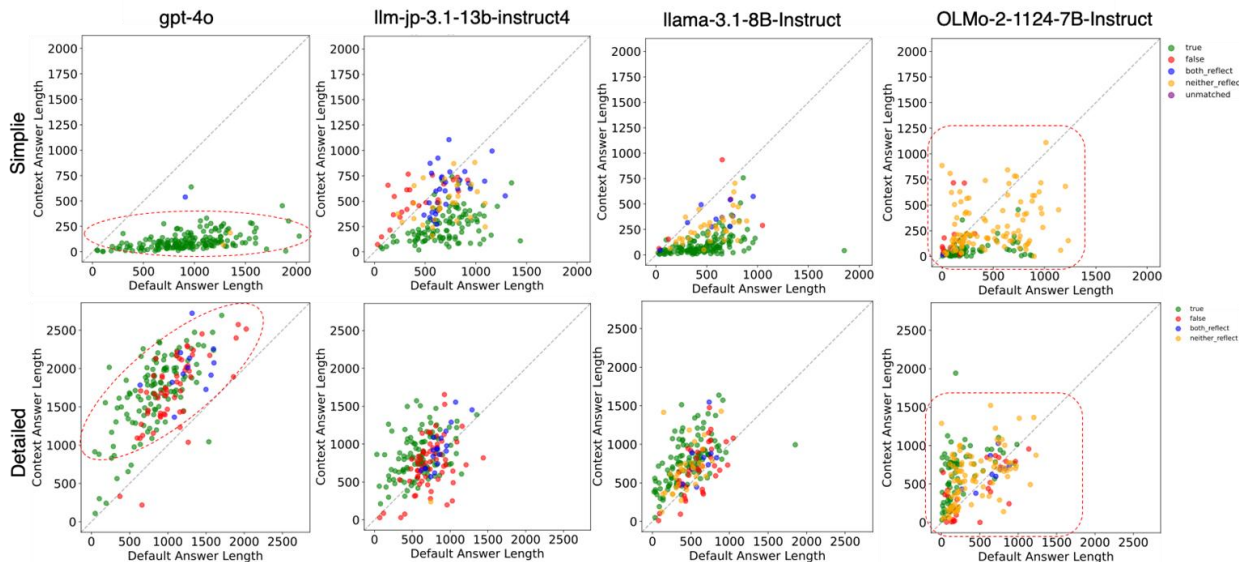


図4 デフォルト回答とSimple・Detailedコンテキスト回答の回答長の比較

(横軸：デフォルト回答、縦軸：コンテキスト回答)

# ③ 生成系タスクの自動評価においてチェックリストの使用は有効なのか？（古橋+、NLP2025, EMNLP2025 Main）



## ● 背景:

- 生成系タスクの自動評価において、**人手評価との相関を向上させる**ため、先行研究では評価基準を細分化した「**チェックリスト**」という枠組みを導入していた。
- 一方、**どのようなチェックリストが自動評価に対して有効であるかなどの調査は不十分**。

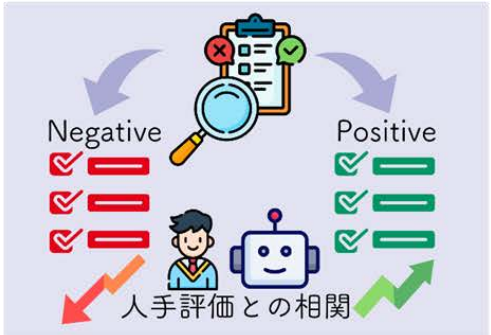
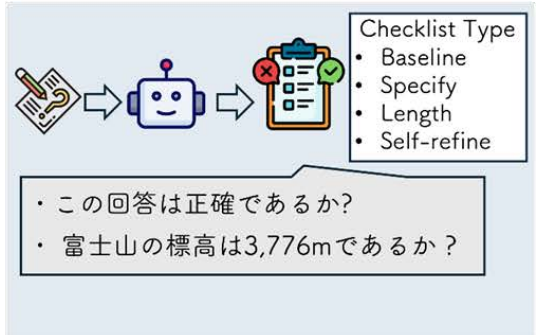
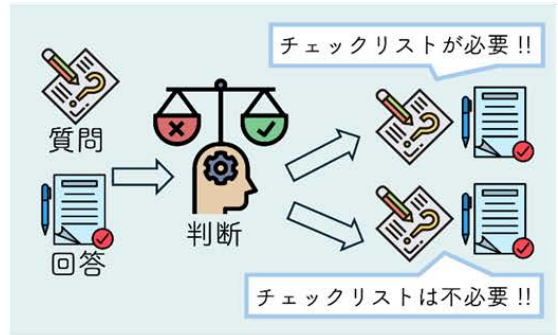
## ● 本研究:

- 2種類のデータ形式（ペアワイズ、スコアリング）に対して**6種類**の生成方法でチェックリストを生成。
- **8種類**の評価モデルを使用して、**回答の評価**

RQ1: チェックリストが必要かどうかを判断することができるのか？

RQ2: 有効なチェックリストをどのように作成することができるのか？

RQ3: どのチェックリストが人手評価との相関に寄与するのか？



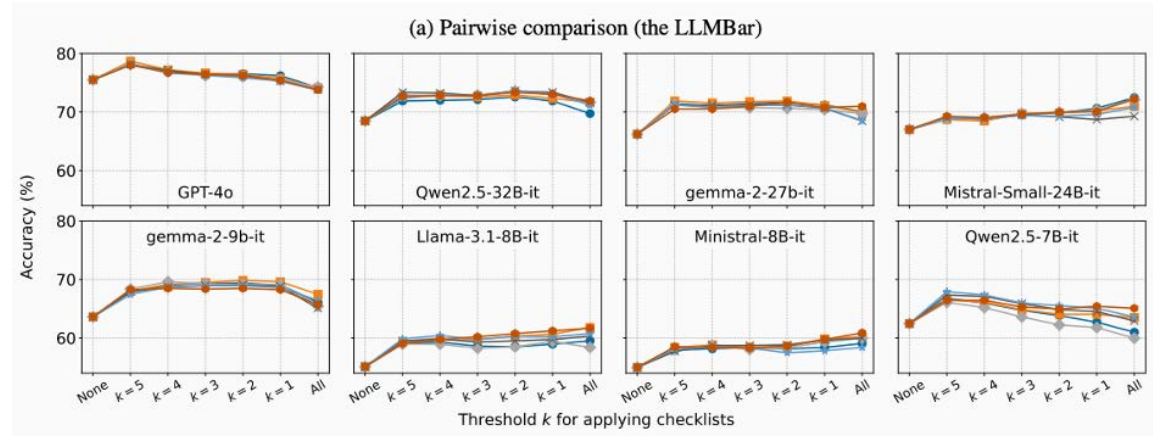
# ③ 生成系タスクの自動評価においてチェックリストの使用は有効なのか？（古橋+, NLP2025, EMNLP2025 Main）



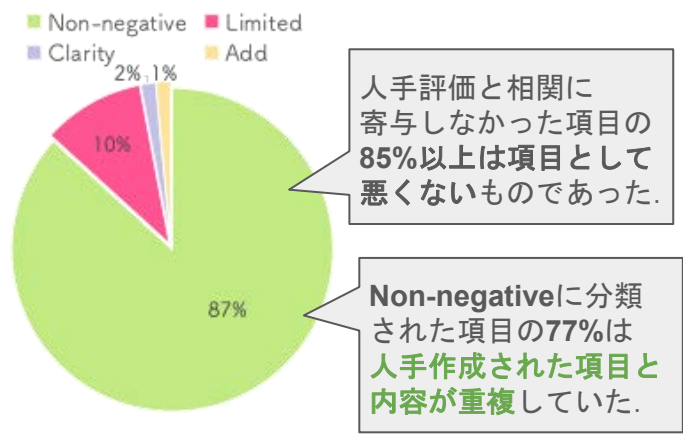
## ● 結果・考察:

- 一貫して良い・悪いチェックリスト生成方法は**見つからなかった**.
- 人手評価との相関に寄与しない項目でも、先行研究で人手で作成された項目と内容が重複していた。
- 人手評価を行う際から、評価観点をより明確にすることが大事。

ペアワイズ評価における評価結果



人手評価と相関に**寄与しなかった項目**の特徴.



# ④

## 大学入試問題を評価データセットとして整備（高見+）

センター試験のXMLデータをLLM評価データとして整備

- 設問と選択肢をJSON形式に変換(物理・化学・生物各5年分)
- 画像の有無を明示し、画像ファイルも格納
- 人（受験生）の全国規模の正答率などの統計情報との比較が可能。
  - 学校等の教育機関では機密性の高い情報を扱うため、local環境でも稼働できるLLMで比較評価

Kyosuke Takami, Satoshi Sekine, and Yusuke Miyao. Evaluating Local LLMs on Japanese National University Entrance Examination Dataset in Comparison with Student Performance. Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025

<https://doi.org/10.5281/zenodo.15870238>

```
"source": "National Center For University Entrance Examination",
"subject": "Kagaku(main exam)",
"year": "1993",
"question_id": "Q14",
"label": "問2",
"text": "\n\n次の問い(問1~3)に答えよ。 \n\n\n次の記述①~⑤のうちから、正しいものを一つ選べ。 \n",
"choices": {
  "choice1": "①水溶液中での酢酸の電離度は、その濃度が小さくなるにつれて、小さくなる。",
  "choice2": "②純水の電離度は、室温で $1 \times 10^{-7}$ である。",
  "choice3": "③一定温度の酸や塩基のうすい水溶液では、水のイオン積はpHによらず一定である。",
  "choice4": "④pH 4の塩酸とpH12の水酸化ナトリウム水溶液とを同体積ずつ混合すると、その溶液のpHは8となる。",
  "choice5": "⑤酢酸水溶液に水酸化ナトリウム水溶液を加えると、溶液中の酢酸イオンの濃度が減少する。",
},
"answer_style": "multipleChoice",
"answer_type": "sentence",
"knowledge_type": "KS_DM_C",
"need_image": "yes",
"image_files": {
  "Center-1993-Main-Kagaku-003.png"
},
"correct_answer": "3",
"score": "4",
}
```

データセット：<https://github.com/KyosukeTakami/center-examination-jp>

4

大学入試センター試験データを用いた日本語LLMの評価比較  
ローカル環境で稼働可能な13B程度のモデルの有効性を検討  
受験生（人）との比較を通じたパフォーマンス検証

\*今回はテキストのみで評価

\*（カッコ内画像を必要としないテキストのみで解ける問題）

Subject	Number of question	llm-jp-3-13b-instruct3	gemma-3-12b-it	DeepSeek-R1-Distill-Qwen-14B	Student average score (weighted/normalized score)
<u>Chemistry</u> 2009	25 (13)	17.86 (18.75)	39.29 (50.00)	32.14 (43.75)	69.54
2005(IB)	29 (8)	31.03 (25.00)	37.93 (50.00)	34.48 (62.50)	66.06
2001(IB)	29 (6)	24.10 (16.67)	31.03 (50.00)	24.14 (33.33)	58.51
1997(IB)	23 (2)	17.30 (50.00)	39.13 (50.00)	<b>34.78</b> <b>(100.0)</b>	62.93
1993	24 (9)	<b>37.50</b> <b>(55.56)</b>	<b>45.83</b> <b>(66.67)</b>	29.17 (44.44)	58.69

化学：1993年の一部設問で人間に相当する成績  
<https://doi.org/10.5281/zenodo.15870238>

Kyosuke Takami, Satoshi Sekine, and Yusuke Miyao. Evaluating Local LLMs on Japanese National University Entrance Examination Dataset in Comparison with Student Performance. Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025

# ⑤ 日本語大規模言語モデルの有用性と安全性の両立に向けたチューニング手法の検討（勝又+）

- 背景：LLM の安全性と有用性の中でトレードオフの関係が報告
- 本研究：日本語 LLM の安全性向上に向けて、有用性をできる限り損なわず、学習データを工夫した検証（SFT、DPO）を実施
  - SFT: 有用性データと比べて、量が少ない安全性データをオーバーサンプリング（n倍）
  - DPO: DPO 向けの安全性データと有用性データを作成



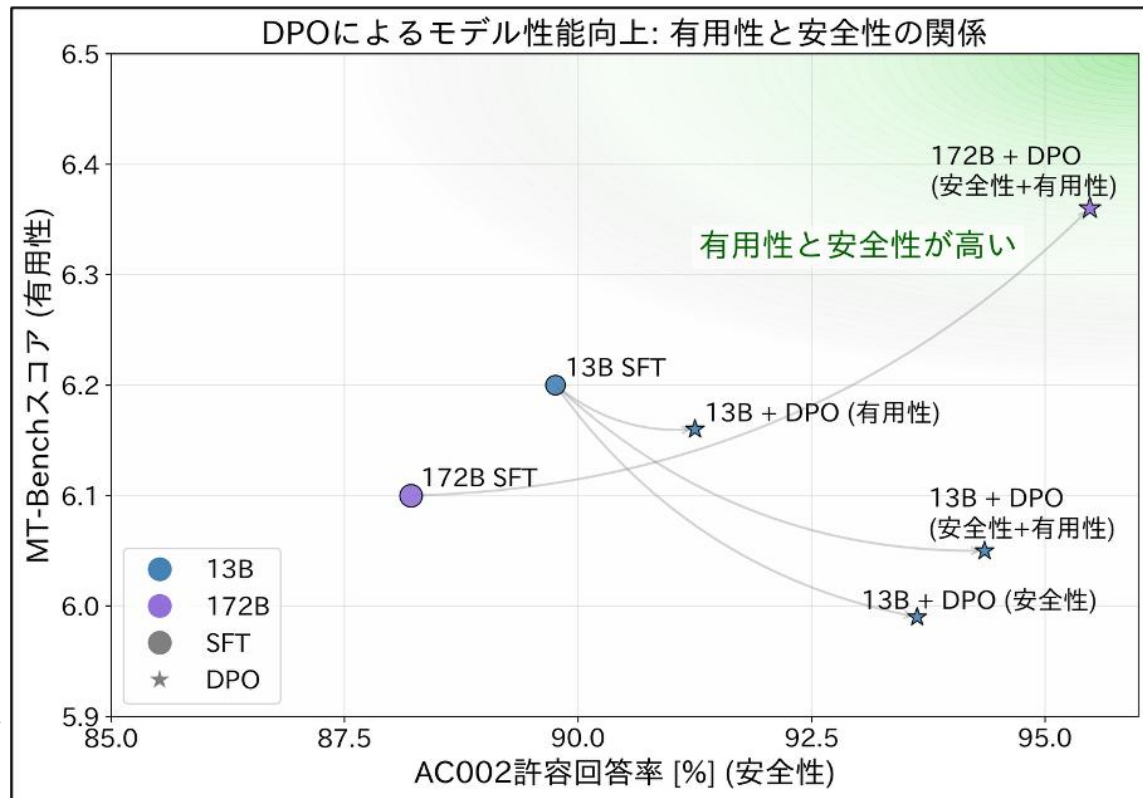
SFT 概要図



DPO 概要図

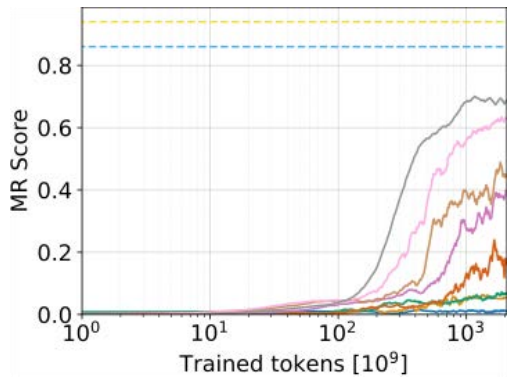
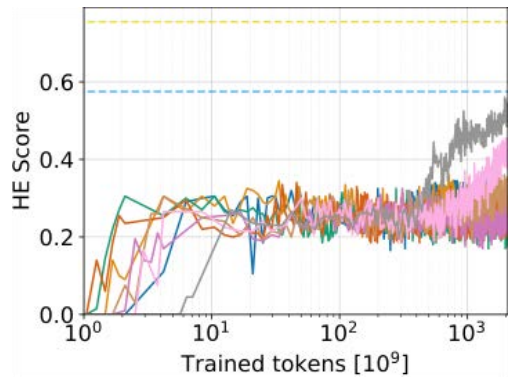
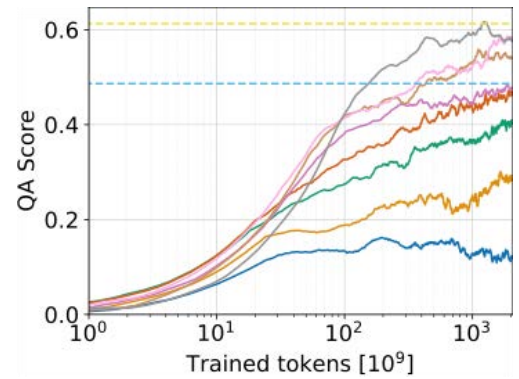
# ⑤ 日本語大規模言語モデルの有用性と安全性の両立に向けたチューニング手法の検討 (勝又+)

- DPO の検証結果
  - SFT 実施済み 13B LLM と 172B LLM に対して、DPO を実施
- 安全性について
  - DPO は有効に機能
  - 安全性データと有用性データを混ぜた際が一番良い結果
- 有用性について
  - 安全性データのみだと、有用性が大きく減少する
  - 有用性データを混ぜることでこの減少を緩和



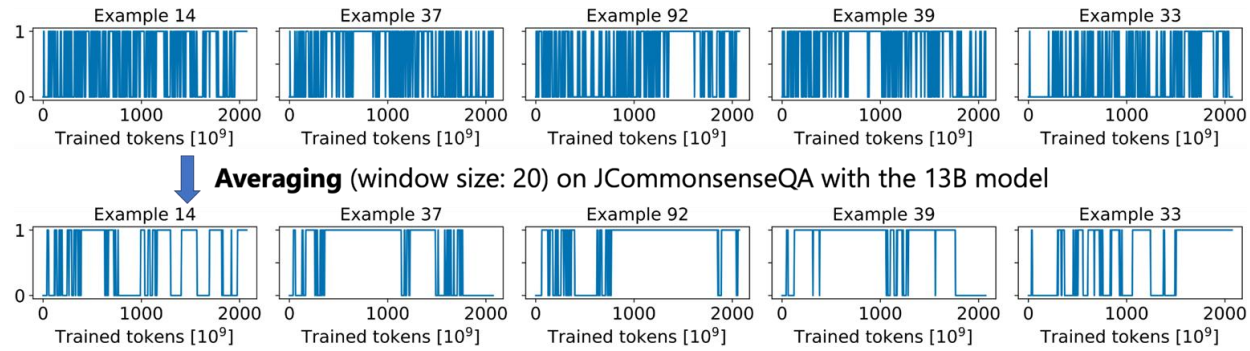
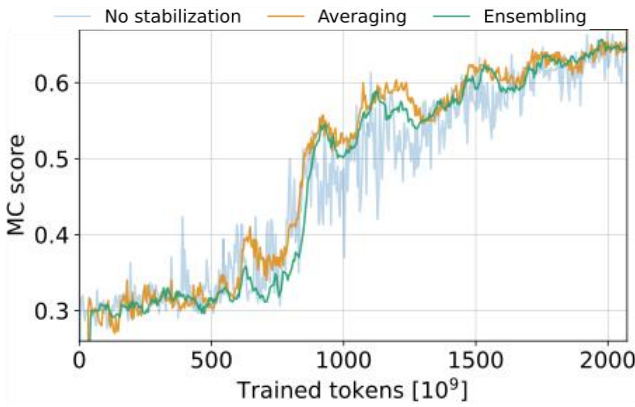
# ⑥ 日本語大規模言語モデルの事前訓練の分析 (西田+)

- 研究1：下流タスク性能の変動の軌跡を類型化する (NLP2025)
  - 日本語LLMがどのようにタスク遂行能力を獲得するかを理解するための初期検討として、モデルサイズやタスク、評価指標の観点から下流タスク性能の変動の軌跡の類型化を試みた
- 結果:
  - 十分大きなサイズのモデルにおいて軌跡は**漸進型**、**二段階型**、**後発型**の3種類に類型化できる
    - 漸進型：学習初期から後期にかけて徐々にスコアが向上 (EL, FA, MT, QA, RC)
    - 二段階型：学習初期に急激にスコア上昇→中期に停滞→後期に再び徐々に上昇 (HE, MC, NLI)
    - 後発型：学習初期は全くタスクが解けず、学習後期から徐々にスコアが向上し始める (MR)
  - 学習過程の類型はタスクや評価尺度の性質で特徴づけられる



# ⑥ 日本語大規模言語モデルの事前訓練の分析 (西田+)

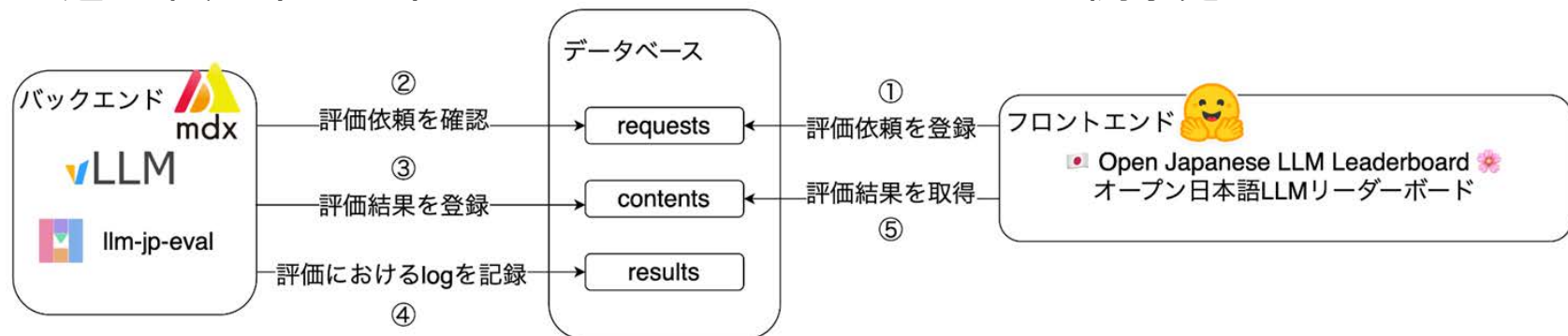
- 研究2: 下流タスク性能の不安定性を緩和する (SIG-NL263, **EMNLP2025 Findings**)
  - LLMの事前学習過程における下流タスクの性能には短期的な振動 (不安定性) がみられる
  - モデルサイズを大きくしても不安定性は緩和されないことを観察した
  - この不安定性は評価の頑健性や信頼性を損なう要因となるため、後処理での緩和を試みた
- 手法: 隣接するチェックポイントの統合 (平均化・多数決)
- 結果:
  - タスク全体および個別の事例で不安定性の緩和を確認した



7

# オープン日本語大規模言語モデルリーダーボードの構築 (Han+)

- **背景**：日本語大規模言語モデル（LLM）の性能評価の重要性が増大
  - 日本語評価ベンチマークによる、複数LLMの評価結果を一覧できるリーダーボードの必要性
- **本研究**：ユーザー参加型のリーダーボードを構築し、24年12月から公開
  - 評価対象：25年11月現在、862個のモデル
  - 評価手法：日本語LLMの評価ツールである、llm-jp-eval v1.4.1を用いて評価
  - mdxの計算ノードを用いて、HuggingFace Hubに存在する70B級のLLMまで評価可能
- 近日中、Open Japanese LLM Leaderboard v2 の公開予定



7

# オープン日本語大規模言語モデルリーダーボードの構築 (Han+)

## ● リーダーボード及び分析図の例

<https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard>



**Open Japanese LLM Leaderboard**  
オープン日本語LLMリーダーボード

LLM.jpによるオープン日本語LLMリーダーボードは、古典的なものから最新のものまで16種類以上のNLPタスクを用いて日本語大規模言語モデル(LLM)の性能を評価します。オープン日本語LLMリーダーボードは、日本の国立情報学研究所を中心に日本語LLMの研究開発を行う組織連携プロジェクト(LLM.jp)のオープンソース貢献者によって構築されました。

「LLM Benchmark」ページでは、疑問行「?」はHugging Faceのモデルカードで不明なパラメータを示しています。データセットに関する詳細情報については、「About」ページを参照するか、LLM.jpのウェブサイトをご覧ください。また、「Submit here」ページでは、あなたのモデルの性能を評価し、リーダーボードに参加することができます。

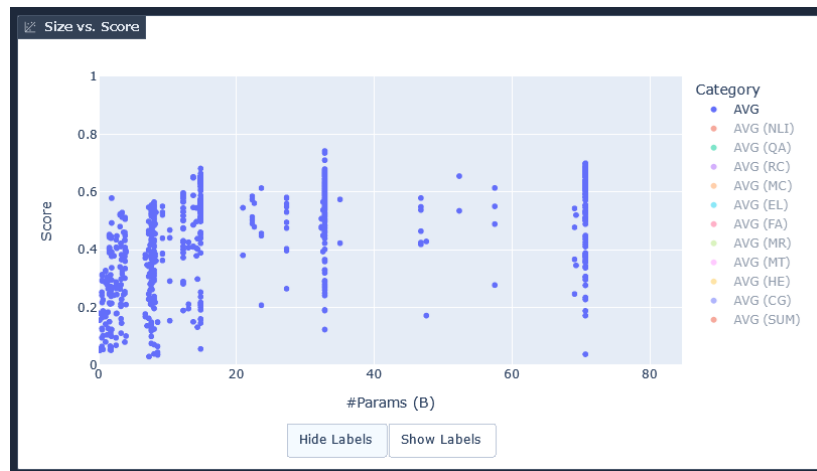
[LLM Benchmark](#) [About](#) [Submit here!](#)

Search for your model (separate multiple queries with ",") and press ENTER...

Column Filter

Model Filter

T	Model	AVG (B)	AVG (NLI)	AVG (QA)	AVG (RC)	AVG (MC)	AVG (EL)	AVG (FA)	AVG (MR)
●	deep-analysis-research/flux-japanese-Queen2.5-320	0.742	0.8803	0.596	0.9202	0.9133	0.6993	0.5176	0.94
●	deep-analysis-research/flux-japanese-Queen2.5-320	0.735	0.8861	0.5793	0.907	0.9124	0.7117	0.4786	0.93
●	deep-analysis-research/O2II-japanese-Queen2.5-320	0.71	0.8793	0.5897	0.9885	0.9139	0.6782	0.4321	0.93
●	Saxo/linkbricks-Horizon-AI-Japanese-Superb-V4-78	0.6995	0.7741	0.7523	0.926	0.9894	0.5949	0.3418	0.93
●	Saxo/linkbricks-Horizon-AI-Japanese-Pro-V5-788	0.6992	0.7843	0.7473	0.9257	0.9171	0.6142	0.354	0.94
●	nitky/llama-3.1-SuperSeaFlow-788-Instruct-v0.1	0.6981	0.7734	0.7385	0.9241	0.9174	0.5974	0.3439	0.93
●	Saxo/linkbricks-Horizon-AI-Japanese-Avengers-V4	0.6969	0.8123	0.7681	0.9228	0.9178	0.6246	0.3345	0.95
●	Saxo/linkbricks-Horizon-AI-Japanese-Advanced-V4	0.6988	0.7792	0.739	0.9239	0.911	0.61	0.3478	0.93
●	Saxo/linkbricks-Horizon-AI-Japanese-Advanced-V2	0.6869	0.7842	0.7384	0.9231	0.9871	0.6121	0.3492	0.94
●	Saxo/linkbricks-Horizon-AI-Japanese-Advanced-V1	0.6865	0.8849	0.7283	0.9272	0.9119	0.6866	0.3313	0.94
●	Saxo/linkbricks-Horizon-AI-Japanese-Superb-V3-78	0.6862	0.7789	0.765	0.9244	0.9132	0.5842	0.3358	0.92
●	Saxo/linkbricks-Horizon-AI-Japanese-Advanced-V1	0.6847	0.8886	0.7283	0.9263	0.9118	0.5985	0.3351	0.94



⑧

## llm-jp-judge: 生成自動評価を行うためのPythonツール (中山+、NLP2025 優秀賞)

- 日本語LLM-as-a-Judge評価を統一的に行うことのできるツールllm-jp-judgeを提案
  - MT-Bench (日/英) : 2ターン対話の評価
  - 品質評価 (日) : 応答の品質を5段階評価
  - 安全性評価 (日) : 応答安全性を5段階評価
- 人手によるメタ評価を実施し、llm-jp-judgeによる評価の妥当性を検証
- オープンソースとして公開中

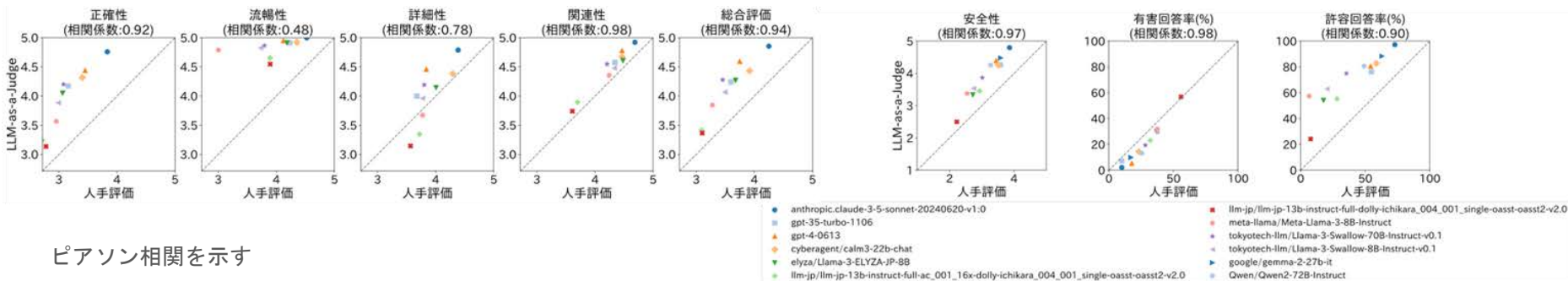
<https://github.com/llm-jp/llm-jp-judge>



⑧

# llm-jp-judge: 生成自動評価を行うためのPythonツール (中山+, NLP2025 優秀賞)

- データセット
  - 品質評価: 10モデルの応答に対してllm-jp-judgeと同じ基準で評価したデータ(101件)
  - 安全性評価: 12モデルの応答に対してllm-jp-judgeと同じ基準で評価したデータ(181件)
- 評価モデル: GPT-4o
- 品質評価
  - 正確性、詳細性、関連性、総合評価で人手によるモデル順位と高い相関
  - 正確性と流暢性でLLM-as-a-Judgeによる評価が甘い傾向
- 安全性評価
  - 人手によるモデル順位と高い相関
  - 有用回答率(安全性評価が4以上の割合)でLLM-as-a-Judgeによる評価が甘い傾向



ピアソン相関を示す

# サーベイ論文（2024年夏）

788

人工知能 39巻6号（2024年11月）

特集 「生成 AI とベンチマークデータセット・評価手法」

## 大規模言語モデルの評価とその課題

Evaluation of Large Language Models and Its Issues

菅原 朔  
Saku Sugawara  
国立情報学研究所  
National Institute of Informatics  
saku@nii.ac.jp

村脇 有吾  
Yugo Murawaki  
京都大学  
Kyoto University  
murawaki@i.kyoto-u.ac.jp

宮尾 祐介  
Yusuke Miyao  
東京大学  
The University of Tokyo  
yusuke@is.s.u-tokyo.ac.jp

**Keywords:** large language models, evaluation, benchmarking.

### 1. はじめに

OpenAI による ChatGPT の登場以降、自然言語処理分野にとどまらず多くの分野に大規模言語モデル (LLM) の利用が広がり、産学問わず盛んに研究開発が進んでいる。そのどのような場面において、適切な意

る。しかし、これまでの論文は単に既存のデータセットや指標をまとめるものが中心であり、それらの課題まで踏み込んでいるものは多くない。また、日本語で書かれたサーベイはこれまでに存在せず、日本国内の多くの方が評価方法の概要や課題についてアクセスしづらい状況にある。

本記事の目的は、これまでの LLM の評価方法について

## まとめ

- 異なる組織・分野から多様な研究者が集まっています
- 透明性・信頼性に関わる成果を着実に出しており、今後さらに展開していきます
- RA など継続的に募集しています（お声がけください）