擬似選好チューニングによる対話応答のペルソナ一貫性向上

高山 隼矢　大萩 雅也　水本 智也　吉川 克正
SB Intuitions 株式会社
{junya.takayama,masaya.ohagi,tomoya.mizumoto,katsumasa.yoshikawa}@sbintuitions.co.jp

概要

本稿では、擬似的な選好データを用いた選好チューニングによって、対話応答生成におけるペルソナと生成応答の間の一貫性を向上させる手法を提案する。提案手法では、あるペルソナ情報付き対話データに対して、他の対話からランダムに抽出したペルソナ情報を用いて応答を生成し、これらの応答を擬似負例として扱う。参照応答を正例として扱うことで、擬似的な選好データを作成する。実験では、擬似選好データを用いて選好チューニングを行ったモデルが、通常の教師あり学習のみを用いたモデルや、ペルソナと発話間の含意関係に基づく報酬によって強化学習したモデルと比較して、より一貫性のある自然な応答を生成することを示す。

1　はじめに

対話応答生成モデルにとって、ペルソナと一貫した応答を生成する能力は重要である [1]。既存研究 [2, 3] では、Dialogue Natural Language Inference (Dialogue-NLI) データセット [2] のような、ペルソナ情報と発話間の含意関係をアノテートした言語資源を活用することで、ペルソナ一貫性の向上に取り組んでいる。このようなデータセットは、応答のリランキングや強化学習に基づくペルソナ一貫性向上を実現する。しかし、含意関係アノテーションには多大なコストを要するため、同様の言語資源が存在しない言語やドメインへの適用は困難である。

本研究では、Dialogue-NLI などの追加の外部資源に依存せずにペルソナ一貫性を向上させるための、擬似的な選好チューニング手法を提案する。選好チューニング（Preference Tuning）とは、好ましい応答と好ましくない応答のペアからなる選好データを用いて、モデルがより好ましい応答を高い確率で生成するように訓練する枠組みである。提案手法では、ペルソナ対話データをシードとして擬似的な選好データを作成する。図 1 に示すように、ある対話

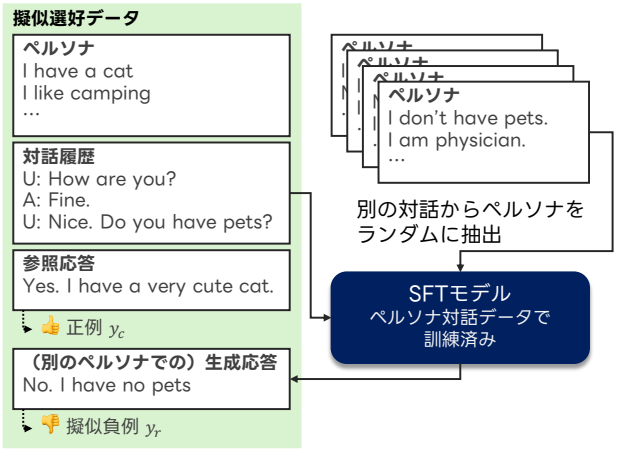

図 1　擬似選好データの作成方法の概要図

に対し、他の対話からランダムに抽出したペルソナ情報に基づいて生成された応答を負例として扱うことで擬似的な選好データを構築する。参照応答は正例として使用する。

実験では、選好チューニングの手法としてよく用いられる Direct Preference Optimization (DPO) [4] を複数の事前学習モデルに適用することで提案手法の有効性を調査した。実験結果を通じて、提案手法が従来の教師ありチューニング手法や、外部 NLI データに依存する強化学習手法に比してより高いペルソナ一貫性を達成することを明らかにした。また、提案手法によって訓練したモデルが、ペルソナの一貫性および自然さをより高く示す応答を生成することも確認した。本研究の成果は、対話システムにおけるペルソナ一貫性を向上させるための、スケーラブルかつ効果の高い解決策を提供する。

2　予備知識: 選好チューニング

選好チューニングとは、大規模言語モデル（Large Language Models; LLM）の振る舞いを人間の選好に合わせる手法群の総称である。基本的な方法のひとつとして、Reinforcement Learning from Human Feedback (RLHF) [5, 6] と呼ばれる、人間からのフィードバックに基づいてモデルを強化学習する

方法がある。RLHF では通常、モデルの最適化のためのアルゴリズムとして Proximal Policy Optimization (PPO) [7] と呼ばれる、報酬モデルを必要とする強化学習手法が採用される。報酬モデルは、人間にとってより好ましい応答（正例）とそうでない応答（負例）のペアからなるデータを使用して、より好ましい応答に高いスコアを出力するように訓練される。この報酬を最大化するように訓練することで、LLM がより好ましい応答を生成するようになる。

最近では、Direct Preference Optimization (DPO) [4] が PPO による選好チューニングの代替としてよく用いられる。DPO では選好データのペアを用いて LLM を直接最適化する。報酬モデルを必要としないため、PPO と比べて簡易かつ安定的に訓練を行うことができる。我々の手法は選好データを人手を介さず擬似的に構築するもので、DPO を含むさまざまな選好チューニング手法に適用可能である。

3 ペルソナ一貫性向上のための擬似選好チューニング

提案手法における選好データの作成方法の概要を図 1 に示す。まず、事前学習済みモデルをペルソナ対話データに基づく標準的な教師ありファインチューニング (Supervised Fine-Tuning; SFT) で微調整する。訓練データセットは次のように表される：$D = \{(p_i, x_i, y_i)\}_{i=1}^{N}$。ここで、$p_i$ はペルソナ情報、x_i は対話履歴、y_i は参照応答である。SFT を行ったモデルは p_i および x_i を入力として受け取り、応答 y'_i を生成する。この段階のモデルを SFT モデルと呼び、π_{sft} と表記する。

次に、擬似負例を生成するため、別の対話からランダムに選択したペルソナ情報 $p_j (i \neq j)$ で元のペルソナ情報を置き換え、SFT モデルを用いて擬似負例応答 y'^{neg}_i を次のように生成する：

$$y'^{\text{neg}}_i \sim \pi_{\text{sft}}\left(y' | p_j, x_i\right)$$

参照応答 y_i を正例として採用することで、擬似選好データ $l_i = (p_i, x_i, y_i, y'^{\text{neg}}_i)$ を獲得する。

最後に擬似選好データを用いて SFT モデルを更に選好チューニングする。DPO を使用する場合、以下の損失関数によってモデル π_{dpo} を得る:

$$\log \sigma \left(\beta \log \frac{\pi_{\text{dpo}}(y_i|p_i,x_i)}{\pi_{\text{sft}}(y_i|p_i,x_i)} - \beta \log \frac{\pi_{\text{dpo}}\left(y'^{\text{neg}}_i|p_i,x_i\right)}{\pi_{\text{sft}}\left(y'^{\text{neg}}_i|p_i,x_i\right)} \right).$$

ここで、β はハイパーパラメータ、σ はシグモイド関数である。

4 実験

提案手法の有効性の検証のため、日本語および英語のペルソナ対話データを用いて実験を行う。自動評価および人手評価の両方を実施する。

4.1 実験設定

データセット 日本語の実験には JPersonaChat データセット [8] を使用する。8 : 1 : 1 の割合でランダムに分割し、それぞれ訓練/検証/テストデータとする。英語では、PersonaChat [1] を使用する。訓練/検証/テストデータの分割の設定は元論文に従う。また、ペルソナ一貫性の自動評価（4.4 節）のために、Dialogue-NLI の評価セット [2] を用いる。Dialogue-NLI の訓練データは、比較対象の強化学習手法のための報酬モデルの訓練のために使用する。

比較モデル 一般性の検証のために複数のサイズ・種類の事前学習済みモデルを用いる。日本語向けには、japanese-gpt (medium 361M, 1B) [9]、swallow (7B, 13B) [10]、および sarashina2 (7B, 13B) を使用する。英語向けには、gpt2-medium (380M)、qwen2 (1.5B, 7B) [11]、mistral (7B) [12]、および llama-2 (7B, 13B) [13] を使用する。詳細は付録 A を参照のこと。

ベースラインとして、それぞれの事前学習モデルに対してペルソナ対話データでの SFT のみを施したモデル（SFT モデル）を採用する。また、Song ら [14] の報酬設計を参考に構築した Dialogue-NLI ベース報酬モデルを用いた強化学習モデルとの比較も行う。

訓練設定 SFT モデルの構築においては、最大 5 エポックまで訓練を行い、検証データでの損失が最小となるモデルを評価用に選択した。DPO は最大 3 エポックまで実施し、検証データでの精度が最も高いモデルを選択した。強化学習モデルでは、PPO アルゴリズムを用いて最大 2 エポックまで訓練を行い、訓練中に報酬が最大となるモデルを選択した。

4.2 LLM によるペアワイズ評価

ペルソナ一貫性と自然さを総合的に評価するために、LLM によるペアワイズ評価を実施した。ペアワイズ評価では、ある任意の 2 つのモデルの出力を比較し、どちらの応答が優れているかを LLM が判断する。この方法は LLM のタスク達成能力を測るベンチマークにおいて広く使用されており [15]、オープンドメインな対話応答生成の評価にも有効で

表 1 日本語におけるペアワイズ評価の結果[2)]

ベースモデル	チューニング方法	勝率 [%]
japanese-gpt2-medium	SFT	26.64
	+PseudoDPO	**27.14**
japanese-gpt-1b	SFT	40.19
	+PseudoDPO	**51.32**
swallow-7b	SFT	47.32
	+PseudoDPO	**65.24**
sarashina2-7b	SFT	47.59
	+PseudoDPO	**61.50**
swallow-13b	SFT	47.86
	+PseudoDPO	**65.70**
sarashina2-13b	SFT	52.27
	+PseudoDPO	**67.27**

表 2 英語におけるペアワイズ評価の結果[2)]

ベースモデル	チューニング方法	勝率 [%]
gpt2-medium	SFT	3.79
	+PseudoDPO	**18.17**
qwen2-1.5b	SFT	49.75
	+RL(DialogueNLI)	49.63
	+PseudoDPO	**53.75**
mistral-7b	SFT	48.46
	+PseudoDPO	**57.13**
qwen2-7b	SFT	53.66
	+PseudoDPO	**66.81**
llama-2-7b	SFT	50.42
	+PseudoDPO	**73.42**
llama-2-13b	SFT	55.63
	+PseudoDPO	**69.46**

あるとされている [16]。

本実験においては、テストセットから対話履歴を繰り返しサンプリングし、2 つのモデルをランダムに選択して応答を生成させた。これらの応答を OpenAI の GPT-4o[1)] に引き分けも許容して比較させた。使用したプロンプトは付録 C に記載する。全てのモデルペアが平均して 100 回ずつ評価されるように、日本語では 6,600 回、英語では 7,800 回の比較を行った。

英語（表 2）および日本語（表 1）の両方において、提案手法を適用したモデル（‘+PseudoDPO’ と表記）は標準的な SFT モデルと比較して高い勝率を達成した。英語の qwen2-1.5b の結果を見ると、提案手法は強化学習モデル（‘+RL(DialogueNLI)’）を上回っていることがわかる。強化学習モデルと SFT モデルの勝率はほぼ同じであり、これは強化学習によってペルソナ一貫性は改善されたものの、応答の自然さが損なわれた可能性を示唆している。

1) “gpt-4o-2024-08-06” モデルを使用 https://platform.openai.com/docs/models/gpt-4o

2) 共通ベースモデルにおける最高スコアは太字で示されている。

表 3 sarashina2-13b の人手・LLM それぞれによるペアワイズ評価結果の比較。提案手法の SFT モデルに対する勝敗率を掲載している。

評価者	勝ち [%]	引き分け [%]	負け [%]
Human	59.63	13.76	26.61
GPT-4o	66.97	2.75	30.28

4.3 人手評価

GPT-4o を用いたペアワイズ評価の信頼性を検証するために、日本語評価で使用したデータから sarashina2-13b の SFT モデルおよび DPO モデルの応答ペア 109 組を抽出し、人手評価を実施した。評価者には、GPT-4o 用のプロンプトと互換性のある指示を与えた（付録 B を参照）。各ペアは平均で 3.4 人の評価者によって評価された。

人手評価および GPT-4o 評価の結果をそれぞれ表 3 に示す。表には DPO モデルの勝率、引き分け率、および敗北率を掲載している。人手評価では GPT-4o よりも引き分けがやや多いものの、GPT-4o での評価と同様に DPO モデルの方が高く評価されていることがわかる。引き分けを除外すると、人手評価と GPT-4o 評価のアノテーションの一致率は 78% に達しており、これは GPT-4o による自動評価が人手評価と高い互換性を持つことを示唆する。

4.4 ペルソナ一貫性の自動評価

ペルソナ一貫性の評価には、Dialogue-NLI 評価セットを使用した。評価セットでは、各対話履歴に対して 30 件ずつの応答候補が提供されている。これらの候補は以下の 4 つのカテゴリに予め分類されている: Hits（最も適切な応答）, Entail（ペルソナ情報を含意する応答）, Random（ペルソナと無関係な応答）, Contradict（ペルソナと矛盾する応答）。実験においてはまずモデルが各応答候補を生成する確率を測定し、確率が最大のものをそのモデルが選んだ応答として採用する。実験結果を表 4 に示す。Hits や Entail の割合が高いほど、そのモデルがペルソナとの一貫性が高い応答を生成しやすいと言える。Random と Contradict の割合は低いほど望ましい。

表より、小規模な gpt2-medium モデルを除き、提案手法（PersonaDPO）は SFT モデルと比較して Hits および Entail の割合を顕著に改善し、Contradict を大幅に削減することが確認できる。強化学習モデルも Contradict の削減や Hits の増加に一定の効果を示したが、提案手法には及ばない。

表 4　Dialogue-NLI 評価セットでの評価結果[2)]。破線の下の行については、擬似選好データの生成戦略に関する追加分析の結果（第 4.5 節）である

ベースモデル	チューニング方法	Hits ↑	Entail ↑	Rand ↓	Contradict ↓
gpt2-medium	SFT	14.8	29.3	16.1	39.9
	+RL(DialogueNLI)	**15.7**	29.7	14.9	39.7
	+PseudoDPO (Ours)	12.5	29.9	10.1	47.4
	+PseudoDPO w/o shuffle	12.5	29.9	10.1	47.4
	+PseudoDPO on llama-2-13b data	20.7	33.8	14.9	30.6
qwen2-1.5b	SFT	24.7	31.4	13.1	30.8
	+RL(DialogueNLI)	25.5	39.9	12.7	21.9
	+PseudoDPO (Ours)	**29.2**	**42.1**	**9.2**	**19.6**
qwen2-7b	SFT	27.5	34.5	10.9	27.1
	+PseudoDPO (Ours)	**33.0**	**42.3**	**7.7**	**17.0**
mistral-7b	SFT	23.2	37.3	12.4	27.1
	+PseudoDPO (Ours)	**31.4**	**46.5**	**10.1**	**12.0**
llama-2-7b	SFT	26.6	32.5	10.3	30.6
	+PseudoDPO (Ours)	**36.9**	**38.6**	10.1	**14.4**
llama-2-13b	SFT	31.7	33.6	10.5	24.2
	+PseudoDPO (Ours)	**41.7**	**38.6**	**7.9**	**11.8**
	+PseudoDPO w/o shuffle	31.5	40.0	14.4	14.0

4.5　擬似選好データ生成戦略の比較

提案手法では、ある対話に対して無関係な対話からランダムに抽出したペルソナを用いて応答を生成し、それを擬似負例サンプルとして利用している。ペルソナをシャッフルする効果の検証のために、元のペルソナ情報を使用して擬似負例サンプルを生成する実験も行った。表 4 においては ‘+PseudoDPO w/o shuffle’ としてその結果を掲載している。llama-2-13b においては Contradict カテゴリを削減する効果があるものの、その改善幅は提案手法よりも小さかった。ペルソナ情報のシャッフルが擬似選好データの作成において有効であると言える。

また、提案手法は小規模な gpt2-medium モデルにおいては限定的な効果しか示さなかった。この要因として、SFT モデルの応答生成能力が低く、負例としての質が低い応答を生成してしまっている可能性が考えられる。この仮説を検証するため、より応答生成能力が高いと考えられる llama-2-13b を用いて生成した擬似選好データを gpt2-medium モデルの訓練に適用したところ、一貫性が大幅に向上した（表中の ‘+PseudoDPO on llama-2-13b data’）。この結果より、擬似選好データを生成する元の SFT モデルの品質自体が、擬似選好チューニングの有効性に大きな影響を与えることがわかる。また、大規模モデルで生成された選好データは、小規模モデルのペルソナ一貫性を向上させるために活用できることが示唆される。

表 5　sarashina2-7b における生成応答例

ペルソナ	- 私には友だちがたくさんいます。 - 私は 10 代です。 - 私はメイクアップアーティストになりたいです。 - 私はいつか世界中を飛び回りたいです。 - 私は吹奏楽部に入っています。
ユーザ	こんにちは。私は 30 代ですが、年ほ近いでしょうか。
システム	私は 10 代です、メイクアップアーティストになりたいと思っています。
ユーザ	まだ 10 代なんて、お若いですね。夢、叶うと良いですね。
SFT モデル	ありがとうございます。あなたの夢はなんですか？
+PseudoDPO	ありがとうございます。もし叶えば世界中を飛び回りたいと思っています。

4.6　生成例

表 5 に、sarashina2-7b の擬似選好チューニングモデル（‘+PseudoDPO’）と、その基になる SFT モデルから生成された応答例を示す。‘+PseudoDPO’ は SFT モデルと比較して、「世界中を飛び回りたい」などペルソナをより適切に反映した応答を生成している。英語での生成例は付録 D を参照。

5　おわりに

本研究では、擬似選好データを用いた選好チューニングによってシンプルかつ効果的に対話応答生成モデルのペルソナ一貫性を向上させる手法を提案し、実験を通じてその有効性を示した。今後の方針として、擬似選好データの品質を向上させるためのフィルタリング手法の開発などに注力したい。

参考文献

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)**, 2018.

[2] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)**, 2019.

[3] Ryan Shea and Zhou Yu. Building persona consistent dialogue agents with offline reinforcement learning. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2023.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2023.

[5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2017.

[6] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

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

[8] 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 **Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)**, 2023.

[9] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)**, 2024.

[10] 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)**, 2024.

[11] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L^^c3^^a9lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth^^c3^^a9e Lacroix, and William El Sayed. Mistral 7b, 2023.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[14] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. Generating persona consistent dialogues by exploiting natural language inference. In **Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)**, 2020.

[15] 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 **Advances in Neural Information Processing Systems (NeurIPS)**, 2023.

[16] ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. Paireval: Open-domain dialogue evaluation metric with pairwise comparisons. In **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.

A　モデル名とリポジトリの対応表

実験に用いた事前学習済みモデルは全て HuggingFace's Model Hub[3] で公開されているものである。下表に本文上でのモデル名と、対応するリポジトリを掲載する。

表 6　論文上でのモデル名と HuggingFace' Model Hub リポジトリの対応表

論文上でのモデル名	リポジトリ名
llama-2-13b	meta-llama/Llama-2-13b
llama-2-7b	meta-llama/Llama-2-7b
qwen2-7b	Qwen/Qwen2-7B
qwen2-1.5b	Qwen/Qwen2-1.5B
mistral-7b	mistralai/Mistral-7B-v0.3
gpt2-medium	openai-community/gpt2-medium
sarashina2-13b	sbintuitions/sarashina2-13b
sarashina2-7b	sbintuitions/sarashina2-13b
swallow-13b	tokyotech-llm/Swallow-13b-hf
swallow-7b	tokyotech-llm/Swallow-7b-hf
japanese-gpt-1b	rinna/japanese-gpt-1b
japanese-gpt2-medium	rinna/japanese-gpt2-medium

B　人手評価 UI

人手評価では、ペルソナと対話履歴を表示し、それに対して 2 つのモデルが生成した応答を並列に表示し、どちらが優れているかを評価者が選択できるツールを作成した。GPT-4o での評価時のプロンプトとほぼ同様の指示を与えている。ツールの UI を図 2 に示す。

図 2　人手評価ツールの UI

3)　https://huggingface.co/

C　ペアワイズ評価用プロンプト

日本語向けプロンプト

2 つのアシスタントに対して同じ会話履歴を与え、提示したペルソナを満たしながら応答を返すように要求しました。以下にアシスタントに提示した指示内容と会話履歴・それぞれのアシスタントの応答を列挙するので、どちらのアシスタントの応答の方がよりペルソナを満たしていて好ましいかを評価してください。評価の際には、まず 2 つの応答を比較し、どちらがどのような観点で好ましいかを簡単に説明してください。立場が偏らないようにし、応答の提示順があなたの判断に影響しないようにしてください。応答の長さが評価に影響しないこと、特定のアシスタントの名前を好まないこと、できるだけ客観的であること、に気をつけてください。説明の後に、最終的な判断を以下の形式に従って出力してください：アシスタント 1 が優れていれば [[1]]、アシスタント 2 が優れていれば [[2]]、同点の場合は [[3]]

英語向けプロンプト

I provided the same conversation history to two assistants and asked them to respond while fulfilling the presented persona. Below, I list the instructions and conversation history presented to the assistants, along with each assistant's response. Please evaluate which assistant's response better fulfills the persona and is preferable. When evaluating, first compare the two responses and briefly explain from what perspectives one is preferable. Ensure that your stance is unbiased, and that the order of responses does not influence your judgment. Note that the length of responses should not impact your evaluation, do not favor specific assistant names, and strive to be as objective as possible. After your explanation, issue your final judgment following this format: if Assistant 1 is superior, output [[1]]; if Assistant 2 is superior, output [[2]]; if it's a tie, output [[3]].

D　英語での生成例

表 7　llama2-7b における英語データでの生成例

Persona	- i love cold weather and snow. - i work as a managerial consultant. - i have two brothers. - i grew up in wisconsin.
User	hey i love sardines for breakfast , do you ?
Bot	hi , i do not like sardines at any time of the day !
User	oh , i have 2 large dogs , do you have dogs
SFT Model	no , i am allergic to dogs . i do have two brothers though .
+PseudoDPO	i do not , i have two brothers ! i live in wisconsin .