LLM を用いた対話印象評価による対話システム学習とその分析

吉田快 [1,2] 水上雅博 [3] 河野誠也 [2,1]
クルンカライ カナサイ [1] 杉山弘晃 [3] 吉野幸一郎 [4,2,1]
[1] 奈良先端科学技術大学院大学　[2] 理研ガーディアンロボットプロジェクト
[3]NTT コミュニケーション科学基礎研究所　[4] 東京科学大学
yoshida.kai.yf1@is.naist.jp,masahiro.mizukami@ntt.com, seiya.kawano@is.naist.jp
canasai.kruengkrai@riken.jp, h.sugi@ieee.org, yoshino.k.ai@m.titech.ac.jp

概要

RLAIF を対話システム学習に適応する上では、シングルターンの対話だけでなく、対話文脈全体の一貫性、個性、共感性などの対話印象を向上させる必要がある。本研究では、対話印象評価のための報酬モデルの比較、及びその報酬をフィードバックとしたシステムの対話印象の改善を行った。自動評価と人手評価の結果、個々の対話印象の向上だけでなく、応答の自然さも向上することが示された。

1　はじめに

近年の大規模言語モデル (Large Language Model; LLM) の発展により、多くの対話システムも LLM を主としたものに置き換わっているが、特定の目的に沿った対話システムの構築のためには、いまだに学習のためのデータ構築が必要となっている。そのような中で、教師データの構築を行わずに多様な目的で LLM の学習を行うことができる、Reinforcement Learning from AI Feedback (RLAIF) が注目を浴びている [1, 2, 3, 4, 5]。RLAIF では報酬モデルの用意ができれば、教師データの構築を無しに LLM の学習が可能になるという点で、対話システムへの応用が期待される技術である。

RLAIF の既存研究の多くは、生成されたテキストに対して一対一で与えられる評価を仮定している。例えば、Cheng らは推論と翻訳 [1] を目的として、Lee や Bai らは生成文からユーザにとって有害な文を取り除くこと [2, 3] を目的として、LLM によるフィードバックに基づいたモデル学習を行っている。これに対し、対話モデルに対して LLM や報酬モデルによるフィードバックを行おうとする場合、各発話応答の評価だけではなく、対話全体に対するユーザからの評価、つまり対話印象が重要な場合がある。対話全体に対する評価を AI モデルに行わせる取り組みは FED [6] や MEEP [7]、 INCAHARACTER [8]、LLM-Eval [9]、 G-Eval [10]、 LLM-as-a-Judge [11] などいくつか存在する。しかし、zero-shot, few-shot の prompting のみで対話全体に対し特定の観点に立った適切な評価を与えることは困難が伴う。

本研究では対話全体で得られる各対話印象を評価するための最適な戦略の調査のため、プロンプティングベースの手法と少量の対話評価データで教師あり学習（Supervised Fine-Tuning; SFT）を行ったモデルの 2 種類の報酬モデルの比較を行った。この結果、prompting のような報酬モデルに明示的な学習を行わないケースと比較して、SFT を行う方がより適切に対話印象を評価できることが確認された。また、これらの報酬モデルからの出力を向上させるよう対話モデルの適応を行った結果、応答の自然さと対話印象の改善が確認され、適応したモデルが自動評価、人手評価双方で高い評価を得た。

2　対話印象評価のための報酬モデル

2.1　対話評価タスクとデータ

今回対話全体への評価が付与された指標として、JTransformer-Eval [12] に付与されている 12 指標を用いる。JTransformer-Eval は Japanese-Dialog Transformer [12](システム) とユーザ (人間) の対話を対話全体に関わる 12 個の観点から評価付けしたデータセットである。12 個の評価指標の定義とデータの詳細は付録 A に示す。データセットには 1,600 件の対話セッションが含まれている。

2.2　報酬モデルの学習

報酬モデルは図 1 のように、対話コンテキスト C_i およびその応答 R_i とある対話印象指標 E_j を入力と

表 1 評価モデルの予測値と人手評価の相関

Metrics	w/o SFT	GPT-3.5	w/ SFT
Agency	0.0	0.21	**0.79**
Attentiveness	0.11	0.30	**0.85**
Consistency	0.14	0.40	**0.84**
Ease	0.07	0.29	**0.79**
Empathetic	0.10	0.41	**0.85**
Emotion	0.08	0.26	**0.82**
Enjoyability	0.03	0.24	**0.79**
Humanness	0.08	0.32	**0.77**
Personality	-0.06	0.27	**0.76**
Respeak	-	0.25	**0.85**
Topic	-0.15	0.33	**0.77**
Trust	-	0.18	**0.89**

して、対応したスコア S_{i,E_j} を評価するモデルである。i は各対話サンプルを識別するインデックスを表し、j は各評価指標を識別するインデックスを表す。ここでモデルを生成モデルではなく、回帰モデルとして学習をするために、モデルの最終層に線形層を追加している。また、1 つのモデルで 12 種類の指標を評価するように訓練を行う。詳しい学習設定については、付録 B に示す。

2.3 報酬モデルの評価

事前学習済みモデル (w/o SFT)、ChatGPT(gpt-3.5)、JTransformer-Eval を用いて 2.2 節で述べたチューニングを行ったモデル (w/ SFT) との比較を行う。報酬モデルの評価のために、事前に学習と検証、テスト用に 8:1:1 で分割をした JTransformer-Eval データのテストセットを用いる。

テストセットにおける人手評価ラベルと各評価モデルの推論値のスピアマンの相関を表 1 に示す。推論の際に 1 種類のラベルしか出力されなかったために相関の計算ができなかったものについては「-」で置き換えている。始めに、SFT 無しモデルは推論が特定のラベルに偏る傾向がみられ、かなり弱い相関であることが確認された。次に、gpt-3.5 は 0.2 から 0.4 程度の弱い相関を達成している。最後に SFT モデルは、全ての評価値において 0.8 前後の強い正の相関を達成していることが確認できた。以降ではこの SFT により学習されたモデルを報酬モデルとして、対話モデルのチューニングに用いる。

2.4 対話印象評価モデルの分析

人手評価と最も相関が強い Trust を対象に、対話印象評価モデルがどのような入力に対して高い評価を与えているのか分析を行う。分析のため、日本語感情表現辞書 (Japanese Linguistic Inquiry and Word Count; JIWC) [1] に定義される「信頼感を感じる」語を対象に、それらの対話印象評価への寄与度の計算をし、比較を行う。寄与度の計算では、機械学習モデルの解釈性における代表的なアルゴリズムである、Vanilla Gradients [13], Integrated Gradients [14], SHAP [15] によって入力トークンごとの報酬モデルの評価への寄与度を計算し、「信頼感を感じる語」と全トークンの平均値を比較する。

表 2 各トークンの AIF における貢献度

Metrics	全トークン	信頼感を感じる語
Vanilla Gradients	0.034	-0.052
Integrated Gradients	0.500	**1.462**
SHAP	0.001	**0.002**

2.4.1 対話印象評価への寄与度の比較

表 2 にトークンごとの寄与度を示す。Integrated Gradients、 SHAP ともに信頼感を感じる語が全トークンと比較して高い寄与度度を達成していることが確認できた。一方で、Vanilla Gradients は信頼感を感じる語の方が低い結果となった。そのため、評価モデルの教師あり学習において信頼感を感じる単語を高く評価できるように学習できており、フィードバックによってそれらが強調された応答が生成できることが期待される。

3 AI フィードバックによる対話システムの最適化

この章では、AIF による対話システム学習を実現するシステム概要、その中で用いる学習の方法、および評価方法について説明を行う。学習では、12 種類の評価値と 2 種類の学習手法の組み合わせ 24 種類をそれぞれ試し、学習結果を自動評価と人手評価で比較を行う。

3.1 システムの概要

本研究で実施する学習の概要を図 1 に示す。学習では、対話文脈 C_i と、C_i に対する LLM の応答 R_i を結合した新しい対話文脈に対して、ある対話印象指標 E_j に対応した報酬モデルを用いてスコア S_{i,E_j} を付与する。その後、それら (C_i, R_i, S_{i,E_j}) に基づきモデルを最適化アルゴリズムに則って更新を行う。

3.2 AI フィードバックによる LLM の学習

報酬モデルの出力を基に LLM を学習するための最適化の手法として、Proximal Policy Optimization

1) https://github.com/sociocom/JIWC-Dictionary

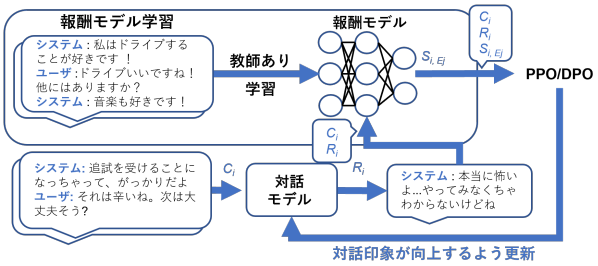

図 1 報酬モデルの学習及び、それを活用した PPO と DPO の実施, C_i は対話履歴, R_i は生成された応答文, S_{i,E_j} は評価対象の対話印象 E_j に対応した評価値を表す

(PPO) [16] と Direct Preference Optimization (DPO) [17] の 2 手法を用いる。DPO の前処理では、対話履歴に対して更新対象のモデルで 2 種類の応答を生成し、それら 2 つの生成文と対話履歴を報酬モデルで評価する。その後、評価値の高い生成文を Accepted、低い方を Rejected として学習に用いる。

学習では RLAIF の汎化性能の確認のため、cyberagent/calm2-7b-chat [2] (calm) と rinna/youri-7b-chat [3] (youri) の 2 種類のモデルを使用する。そして 2 種類の最適化手法と 2 種類のモデル、12 種類の評価値をそれぞれ組み合わせた、48 種類の組み合わせについて学習を行った。学習は、A100 80GB を 8 個用いて、デバイスあたりの batch size を 32 で行い、PPO の epoch はオリジナルの RLHF [18] と同じ 2 を、DPO では 120 を採用して学習を行った。ただし、DPO の最終的なモデルは training loss が最も低い物を選択した。

4 実験設定

4.1 データセットの説明

JEmpatheticDialogues [12] はシステムと人間の 4 ターンの対話が 20,000 件収録されたデータである。強化学習の実施のため、データを訓練と検証、テストに 8:1:1 で分割を行い利用した。

4.2 評価指標

自動評価では、AI フィードバックを基にした学習によって LLM の生成文が学習前モデルと比較して流暢になったか、適応対象とした対話印象の評価値 (S_{i,E_j}; AIF Score) が反映されているかについてそれぞれ評価を行う。流暢性の評価のためには Perplexity (PPL) を、評価値の反映の確認のためには事前に学習した評価モデルを活用する。

4.2.1 人手評価指標

人手評価では、応答が対話履歴に対して自然であったかと、適応対象の対話印象が反映されているかの 2 つの側面で評価を行う。自然性は 5 段階 (5 が最も良く、1 が最も悪い) で評価を行い、評価値の反映度は評価値ごとに学習前と PPO、DPO の 3 種類の生成文をそれぞれ提示し、それらの順位付け (1 が最もよく、3 が最も悪い) を行う。ただし、順位付けは同じ順位を許可している。評価者に提示したインストラクションは付録 D に示す。

5 実験結果

自動評価の結果を表 3 に示す。calm と youri では AIF と PPL ともに DPO が最も良い結果を示した。PPO では AIF が学習前と比較して、わずかな上昇、もしくは減少となっている。一方で、DPO の AIF は平均 0.5 程度の改善に成功している。PPL に関しては、calm は全ての報酬において DPO が最も良い値を達成しているが、youri ではいくつかの指標で学習前の方が良い値となっていた。

次に人手による対話印象の評価結果を表 4 に、自然性の評価結果を表 5 に示す。自然性について、calm では DPO が最も優れており、youri では学習前と DPO が同程度に優れた結果を示している。また、PPO と DPO ともに学習前と比べて自然性が向上していることから RLAIF によって評価値の反映だけではなく自然な応答の学習が可能であることが示唆されている。Rank と Win については calm と youri のどちらも DPO が最も多く良い結果を出している。自動評価と人手評価の結果から、応答の自然性と評価値の反映には DPO が最も優れていることがわかる。

6 議論

PPO と DPO ともに、損失関数に事前学習時のクロスエントロピー誤差とは異なるものを用いる。DPO はベースモデルの生成文に対してモデルを適応するものであるため PPL が改善されるのに対して、PPO は報酬モデルに適応するものであるという観点から PPL が悪化したと考えられる。また PPO は学習前と比較して PPL が悪化し、AIF の値があまり増加しなかった一方で、人手評価では自然性と

2) https://huggingface.co/cyberagent/calm2-7b-chat
3) https://huggingface.co/rinna/youri-7b-chat

表 3　自動評価結果、 AIF は 11 段階の対話印象の評価値を表す

	calm						youri					
	w/o tuning		PPO		DPO		w/o tuning		PPO		DPO	
	AIF	PPL	AIF	PPL	AIF	PPL	AIF	PPL	AIF	PPL	AIF	PPL
Agency	5.74	42.12	5.74	42.83	**6.50**	**36.74**	5.43	25.56	5.29	28.40	**5.91**	**23.36**
Attentiveness	5.03	39.72	5.03	41.32	**5.66**	**33.98**	4.85	23.57	4.76	46.75	**5.17**	**23.43**
Consistency	4.38	39.01	4.38	41.62	**4.93**	**31.94**	4.01	22.65	4.04	63.92	**4.46**	**22.29**
Ease	6.71	38.74	6.70	41.46	**7.24**	**34.21**	6.60	**22.28**	6.22	913.52	**6.96**	22.94
Empathetic	4.77	38.62	4.80	41.48	**5.17**	**30.41**	4.82	22.14	4.85	40.98	**5.13**	**19.22**
Emotion	4.62	38.56	4.62	40.60	**5.32**	**31.67**	4.44	22.06	4.54	50.25	**5.02**	**16.77**
Enjoyability	5.32	38.55	5.33	41.40	**5.93**	**29.73**	5.21	**21.98**	5.18	20.96	**5.54**	23.92
Humanness	6.31	38.55	6.34	42.44	**6.80**	**33.34**	6.18	21.96	6.10	41.93	**6.48**	**20.85**
Personality	6.00	38.55	6.02	41.10	**6.59**	**31.93**	5.83	21.95	5.82	35.61	**6.37**	**12.53**
Respeak	4.81	38.55	4.85	40.54	**5.46**	**32.51**	4.68	21.95	4.65	23.97	**5.10**	**20.59**
Topic	4.92	38.55	4.92	42.14	**5.53**	**34.44**	4.75	**21.95**	4.93	37.61	**5.23**	26.91
Trust	4.29	38.55	4.29	41.85	**4.76**	**31.32**	4.07	21.95	4.07	32.60	**4.47**	**19.04**

表 4　対話印象の反映度の順位付け評価結果, Rank は 3 段階の順位、Win は学習前と比較して同順以上である割合を表す

	calm					youri				
	w/o tuning	PPO		DPO		w/o tuning	PPO		DPO	
	Rank	Rank	Win	Rank	Win	Rank	Rank	Win	Rank	Win
Agency	1.76	**1.75**	**0.59**	1.84	0.57	1.62	**1.58**	**0.69**	1.59	0.68
Attentiveness	1.95	1.85	0.65	**1.55**	**0.71**	2.01	1.84	0.64	**1.5**	**0.78**
Consistency	1.91	1.85	0.61	**1.69**	**0.67**	**1.48**	1.66	0.65	1.5	**0.68**
Ease	1.74	**1.68**	**0.58**	2.01	0.47	**1.73**	2.0	0.45	1.9	**0.52**
Empathetic	2.19	1.84	0.69	**1.44**	**0.84**	1.97	1.8	0.65	**1.63**	**0.71**
Emotion	2.07	1.82	0.66	**1.47**	**0.78**	1.8	1.79	0.64	**1.68**	**0.65**
Enjoyability	2.33	1.82	0.74	**1.57**	**0.79**	2.0	**1.76**	**0.68**	**1.76**	0.6
Humanness	1.99	**1.73**	**0.61**	1.91	0.58	1.7	2.26	0.39	**1.49**	**0.64**
Personality	2.14	1.82	0.73	**1.36**	**0.83**	2.01	1.71	0.68	**1.58**	**0.71**
Respeak	2.15	**1.79**	0.65	1.8	**0.66**	1.87	1.82	0.63	**1.51**	**0.76**
Topic	1.97	**1.82**	**0.61**	1.86	0.59	1.77	1.85	0.6	**1.36**	**0.79**
Trust	2.1	1.87	0.66	**1.56**	**0.76**	1.9	1.79	0.64	**1.54**	**0.74**

表 5　5 段階の応答の自然性の人手評価結果

	calm			youri		
	w/o tuning	PPO	DPO	w/o tuning	PPO	DPO
Agency	2.48	2.49	**2.85**	2.12	**2.25**	1.96
Attentiveness	2.45	**2.66**	2.64	2.33	1.67	**2.44**
Consistency	2.42	**2.68**	**2.6**	1.91	1.52	**1.92**
Ease	2.48	2.37	**2.61**	1.97	1.77	1.68
Empathetic	2.39	**2.75**	**3.14**	1.93	1.47	**1.97**
Emotion	2.46	2.51	**2.92**	**2.03**	1.68	1.99
Enjoyability	2.48	**2.63**	**2.63**	1.95	1.71	1.89
Humanness	2.43	2.36	**2.6**	**2.02**	1.8	1.98
Personality	2.48	**2.71**	2.51	**2.08**	1.7	**2.09**
Respeak	2.53	2.58	**2.86**	2.26	**2.57**	2.36
Topic	**2.5**	2.42	2.44	**2.13**	1.54	1.77
Trust	2.49	**2.78**	**3.08**	2.17	1.91	**2.46**

Win において学習前の評価を上回った。そのため DPO に劣るものの応答の自然性と評価値の学習は可能であることが示唆されている。

AIF による評価では、応答として自然ではあるが Emotion や Topic などの評価値があまり反映されていない応答であっても、ある程度自然であれば高い評価を得てしまうことが確認された。これらは、「はい、そうですね」や「いいと思います」というような Dull Response と呼ばれるものが多かった。特に、PPO では学習を進めるにつれてモデルの出力が Dull Response に偏る現象が確認された。これは、JTransformer-Eval の対話印象評価において、自然性も暗黙的に各評価として内包されてしまっているのが原因であると考えられる。以上のことから、よりよい対話モデルの適応のためには報酬とした評価値をより純粋に評価する方法と、生成文の多様性を考慮することが重要であると考えられる。

7　まとめ

本研究では、LLM をベースに SFT した報酬モデルを用いて対話全体に対する対話印象を評価させ、その評価値を基に LLM をベースとした対話モデルの適応を行うことで、個々の対話印象の評価指標に対応した対話モデル適応が可能か検証を行った。自動評価と人手評価では、DPO による学習が最も良い結果を示し、学習によって報酬とした評価値だけでなく生成文の自然性も向上することが確認された。一方で、Dull Response のような本来好ましくない応答であっても、それが自然な応答であれば評価モデルが高い評価を与えるなどの課題が確認されたため、それらの対策が必要となることも判明した。

謝辞

本研究の 2 章は NTT コミュニケーション科学基礎研究所と理化学研究所の共同研究によるものです。

参考文献

[1] Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. Language model self-improvement by reinforcement learning contemplation, 2023. arXiv:2305.14483.

[2] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2023. arXiv:2309.00267.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, and Amanda Askell et al. Constitutional ai: Harmlessness from ai feedback, 2022. arXiv:2212.08073.

[4] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[5] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, and Aaron T Parisi et al. Beyond human data: Scaling self-training for problem-solving with language models. **Transactions on Machine Learning Research**, 2024. Expert Certification.

[6] Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, **Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 225–235, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[7] Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 2078–2100, Singapore, December 2023. Association for Computational Linguistics.

[8] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. In-Character: Evaluating personality fidelity in role-playing agents through psychological interviews. 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. 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Yen-Ting Lin and Yun-Nung Chen. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In Yun-Nung Chen and Abhinav Rastogi, editors, **Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)**, pp. 47–58, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[11] 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. arXiv:2306.05685.

[12] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In **2022 IEEE Spoken Language Technology Workshop (SLT)**, pp. 685–691, 2023.

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

[14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[15] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems 30**, pp. 4765–4774. Curran Associates, Inc., 2017.

[16] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. arXiv:1707.06347.

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. arXiv:2305.18290.

[18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

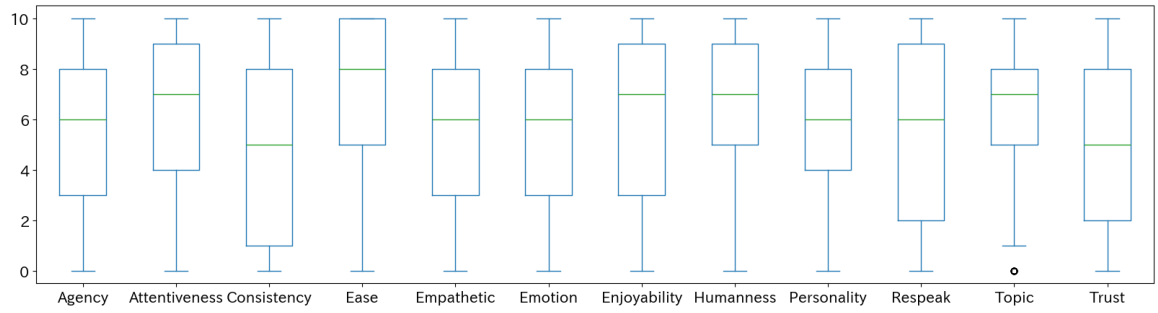

図 2　対話印象評価値の分布

表 6　評価値とその提示文

Metric name	Questionnaire
Agency	システムは自身の考えをもって話していると感じた
Attentiveness	システムはあなたに興味を持って積極的に話そうとしていた
Consistency	システムの発話は矛盾せず一貫していた
Ease	簡単に対話を続けることができた
Empathetic	システムの発話に共感できた
Emotion	システムは感情を持っていると感じた
Enjoyability	システムとの対話は楽しかった
Humanness	システムの発話は人間らしく自然だった
Personality	システムの個性・人となりが感じられた
Respeak	またこのシステムと話したい
Topic	システムには話したい話題があると感じた
Trust	システムの話したことは信頼できると感じた

A　JTransformer-Eval の詳細

12 個の評価指標の定義は表 6 の通りである。各対話印象評価値の分布を図 2 に示す。

次に、各対話印象評価値の相関を図 3 に示す。Respeak, Humanness, Ease, Enjoyability, Empathetic がそれぞれ相関が 0.8 以上であるなどかなり高い相関があることが確認できることから、これらが総合的な対話印象へと寄与していることが考えられる。

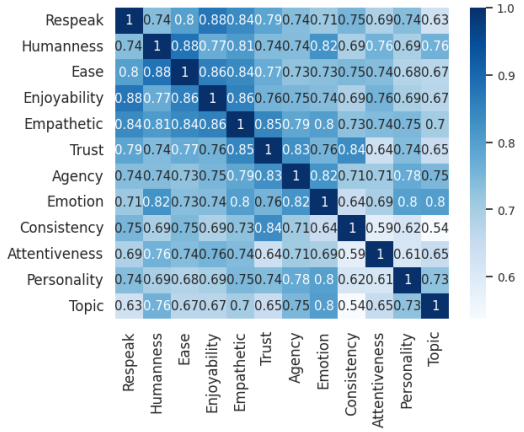

図 3　対話印象ごとの相関

B　評価モデルの学習設定

学習では、損失関数に Mean Squared Error を使用し、epoch が 10、デバイス当たりの batch size を 16 とし A100 80GB を 8 個使用した。そのため、最終的な batch size は 128 となっている。ベースモデルには tokyottech/swallow-7b-instruct-hf [4] を使用した。最終的なモデルは validation loss によって選択を行った。

C　事前学習済みモデルの推論方法

事前学習済みモデルの推論は式 (1) のように、評価プロンプトに対するスコアラベルの生成確率に各スコアラベルを掛けたものの平均を用いた。

$$AIF = \frac{1}{11}\sum_{i=0}^{10} i \times P(C|i) \tag{1}$$

D　人手評価の指示文

評価者に提示したインストラクションはそれぞれ次の通りである。

自然性の評価

3 種類の応答について、応答が会話履歴に対して自然な応答であったかをそれぞれ 5 段階で評価してください。1 から 5 の目安は以下の通りです。5 自然、4 やや自然、3 どちらともいえない、2 やや不自然、1 明らかに不自然

対話印象の評価値の反映度の評価

表 6 の各評価値について質問を基に、もっともそう感じたものから順に 1 から始まる順位を付けてください。同じ程度に感じたものは同順でよいです。この評価では、発話の自然さや流暢さは考慮せず報酬ごとの評価基準にのみ基づいて評価を行ってください。

このインストラクションに従って、1 人の評価者が各評価指標ごとに 100 件の評価データに対して評価を行った。

4)　https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-hf