JAMSE：日本語 LLM 評価用の高品質な少サンプル日本語ベンチマークの作成および評価 – GENIAC LLM 開発コンペティションからの知見 –

山崎 友大[1] 谷口 仁慈[2] 山際 愛実[3] 原田 憲旺[4] 小島 武[4] 岩澤 有祐[4] 松尾 豊[4]
[1] 京都大学 [2] 琉球大学 [3] 三重大学 [4] 東京大学
[1]yamazaki.yudai.82m@st.kyoto-u.ac.jp
[2]k248443@eve.u-ryukyu.ac.jp
[3]424M248@m.mie-u.ac.jp
[4]{keno.harada,t.kojima,iwasawa,matsuo}@weblab.t.u-tokyo.ac.jp

概要

日本語 LLM の開発が盛んな一方，日本語評価ベンチマークの数や質は十分でない．また，既存のベンチマークはサンプル数が多く，評価コストが高いという問題がある．本稿では，高品質かつ少サンプルな評価用日本語ベンチマーク JAMSE を提案する．JAMSE は，7 つのベンチマークから構成されており，1 ベンチマークあたり 100 サンプルである．そのため，日本語 LLM の言語理解能力と生成能力を低コストで評価することができる．国内外の継続学習モデルや GENIAC コンペティションで開発されたモデルで評価を行ったところ，Nejumi Leaderboard Neo による評価結果と強い相関が確認された．[1)]

1 はじめに

OpenAI 社の GPT モデルをはじめ，世界中で新たな大規模言語モデル（LLM）が開発され続けている．日本でも，高性能な日本語 LLM を目指した開発が進められ，発表されるモデルの数は加速度的に増加してきた．

こうした新規モデルの性能向上に合わせて，日本語の評価ベンチマークも継続的に整備されることが重要である．しかし，JGLUE[1] や Rakuda ベンチマーク[2)]など，いくつかのベンチマークが発表されてはいるものの，依然としてその数は少ない．

また，既存のベンチマークはサンプル数が非常に多く，10^5 程度のサンプルが含まれることも珍しくない．このようなベンチマークを LLM 評価に用いると，GPU 計算に 4,000 時間以上かかるという報告もあり [2]，計算コストが問題となっている．

そこで本研究では，高品質かつ少サンプルの日本語ベンチマーク JAMSE （JApanese Minimum-Sized Evaluation benchmark）を提案する．JAMSE は，一問一答タスクを含む 6 つのベンチマークと，文章生成タスクを含む 1 つのベンチマークの計 7 ベンチマークから構成されている．そのため，JAMSE で高いスコアを獲得するには，幅広い知識と高度な言語理解能力に加え，指示に従って適切な回答を生成する能力が要求される．また，1 ベンチマークあたり 100 サンプルなので，評価コストは極めて小さい．

なお，JAMSE 作成の背景には，「経済産業省が主導する基盤モデルの開発に必要な計算資源に関する支援や関係者間の連携を促す『GENIAC』プロジェクト」の一環として行われた，松尾・岩澤研究室 LLM 開発コンペティション（GENIAC コンペティション）がある．本稿では，JAMSE の評価例として，国内外の継続学習モデルに加え，コンペティションで開発された日本語 LLM[3)]による評価結果を報告する．

2 関連研究

2.1 日本語ベンチマークの整備

これまで，様々な日本語 LLM が公開されてきた．特に近年は，大学や企業が独自モデルの開発に注力しており，Weblab[4)]，Swallow[3]，ELYZA[4]，

1） 本研究で新たに翻訳されたデータセットは以下で公開されている: https://huggingface.co/collections/weblab-GENIAC/jamse-66e49d2241e4987cc0283036
2） https://yuzuai.jp/benchmark
3） https://huggingface.co/weblab-GENIAC
4） https://huggingface.co/matsuo-lab/weblab-10b-instruction-sft

PLaMo[5], CALM3[5] などが発表されてきた．これにともなって，日本語ベンチマークの拡充も図られており，JGLUE，Rakuda ベンチマーク，llm-jp-eval[6]，ELYZA-tasks-100[7] などが公開されている．しかし，海外のベンチマークの豊富さや作成ペースと比較すると，日本語ベンチマークの整備は依然として遅れをとっており，日本語 LLM の性能を正しく評価できる高品質なベンチマークの作成が急務である．

2.2 少サンプルによるモデル評価

英語ベンチマークでは，全サンプルによる評価を少サンプルによる評価で代替できることが報告されている．Vivek らは，Anchor Point Selection を提案している [8]．これは、大規模なデータセットを信頼度という指標に基づいてクラスタリングし，クラスタの中心点を用いてデータセット全体にわたるモデルの挙動を推定する手法である．いくつかの分類タスクにおいて，10^2 個の中心サンプルを用いた場合と全データを用いた場合で，分類の一致率は平均 80% となり，少サンプルでもモデルの性能を十分推定できることが分かった．Polo らは，項目応答理論（IRT）[9] を取り入れた推定手法を提案している [10]．この手法では，サンプルにフィッティングした IRT モデルによる表現を用いてクラスタリングを行い，その中心点によって，モデルの評価を推定する．検証に用いた全てのベンチマークにおいて，10^2 個程度の中心サンプルによる正答率と全サンプルによる正答率が，誤差 2% 以内で一致した．

いずれの提案手法でも，10^2 個程度の代表サンプルによってデータセット全体の評価を推定できると報告されている．また，提案手法には劣るものの，比較として行われたランダムサンプリングによる評価でも十分な推定性能を発揮することが分かっており，Polo らの検証では，代表サンプルと全サンプルの誤差は 6% 以内に収まっている．したがって，10^2 個程度のランダムサンプリングによるサブセットでも評価データセットとして機能すると考えられる．

3 JAMSE の構築

3.1 選定ベンチマーク

JAMSE には，日本語ベンチマークとして広く用いられている llm-jp-eval，Elyza-tasks-100 に加え，海外モデルの評価の試金石となっている Open LLM Leaderboard v1[11] から，GSM8K[12] を除いた 5 つのベンチマークを取り入れた[6]．

llm-jp-eval　llm-jp-eval は，既存の日本語評価データセット 13 個から構成される生成問題のベンチマークであり，自然言語推論，QA，文章補完など 8 個のカテゴリを幅広くカバーしている．

ARC　ARC[13] は，高度な知識や推論性能を評価するベンチマークである．自然科学に関する 7,787 問の 4 択問題で構成されている．単純な検索や単語の相互関係から解答を導くことが難しい問題が含まれており，LLM の深い知識や推論能力が要求される．

MMLU　MMLU[14] も，高度な知識や推論性能を評価するベンチマークである．57 の分野（STEM，人文科学，社会科学など）をカバーした，17,844 問の 4 択問題で構成されている．高いスコアを獲得するには，幅広い知識と専門家レベルの問題解決能力が必要となる．MMLU を日本の文化に適合するように再構築した JMMLU[15] が作成されている．

TruthfulQA　TruthfulQA[16] は，モデルの真実性を評価するベンチマークである．健康，法律，金融，政治を含む 38 のカテゴリーにまたがる 817 の質問で構成されており，質問の内容が真実か否かを正確に判断する能力が求められる．TruthfulQA も，日本語で再構築した JTruthfulQA[17] が作成されている．

WinoGrande　WinoGrande[18] は，常識的な推論性能を評価するベンチマークである．日常の出来事に関する短文の中に代名詞が含まれており，対応する先行詞を選択する，43,972 問の 2 択問題で構成されている．感情分析による解答が困難な問題が含まれるため，人間レベルの常識が必要とされる．

HellaSwag　HellaSwag[19] も WinoGrande 同様，常識的な推論性能を評価するベンチマークである．動画や物の使用方法に関するキャプションが与えられ，末尾に続く文章として最も妥当なものを選択する，59,950 問の 4 択問題で構成されている．選択肢が敵対的手法によって生成されているため，単語の関連から解答することが難しいタスクである．

5） https://huggingface.co/cyberagent/calm3-22b-chat

6） 複数ステップの数学的推論を要する GSM8K，コンペティションで開発された 10B 程度のモデルにとって難しく，ベンチマークとして機能しないと推測されたため除外した．

表 1　ARC における機械翻訳と人手修正の例.

原文	機械翻訳	人手修正
What chemical symbol represents the element copper? A.C, B.Co, C.Cp, D.Cu	銅元素を表す化学記号は何ですか? A.C, B. コ, C.CP, D. と	銅の元素を表す化学記号は次のうちどれですか？ A.C, B.Co, C.Cp, D.Cu
Some birds fly south in the fall and return in the spring. This is an example of A.migration, B.camouflage, C.hibernation, D.growth	秋に南に飛んで春に戻ってくる鳥もいます。これは一例です A. 移行, B. 迷彩, C. 冬眠, D. 成長	秋に南に飛んで春に戻ってくる鳥がいます。これはなんの例ですか？ A. 移動, B. 擬態, C. 冬眠, D. 成長

ELYZA-tasks-100　ELYZA-tasks-100 は，回答生成能力を評価するベンチマークである．択一式の上記 6 つと異なり，Elyza-tasks-100 は，回答が一意に定まらない生成問題 100 問から構成される．質問内の要求には複雑な指示やタスクが含まれており，適切で正確な文章を生成することが求められる．回答の評価（0-5）は，GPT-4 による絶対評価で行われる．

3.2　翻訳

llm-jp-eval や Elyza-tasks-100，すでに日本語版が公開されていた JMMLU と JTruthfulQA は，それぞれからランダムに 100 サンプルずつ抽出して利用した．ARC，WinoGrande，HellaSwag は，サブシナリオごとの偏りがないようにサンプル数のバランスをとったうえで，それぞれからランダムに 100 サンプルずつ抽出し，機械翻訳（Google Translate）で翻訳した．翻訳後の文章は，翻訳内容が原文の意図と異なっているものや，出力フォーマットが統一されていないものが散見されたため，必要に応じて著者らが人手で修正を施した（表 1）．本研究で新たに翻訳したベンチマークは，それぞれ，JARC，JWinoGrande，JHellaSwag とした．

4　モデル評価

4.1　評価データおよび評価指標

評価データ　一問一答タスクに，llm-jp-eval，JARC，JMMLU，JTruthfulQA，JWinoGrande，JHellaSwag，文章生成タスクに，Elyza-tasks-100 を用いた．Elyza-tasks-100 は 0 ショットとし，その他のベンチマークは 4 ショットとした．few-shot には各ベンチマークの 100 サンプルからランダム抽出したものを使用し，評価は残りの 96 サンプルで行った．

評価指標　一問一答タスクは，各ベンチマークの正答率 $\in [0, 1]$ の平均をとり，Score_{ch} とした．文章生成タスクは，Elyza-tasks-100 の評価平均を $[0, 1]$ の範囲に正規化して $\text{Score}_{\text{gen}}$ とした．そして，Score_{ch} と $\text{Score}_{\text{gen}}$ の平均を総合スコア $\text{Score}_{\text{total}}$ とした．

4.2　各モデルによる評価結果

本稿では，国内外の継続学習モデルに加え，GENIAC コンペティションで開発された日本語 LLM に対する評価結果を示す．継続学習モデルには，GPT-3.5-turbo，Llama-3-youko-8b（Llama-8b），Weblab-10b-instruction-sft（Weblab-10b）を採用した．

表 2 に，各モデルの評価結果を示す．上段が継続学習モデル，下段が GENIAC プロジェクトで開発されたモデルである．GENIAC モデルは，「Hugging Face 上のモデル名（チーム名）」と表記する．GPT-3.5-turbo は，Score_{ch}，$\text{Score}_{\text{gen}}$，$\text{Score}_{\text{total}}$ だけでなく，ほとんどのベンチマークでも最大のスコアを獲得し，Llama-8b がそれに続いた．GENIAC モデルは，GPT-3.5-turbo には及ばなかったものの，9 モデル中 4 モデルは，$\text{Score}_{\text{total}}$ および $\text{Score}_{\text{gen}}$ で Llama-8b に匹敵する結果となった．また，7 モデルは，Weblab-10b よりも $\text{Score}_{\text{total}}$ が高くなった．

各ベンチマークのスコアを見ると，llm-jp-eval，JARC，JMMLU，Elyza-tasks-100 のスコアが高いモデルほど $\text{Score}_{\text{total}}$ も高い傾向にあることが分かる．JWinoGrande と JHellaSwag は，モデル間でスコアに大きな差が見られず，いずれもチャンスレート（それぞれ 0.50 と 0.25）付近のスコアになった．これは，両ベンチマークが，評価対象モデルにとって難しいタスクであったことを示唆している．JTruthfulQA は，GENIAC モデルのスコアがチャンスレート（およそ 0.23）付近なのに対し，GPT-3.5-turbo や同サイズの Llama-8b のスコアは非常に高くなった．この原因として，学習データやトレーニング手法，アーキテクチャの違いのほか，データリークなどの可能性も考えられる．

実行時間に着目すると，ほとんどのモデルで数十分 ~2 時間程度であり，最大でも 5 時間であった．全サンプルを用いた従来の評価では，数百 ~ 数千時間を要していたことから，大幅なコスト削減を達成できたと言える．

表 2　モデルの評価結果．ベンチマーク名は略称を用いた．$\text{Score}_{\text{total}}$ 順に並べており，太字が最も高いスコア，下線が次点のスコアを示している．いずれも小数第 3 位を四捨五入しているが，太字・下線は実際のスコアに基づいている．

		一問一答							文章生成		Runtime
	$\text{Score}_{\text{total}}$	Score_{ch}	llm-jp-eval	JARC	JMMLU	JTQA	JWG	JHS	$\text{Score}_{\text{gen}}$	Elyza	(min.)
GPT-3.5-turbo	**0.64**	**0.55**	**0.65**	**0.70**	**0.53**	**0.65**	**0.53**	**0.27**	**0.73**	**3.67**	20
Llama-3-youko-8b	0.47	0.50	0.62	0.63	0.50	0.61	0.46	0.21	0.43	2.17	195
Weblab-10b-instruction-sft	0.29	0.29	0.37	0.18	0.28	0.20	0.45	0.25	0.30	1.50	134
Tanuki-8x8B-dpo-v1.0 (GENIAC たぬき)	**0.45**	**0.40**	**0.55**	**0.49**	**0.44**	0.20	0.45	0.25	0.50	2.48	109
Tanuki-8B-dpo-v1.0 (GENIAC たぬき)	0.44	0.32	0.31	0.40	0.29	**0.22**	0.46	0.25	**0.56**	**2.80**	45
team_hatakeyama_phase1 (GENIAC たぬき)	0.42	0.37	0.43	0.32	0.39	0.19	**0.60**	0.26	0.48	2.38	31
team_kawagoshi_submit (GENIAC ビジネス)	0.40	0.36	0.43	0.39	0.31	**0.22**	0.53	0.27	0.45	2.23	15
team_ozaki_submit (GENIAC 天元突破)	0.35	0.28	0.36	0.26	0.25	0.17	0.44	0.22	0.41	2.06	36
team_kumagai_submit (GENIAC 甲)	0.31	0.27	0.21	0.32	0.23	0.11	0.45	0.27	0.35	1.76	119
team_kumagai_submit (GENIAC Kuma)	0.31	0.31	0.31	0.23	0.24	**0.22**	0.55	**0.30**	0.32	1.60	52
team_nakamura_submit (GENIAC JINIAC)	0.25	0.24	0.19	0.10	0.18	0.19	0.52	0.26	0.26	1.31	22
team_sannai (GENIAC Zoo)	0.11	0.01	0.07	0	0	0	0	0	0.21	1.04	317

表 3　JAMSE と Nejumi Leaderboard Neo の評価比較．

	JAMSE			Nejumi Leaderboard Neo		
	$\text{Score}_{\text{total}}$	Score_{ch}	$\text{Score}_{\text{gen}}$	$\text{Score}'_{\text{total}}$	$\text{Score}'_{\text{ch}}$	$\text{Score}'_{\text{gen}}$
GPT-3.5-turbo	**0.64**	**0.55**	**0.73**	**0.66**	**0.54**	**0.78**
Llama-3-youko-8b	0.47	0.50	0.43	0.34	0.31	0.36
Weblab-10b-instruction-sft	0.29	0.29	0.30	0.15	0.10	0.19
Tanuki-8x8B-dpo-v1.0 (GENIAC たぬき)	**0.45**	**0.40**	0.50	**0.50**	0.27	**0.73**
Tanuki-8B-dpo-v1.0 (GENIAC たぬき)	0.44	0.32	**0.56**	0.44	0.16	0.72
team_hatakeyama_phase1 (GENIAC たぬき)	0.42	0.37	0.48	0.42	0.37	0.46
team_kawagoshi_submit (GENIAC ビジネス)	0.40	0.36	0.45	0.38	**0.40**	0.35
team_ozaki_submit (GENIAC 天元突破)	0.35	0.28	0.41	0.35	0.38	0.31
team_kumagai_submit (GENIAC 甲)	0.31	0.27	0.35	0.27	0.18	0.36
team_kumagai_submit (GENIAC Kuma)	0.31	0.31	0.32	0.21	0.23	0.19
team_nakamura_submit (GENIAC JINIAC)	0.25	0.24	0.26	0.11	0.09	0.13
team_sannai (GENIAC Zoo)	0.11	0.01	0.21	0.09	0.07	0.10

4.3　Nejumi Leaderboard Neo との比較

GENIAC コンペティションでは，Nejumi Leaderboard Neo[7] の評価結果も加味した．Nejumi Leaderboard Neo は，llm-jp-eval（一問一答）1200 サンプルの 0shot スコア $\text{Score}'_{\text{ch}}$ と Japanese MT-Bench[20]（文章生成）80 サンプルの 0shot スコア $\text{Score}'_{\text{gen}}$ の平均をとった $\text{Score}'_{\text{total}}$ によって評価を行う．

表 3 に，JAMSE と Nejumi Leaderboard Neo による評価結果を示す．ほとんどのモデルで評価結果に相関があり，Score_{ch}，$\text{Score}_{\text{gen}}$ が高いモデルは，$\text{Score}'_{\text{ch}}$，$\text{Score}'_{\text{gen}}$ も高い傾向にあることが分かる．この傾向に反して，Tanuki-8x8B-dpo-v1.0 と Tanuki-8B-dpo-v1.0 は，$\text{Score}'_{\text{ch}}$ が低く，$\text{Score}'_{\text{gen}}$ が高くなったが，これは，モデル開発者が一般的知識の学習よりも指示追従を優先させたためであると考えられる[8]．したがって，JAMSE は日本語ベンチマークとして，Nejumi Leaderboard Neo と同程度の評価性能を有すると考えられる．

5　おわりに

本稿では，高品質かつ少数のサンプルで構成された日本語ベンチマーク JAMSE を提案した．JAMSE による評価は，Nejumi Leaderboard Neo による評価と強く相関しており，日本語 LLM の言語理解能力および回答生成能力を評価するベンチマークとして有用である．また，評価に要した時間は，ほとんどのモデルで数十分 ~2 時間程度であり，計算コストを極めて小さく抑えられた点でも価値がある．

一方で，評価ベンチマークは LLM の性能向上に合わせて日々刷新されている．本研究で参考にした Open LLM Leaderboard v1 は v2[21] に更新され，より複雑で広い知識をカバーしたベンチマークが採用されている．Nejumi Leaderboard 3 [9] も公開されており，新たに安全性能やドメイン特化性能（未実装）が評価に組み込まれている．

JAMSE の作成から半年以上が経過しているため，特長を維持しつつ，日本語 LLM の性能向上に合わせた拡張・改良を今後の課題としたい．

7）https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0
8）https://zenn.dev/matsuolab/articles/377f7ae8b1169e
9）https://wandb.ai/wandb-japan/llm-leaderboard3/reports/Nejumi-LLM-3--Vmlldzo3OTg2NjM2

謝辞

この成果は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の助成事業「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の結果得られたものである．

参考文献

[1] 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.

[2] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

[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 **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[4] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023.

[5] Inc Preferred Networks. Plamo-13b, 2023.

[6] Han Namgi, 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, Chen Bowen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会 (NLP2024), 3 2024.

[7] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-tasks-100: 日本語 instruction モデル評価データセット, 2023.

[8] Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1576–1601, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[9] Frederic M Lord and Melvin R Novick. **Statistical theories of mental test scores**. IAP, 2008.

[10] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples, 2024.

[11] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard (2023-2024). https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard, 2023.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[13] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[15] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance, 2024.

[16] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[17] 中村友亮, 河原大輔. 日本語 truthfulqa の構築. 言語処理学会第 30 回年次大会, pp. 1709–1714, 2024.

[18] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[21] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.