Japanese MT-bench++: より自然なマルチターン対話設定の日本語大規模ベンチマーク

植松拓也 [1]　福田創 [1]　河原大輔 [1]　柴田知秀 [2]
[1] 早稲田大学　[2]LINE ヤフー株式会社
{takuya1009@akane.,so.fukuda@akane.,dkw@}waseda.jp　tomshiba@lycorp.co.jp

概要

大規模言語モデル (LLM) の能力を網羅的に評価するのは大変に難しい課題である。LLM のベンチマークの一つに、マルチターンの対話的タスク遂行能力を評価する MT-bench があり、日本文化に合うように改編された Japanese MT-bench も構築されている。しかし、これらのデータセットは 80 問と小規模であることと、2 ターン目の質問が 1 ターン目の回答に依存していないという問題がある。我々はクラウドソーシングを用いることにより、5,000 問程度にまで大規模化し、より広範に評価を行えるベンチマークを構築する。1 ターン目の質問はワーカー、回答はワーカーと LLM によって作成することにより、多様な回答を得る。2 ターン目の質問を作成する際は 1 ターン目の各回答に対して作成し、より自然な対話設定となるようにする。

1　はじめに

大規模言語モデル (LLM) は、人間のアシスタントとしてさまざまなタスクや問題解決に利用されている。これらは多くの場合、人間と LLM の間でマルチターンにわたる対話を行うことによって成し遂げられる。LLM におけるこのようなマルチターンの対話的タスク遂行能力を評価するために、MT-bench [1] に代表されるベンチマークが構築、利用されている。MT-bench は 2 ターンからなる 80 対話について LLM の回答を評価するベンチマークである。日本文化に合うように改編された Japanese MT-bench[1)] も日本語における LLM の評価で用いられている。

しかし、これらのベンチマークは 2 ターン目の質問が 1 ターン目の回答に依存していないという問題と、80 対話からなり小規模であるという問題がある。このため、トピックは同じであるが関連度の低いシングルターンが連続していることが多く、また小規模であるために、広範なトピックや観点からのマルチターン能力を正確に測ることができない。

我々はクラウドソーシングを用いることにより、約 5,000 対話に大規模化し、より広範な評価を行えるベンチマークである Japanese MT-bench++を構築する。2 ターン目の質問は 1 ターン目の回答に対して作成し、より自然な対話設定となるようにする。各ターンの質問はワーカー、回答はワーカーと LLM によって作成することにより多様な対話を得る。

さらに、構築した Japanese MT-bench++を用いて、さまざまな LLM についてマルチターン対話における性能を比較し、結果を分析する。

2　関連研究

MT-bench は、Writing、Roleplay、Reasoning、STEM、Humanities、Math、Coding、Extraction の 8 つのカテゴリをもち、カテゴリごとに 10 件のマルチターン対話から構成される。各対話はユーザーとアシスタントの 2 ターンの対話を想定しており、ユーザーの 2 ターン分の質問や指示が含まれる。アシスタントの回答は含まれておらず、評価対象の LLM が生成し、2 ターン目の回答について LLM-as-a-Judge で評価する設定である。LLM-as-a-Judge とは GPT-4 などの強力な LLM で回答を自動評価する枠組みである。MT-bench は日本語版、中国語版などの他言語版も作成されている。MT-bench はさまざまな言語で広く利用されているが、1 節で述べた問題がある。

MT-bench と比較して、大規模かつ自然な対話設定であるマルチターン対話ベンチマークとして、Baize [2]、UltraChat [3]、MT-Eval [4]、MT-bench101 [5] などがある。ただし、これらのベンチマークにおける質問や指示は LLM が生成しているため、省略の使用方法などにおいて、人間の質問や指示を必ずし

1) https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge

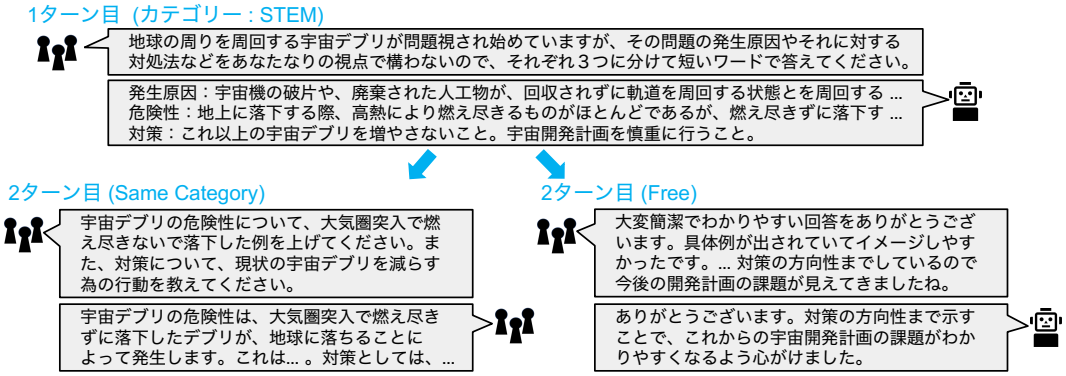

図 1　Japanese MT-bench++の例

も模倣できているわけではない。

上記の問題に対し、MT-bench++ [6] では人間が行う質問や指示を生成できるモデルを学習し、そのモデルが生成した質問を含む 4 万件のマルチターン会話からなるベンチマークを構築している。日本語にはこのようなベンチマークは存在しないが、本研究の Japanese MT-bench++では、これまでの対話の流れを見た人間が質問・指示を作成するため、マルチターン対話として自然であるといえる。

3　Japanese MT-bench++の構築

3.1　Japanese MT-bench++の構成

Japanese MT-bench++は、クラウドソーシング[2]を利用して大規模に構築する。MT-bench の 8 カテゴリのうちワーカーにとって作成の難易度が高い Math、Coding、Extraction[3]を除く 5 カテゴリを対象とする。MT-bench と同様、2 ターンで構成するが、Japanese MT-bench++ではユーザーの質問だけでなく、アシスタントの回答も含める。ユーザーとアシスタントの発話を収集し、各ターンのアシスタント回答に対し人間と GPT-4[4]が評価しスコアを付ける。図 1 に Japanese MT-bench++の例を示す。2 ターン目については Same Category と Free の 2 種類を収集するが、これは次節で説明する。MT-bench と同様、LLM が生成した 2 ターン目の回答を LLM-as-a-Judge で評価する (評価結果は 5 節で述べる)。

3.2　発話の収集方法

各ターンにおいて、ユーザー発話の作成者はワーカーであり、アシスタント発話の作成者はワーカーと LLM[5]である。オリジナルの MT-bench における 2 ターン目のユーザー発話は、1 ターン目と同じカテゴリに関する質問や指示であるが、自然な対話では他のカテゴリに関する発話になる可能性もある。そこで、2 ターン目のユーザー発話では Same Category と Free の 2 種類を収集する。Same Category は、オリジナルの MT-bench と同じく、2 ターン目のユーザー発話が 1 ターン目のカテゴリと同じ質問・指示のみであるが、Free では、2 ターン目のユーザー発話が、他のカテゴリに関する質問・指示や、質問・指示でないものを許容する。これにより、多様な対話を収集することができる。

1 ターン目のユーザー発話は、指定したカテゴリに関する質問や指示を 50 文字以上で書いてもらう。1 ターン目のアシスタント発話は、1 ターン目のユーザー発話に対し回答を生成するように指示する[6]。

2 ターン目のユーザー発話を得る際、1 ターン目の対話をワーカーに与え、それを参照してユーザーとして質問や指示を書いてもらう。1 ターン目の回答が誤っている場合は、回答を修正するような発話を作成してもらう。2 ターン目のアシスタント発話を得る際、それまでの対話を与え、それを参照して回答を書いてもらう。

それぞれの発話をワーカーから得るために使用したインストラクションを付録 A に示す。

3.3　人間と GPT-4 によるスコア付け

GPT-4 の評価に関する分析や評価器の作成、3.4 節で軽量版を作成することを目的に、人間 (ワーカー) と GPT-4 が、1 ターン目と 2 ターン目のそれぞれの

2)　Yahoo!クラウドソーシングを用いた。
3)　Extraction カテゴリでは、ユーザの指示（例: 国名を抽出してください）の対象となるテキストが必要となり、ワーカーに適切に取得するように指示することが難しいと考えた。
4)　gpt-4-turbo-2024-04-09　(https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4)
5)　https://huggingface.co/cyberagent/calm2-7b-chat、https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-ac_001_16x-dolly-ichikara_004_001_single-oasst-oasst2-v2.0 を用いた。
6)　ワーカーのみ 50 文字以上で回答するように指示した。

表 1 各軽量版におけるカテゴリ別の 2 ターン目の回答スコアの平均

	human		roleplay		writing		STEM		reason	
	人間	GPT	人間	GPT	人間	GPT	人間	GPT	人間	GPT
easy	3.65	3.94	3.52	4.02	3.64	3.99	3.73	3.86	3.47	3.77
hard	2.82	2.94	2.83	3.33	3.00	3.33	3.03	2.92	2.69	2.48
lite	3.39	3.59	3.20	3.69	3.33	3.72	3.39	3.53	3.13	3.13

回答について評価しスコアを付ける。MT-bench の評価プロンプトを参考にし、「有用性」、「関連性」、「正確性」、「深み」、「創造性」、「詳細さ」を考慮して総合的に評価するように指示する。ワーカーが評価するため、MT-bench では 10 段階であるが複雑さを減らすため 1-5 の 5 段階で評価する。評価に使用したプロンプトを付録 B に示す。

3.4 軽量版の構築

構築した Japanese MT-bench++は大規模であるという長所があるものの、大規模ベンチマークを用いた評価では LLM による回答生成に時間がかかり、また、LLM-as-a-judge のコストもかかる。そこで、300 対話からなる軽量版を lite 版、easy 版、hard 版の 3 種類構築し、比較的低コストで評価できるようにする。lite 版はランダムに対話を抽出した。easy 版はモデルにとって簡単なベンチマークであり、モデルの 2 ターン目の回答に対する人手評価スコア (2 つの平均) が高い対話を抽出した。hard 版はモデルにとって難しいベンチマークであり、モデルの 2 ターン目の回答に対する人手評価スコア (2 つの平均) が低い対話を抽出した。各軽量版において、各カテゴリの対話数が同じになるように 60 対話ずつ抽出し、Same Category と Free についても均等になるように収集する。各軽量版について、2 ターン目のユーザー発話に対する回答の平均スコアを表 1 に示す。ただし、回答の平均スコアは各 2 ターン目のユーザー発話に対する 3 つの回答すべてのスコアを考慮して求めた。

4 分析

本節では、Japanese MT-bench++の統計情報や、人間と GPT-4 による評価間の相関、2 ターン目のユーザー発話における Same Category と Free の違いについて説明する。人間と LLM が生成した回答の平均長は付録 C に示す。

4.1 規模

Japanese MT-bench++は、5,262 対話からなる。カテゴリごとの対話数について表 2 に示す。

表 2 Japanese MT-bench++の規模

	human	roleplay	writing	STEM	reason
Same Category	615	378	570	420	297
Free	834	444	666	555	483

表 3 人間と LLM が作成した 1 ターン目の回答のスコア

	human		roleplay		writing		STEM		reason	
	人間	GPT	人間	GPT	人間	GPT	人間	GPT	人間	GPT
calm	3.53	3.60	3.38	3.74	3.56	4.01	3.54	3.60	2.69	2.58
llm-jp	3.43	3.40	3.12	3.51	3.32	3.83	3.41	3.29	2.80	2.40
人間	3.29	3.41	3.10	3.24	3.14	3.13	3.33	3.72	3.50	3.86

表 4 人間と LLM が作成した 2 ターン目の回答のスコア

	human		roleplay		writing		STEM		reason	
	人間	GPT	人間	GPT	人間	GPT	人間	GPT	人間	GPT
calm	3.41	3.69	3.30	4.08	3.48	4.20	3.58	3.78	3.05	3.02
llm-jp	3.40	3.69	3.19	3.98	3.37	3.62	3.53	3.68	3.02	2.85
人間	3.05	3.16	3.02	3.22	3.10	3.10	3.10	3.02	3.07	3.27

表 5 人間と GPT-4 による評価間の相関

回答作成者	ピアソン	スピアマン
calm	0.480	0.404
llm-jp	0.453	0.415
人間	0.564	0.545

4.2 人間と GPT-4 による評価

1 ターン目のアシスタント発話に対する評価結果を表 3、2 ターン目のアシスタント発話に対する評価結果を表 4 に示す。

writing カテゴリにおいて、LLM が生成した回答の方が人間が生成した回答よりもスコアが高くなっている。LLM は長い文章を書くことが得意であると考えられる。また、LLM が生成した回答に対し GPT-4 評価の方が人間評価よりもかなりスコアが高くなっており、GPT-4 評価では長さバイアスがあることが推測できる。reasoning カテゴリにおいて、LLM が生成したスコアが低くなっており、クイズや計算問題が苦手であることがわかる。ただ、reasoning カテゴリにおける LLM が生成した 2 ターンのアシスタント発話を見ると、スコアが向上している。1 ターン目の人間の回答、2 ターン目のユーザー発話を参考にして、2 ターン目の回答を改善したことが推測される。roleplay カテゴリにおいて、GPT-4 がつけたスコアの方が人間より高くなっている。人間は、「なりきれているか」だけでなく内容も厳しく評価していると推測される。

また、人間と GPT-4 による評価間の相関を表 5 に示す。人間と GPT-4 の相関は高くなく、GPT-4 による評価にはまだ課題が残る。また、人間と GPT-4 では重点を置く評価ポイントが異なる可能性がある。人間は情報量の多さも評価するポイントであるが、GPT-4 では真実性が高く評価される傾向にある。

表 6 Japanese MT-bench original ならびに lite/easy/hard 版において LLM が生成した 2 ターン目の回答に対するスコア

	Japanese MT-bench original						lite 版					
	human	roleplay	writing	STEM	reason	平均	human	roleplay	writing	STEM	reason	平均
calm3[a]	7.90	8.10	6.00	6.60	5.20	8.04	7.90	7.65	8.17	8.35	7.47	7.91
Swallow[b]	7.50	6.80	5.00	6.40	7.70	6.95	6.08	6.47	6.87	7.15	6.68	6.65
llm-jp-3[c]	7.90	7.60	6.50	6.20	5.30	5.72	6.62	6.67	6.98	6.77	5.53	6.51
Qwen2.5[d]	8.30	8.50	7.80	8.10	8.50	8.54	8.42	8.15	8.43	8.78	8.58	8.47
gemma-2[e]	8.10	8.20	7.50	7.70	6.90	7.75	7.73	7.70	8.10	8.45	7.87	7.97
claude[f]	8.50	8.70	8.40	8.30	8.80	8.74	8.28	8.12	8.38	8.48	8.63	8.38
GPT-4o[g]	8.10	8.50	8.00	8.00	8.50	8.49	8.52	8.08	8.28	8.75	8.48	8.42

	easy 版						hard 版					
	human	roleplay	writing	STEM	reason	平均	human	roleplay	writing	STEM	reason	平均
calm3	8.22	8.15	8.43	8.10	8.13	8.21	6.83	7.65	7.80	7.58	6.88	7.35
Swallow	7.15	6.65	6.82	7.08	6.88	6.92	6.12	6.65	6.48	7.32	6.23	6.56
llm-jp-3	7.08	6.85	7.33	6.95	6.85	7.01	5.32	6.45	6.28	6.43	4.50	5.80
Qwen2.5	8.60	8.40	8.57	8.83	8.60	8.60	7.43	8.23	8.25	8.92	8.70	8.31
gemma-2	8.18	7.98	8.23	8.15	8.10	8.13	6.53	7.60	7.75	8.08	7.78	7.55
claude	8.20	8.37	8.42	8.30	8.38	8.33	8.08	8.27	8.23	8.42	8.75	8.35
GPT-4o	8.45	8.23	8.42	8.75	8.45	8.46	8.25	8.13	8.18	8.72	8.63	8.38

[a] https://huggingface.co/cyberagent/calm3-22b-chat
[b] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1
[c] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct
[d] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[e] https://huggingface.co/google/gemma-2-27b-it
[f] claude-3-5-sonnet-20241022 (https://docs.anthropic.com/en/docs/about-claude/models#model-names)
[g] gpt-4o-2024-08-06 (https://platform.openai.com/docs/models#gpt-4o)

4.3 Free の分析

カテゴリによって他のカテゴリの質問・指示や質問・指示ではない発話の割合を分析すると reasoning と writing は約 3, 4 割だが、STEM は 1 割程度であった。カテゴリによっては、対話の流れと関係のない発話を急にはせず、1 ターン目と同じカテゴリの発話が自然であることがある。そのため、他のカテゴリに関する質問・指示を多く得られなかったと考えられる。多様さに関する分析は今後の課題とする。

5 Japanese MT-bench++を用いた LLM の評価

5.1 実験設定

3.4 節で構築した軽量版を用いて、calm2-7b-chat と llm-jp-13b 以外のモデルで、2 ターン目のユーザーからの指示に対する回答を生成し、GPT-4o で評価した。original Japanese MT-bench と比較するため、1-10 の 10 段階で評価した。評価するときに使用したプロンプトを付録 B に示す。モデルは、GPT-4o、claude-3-5-sonnet、calm3-22b-chat、Llama-3.1-Swallow-70B、llm-jp-3-13b、Qwen2.5-72B-Instruct、gemma-2-27b-it を用いた。

5.2 結果

結果を表 6 に示す。lite 版は、Japanese MT-bench original と比較すると、writing、STEM、reasoning カテゴリにおいてスコアが高くなっている。Japanese MT-bench++は、より自然な対話設定になっているので、2 ターン目の質問や指示に対して回答しやすくなっていると推測できる。また、easy 版と hard 版で比較すると、基本的に easy 版の方がスコアが高く、難易度に差のある軽量版を作成できた。

次にモデルごとに比較すると、calm3 や Swallow、llm-jp-3 は reasoning カテゴリにおいてスコアが低く、推論が苦手であるといえる。Qwen2.5 や claude、GPT-4o はどのカテゴリでも高いスコアを出しており、苦手分野があまりないと推測できる。

6 おわりに

本研究では、Japanese MT-bench++を構築した。クラウドソーシングを利用することで数千問程度まで大規模化した。2 ターン目の質問を作成する際は、1 ターン目の各回答を見て人間が作成することにより自然な対話設定になるようにした。構築したデータセットは公開予定であり、LLM の評価や分析、LLM 評価器の学習に利用することができ、日本語においてこれらの研究が促進されることが期待される。

謝辞

本研究は LINE ヤフー株式会社と早稲田大学の共同研究により実施した。

参考文献

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[2] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 6268–6278, Singapore, December 2023. Association for Computational Linguistics.

[3] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics.

[4] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. 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. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing multi-turn instruction following 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. 9729–9750, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

A ワーカーから発話を得るときのインストラクション

3.2 節において、ワーカーから発話を得るときのインストラクションを表 7 に、そのときに提示した注意点を表 8 に示す。

表 7 ワーカーから発話を得るときのインストラクション (カテゴリ STEM の場合)

1 ターン目のユーザー発話の取得時
科学・テクノロジー・工学など、理系分野に関する問題を考えてください。ただし、数学に関する問題は除いてください。
1 ターン目のアシスタント発話の取得時
科学、テクノロジー、工学など、理系分野に関連している問題が与えられるので、それに対する回答を作成してください。答えだけでなく、答えに関連した情報も追加することで、より詳細に回答してください。
2 ターン目のユーザー発話の取得時 (Same Category)
「問題」とそれに対する「回答」が与えられます。その「回答」を参照し、科学、テクノロジー、工学など、理系分野に関連している「問題」を追加してください。与えられた回答が間違っている場合は、回答を修正するように促すような問題を作成してください。
2 ターン目のユーザー発話の取得時 (Free)
ユーザーとアシスタントの対話の様子を提示します。あなたがユーザーとして、最後のアシスタントの発言に対し、返答を考え、会話を続けてください。ただし、アシスタントの発言が、最初のユーザーの指示に従っていない場合は、ユーザーの指示に従うよう促すような返答をしてください。
2 ターン目のアシスタント発話の取得時 (Same Category)
ユーザーとアシスタントの対話の様子を提示します。あなたがアシスタントとして、ユーザーから最後に与えられる問題に対し、回答を作成してください。
2 ターン目のアシスタント発話の取得時 (Free)
ユーザーとアシスタントの対話の様子を提示します。あなたがアシスタントとして、ユーザーから最後に与えられる発言に対し、つじつまの合う発言を作成してください。

表 8 ワーカーから発話を得るときに与えた注意点

・返答は、50 文字以上書いてください。
・また、「今日」「今年」など、答えるタイミングによって認識の異なる単語は含まないようにしてください。
・ChatGPT などの生成 AI の使用は不可です。
・字数が足りていない、無回答の場合には今後この一連のクラウドソーシングタスクが実施できなくなる可能性があります。

B 回答を評価するときのインストラクション・プロンプト

3.3 節において、ワーカーから評価のスコアを得るインストラクションを表 9 に、GPT-4 が回答を評価する際に使用したプロンプトを表 10 に示す (紙面の都合上、2 ターン目のアシスタント発話を評価するときのインストラクション・プロンプトのみ記載をしている)。

5 節において、モデルの性能を GPT-4o で評価した際に使用した、回答を評価するプロンプトを表 11 に示す。

C 人間と LLM の回答の平均長

人間と LLM が生成した 1 ターン目、2 ターン目のアシスタント発話の平均長について、それぞれ表 12、表 13 に示す。LLM の回答の方が、人間と比較して全てのカテゴリで長いが、人間もある程度長い回答を生成していることが確認できる。

表 9 ワーカーから評価のスコアを得るインストラクション

ユーザーとアシスタントの対話の様子を提示します。ユーザーから最後に与えられる質問 (Free の場合は「発言」) に対する、アシスタントの回答の質を 5 段階で評価し、その理由を書いてください。1 は回答の質が低い、5 は回答の質が高いとします。評価は、回答の「有用性」、「関連性」、「正確性」、「深み」、「創造性」、「詳細さ」を考慮して総合的に判定してください。 ユーザー： {first_user} アシスタント： {first_assistant} ユーザー： {second_user} アシスタント： {second_assistant}

表 10 GPT-4 が回答の評価する際のプロンプト

ユーザーとアシスタントの対話の様子を提示します。ユーザーから最後 (2 回目) に与えられる発言に対する、アシスタントの回答の質を 5 段階で評価し、数字のみを出力してください。1 は回答の質が低い、5 は回答の質が高いとします。評価は、回答の「有用性」、「関連性」、「正確性」、「深み」、「創造性」、「詳細さ」を考慮して総合的に判定してください。 [ユーザー] {first_user} [アシスタント] {first_assistant} [ユーザー] {second_user} [アシスタント] {second_assistant}

表 11 モデルが生成した 2 ターン目の回答を GPT-4o が評価する際のプロンプト

ユーザーとアシスタントの対話の様子を提示します。ユーザーから最後 (2 回目) に与えられる発言に対する、アシスタントの回答の質を 10 段階で評価し、数字のみを出力してください。1 は回答の質が低い、10 は回答の質が高いとします。評価は、回答の「有用性」、「関連性」、「正確性」、「深み」、「創造性」、「詳細さ」を考慮して総合的に判定してください。 [ユーザー] {first_user} [アシスタント] {first_assistant} [ユーザー] {second_user} [アシスタント] {second_assistant}

表 12 人間と LLM が作成した 1 ターン目の回答の平均長

回答作成者	human	roleplay	writing	STEM	reason
calm	367	418	521	426	172
llm-jp	330	290	461	330	228
人間	98.5	75.7	106	96.5	77.6

表 13 人間と LLM が作成した 2 ターン目の回答の平均長

回答作成者	human	roleplay	writing	STEM	reason
calm	293	291	385	355	226
llm-jp	395	330	461	472	205
人間	99.5	85.5	103	94.2	73.8