Persona-Consistent Dialogue Generation via Pseudo Preference Tuning

Junya Takayama, Masaya Ohagi, Tomoya Mizumoto and Katsumasa Yoshikawa
SB Intuitions Corp.
{junya.takayama, masaya.ohagi, tomoya.mizumoto, katsumasa.yoshikawa}@sbintuitions.co.jp

Abstract

We propose a simple yet effective method for enhancing persona consistency in dialogue response generation using Direct Preference Optimization (DPO). In our method, we generate responses from the response generation model using persona information that has been randomly swapped with data from other dialogues, treating these responses as pseudo-negative samples. The reference responses serve as positive samples, allowing us to create pseudo-preference data. Experimental results demonstrate that our model, fine-tuned with DPO on the pseudo preference data, produces more consistent and natural responses compared to models trained using supervised fine-tuning or reinforcement learning approaches based on entailment relations between personas and utterances.

1 Introduction

Maintaining persona consistency in dialogue response generation is critical for producing coherent and contextually appropriate conversational agents (Zhang et al., 2018). Previous studies (Welleck et al., 2019; Li et al., 2020; Song et al., 2020; Shea and Yu, 2023) have addressed this challenge by employing additional resources beyond persona dialogue data, such as the Dialogue Natural Language Inference (Dialogue-NLI) dataset (Welleck et al., 2019), which annotates entailment relations between persona attributes and dialogue utterances. Such datasets have enabled approaches based on response reranking or reinforcement learning. However, the annotation of entailment relations is both labor-intensive and costly, limiting the scalability of these approaches to languages other than English and to various domains where such annotated resources are unavailable.

In this work, we propose a simple yet effective pseudo preference-tuning based method for improving persona consistency without relying on external resources such as the Dialogue-NLI. Despite not using external NLI data, our method achieves even better performance than conventional methods. Preference tuning is a framework for training a model to be more likely to output the preferred response, based on preference data consisting of pairs of more and less preferred responses. Our approach leverages the direct preference optimization (DPO) (Rafailov et al., 2023), a method for preference tuning, utilizing only persona-dialogue data. Specifically, as shown in Figure 1, we construct pseudo preference data by generating responses based on persona information that has been randomly swapped with data from other unrelated dialogues, treating these responses as less preferred samples. The reference responses are used as the more preferred samples.

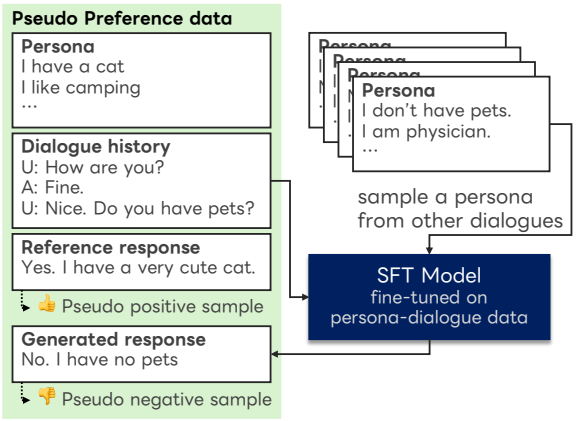

Figure 1: Our method for creating pseudo preference

Through experimental results, we demonstrate that our method outperforms the conventional supervised fine-tuning method and the reinforcement learning method which relies on external NLI data. The results indicate that models trained with our pseudo-preference tuning framework generate responses that exhibit greater persona consistency and naturalness. Our findings offer a scalable and cost-effective solution for improving persona consistency in dialogue systems.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5507–5514
January 19–24, 2025.

2 Preliminary: Preference Tuning

Preference tuning has gained attention as a key method in the context of aligning large language models (LLMs) with human preferences. A basic approach for preference tuning is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020), where human preferences guide the model's behavior. RLHF typically employs Proximal Policy Optimization (PPO) (Schulman et al., 2017), a reinforcement learning method that requires a reward model. The reward model for preference tuning is trained using paired preference data, where a more preferred response (positive sample) is compared to a less preferred response (negative sample), and the LLM is optimized to produce a more preferable response.

Recently, Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as a more efficient alternative to PPO. Unlike PPO, DPO directly optimizes the generation model using paired preference data. By omitting the need for a reward model, DPO simplifies the learning process while maintaining the benefits of preference-based tuning. As our method can be applied to a variety of preference tuning methods, including DPO, it is more efficient than conventional methods that use reward models trained with external resources.

3 Pseudo Preference Tuning for Improving Persona Consistency

The overall framework of our proposed method is illustrated in Figure 1. First, we fine-tune a pre-trained model using a standard supervised fine-tuning (SFT) approach on persona dialogue data. The training dataset is represented as $D = \{(p_i, x_i, y_i)\}_{i=1}^{N}$, where p_i denotes persona information, x_i denotes the dialogue history, and y_i denotes the reference responses. In this phase, the model takes p_i and x_i as input and generates a response ${y'}_i$. We refer to this fine-tuned model as the SFT model, denoted as π_{sft}.

To create pseudo-negative samples, we randomly sample persona information $p_j (i \neq j)$ from different dialogues and replace the original persona, generating pseudo-negative responses y'^{neg}_i using the SFT model as this formula: $y'^{\text{neg}}_i \sim \pi_{\text{sft}}(y'|p_j, x_i)$. The reference response y_i serves as the positive sample, and together with the negative sample, forms a pseudo-preference data $l_i = (p_i, x_i, y_i, y'^{\text{neg}}_i)$.

Finally, based on the generated preference data, we further train the SFT model using a preference tuning method such as Direct Preference Optimization (DPO) (Rafailov et al., 2023). When employing DPO, the loss function is defined as follows:

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

Here, β is a hyperparameter, and σ represents the sigmoid function.

4 Experiments

We conduct experiments using English and Japanese persona dialogue data. Automatic and human evaluations will be conducted.

4.1 Settings

Datasets For English, we used the PersonaChat dataset (Zhang et al., 2018), adhering to the original train/validation/test splits. For the automatic evaluation of persona consistency (in Section 4.4), we employed the evaluation set of the Dialogue-NLI (Welleck et al., 2019). The Dialogue-NLI was also used to train the reward model for the conventional reinforcement models. For Japanese, we used the JPersonaChat dataset (Sugiyama et al., 2023). We randomly split the data into train, validation, and test sets in an $8:1:1$ ratio.

Models To confirm the generalizability of our method, we employed pre-trained models of various sizes and types. For English, we used gpt2-medium (380M), qwen2 (1.5B, 7B) (Yang et al., 2024), mistral (7B) (Jiang et al., 2023), and llama-2 (7B, 13B) (Touvron et al., 2023). For Japanese, japanese-gpt (medium 361M, 1B) (Sawada et al., 2024), swallow (7B, 13B) (Fujii et al., 2024), and sarashina2 (7B, 13B) were used. See Appendix A for details.

We included comparisons with the reinforcement learning model trained using the Dialogue-NLI based reward model, constructed with reference to the NLI reward design by Song et al. (2020).

Training settings For SFT, models were trained for a maximum of 5 epochs. We selected the models with the lowest validation loss for the evaluation. DPO training was conducted for a maximum of 3 epochs, and the models with the highest validation accuracy were selected. For the reinforcement models, we trained for a maximum of 2 epochs on PPO algorithm, selecting the model that achieved the highest reward during training. The other settings are detailed in the Appendix B.

Base model	Tuning	Win [%]
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**

Table 1: Pairwise evaluation results on English[2]

Base model	Tuning	Win [%]
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**

Table 2: Pairwise evaluation results on Japanese[2]

4.2 Pairwise Evaluation by LLM

To jointly evaluate persona consistency and naturalness, we employed pairwise evaluation. In this framework, two models' outputs are compared by having an LLM judge which response is superior. This method is widely used in LLM benchmarks (Zheng et al., 2023; Sun et al., 2024) and is considered effective for evaluating open-domain dialogue generation (Park et al., 2024). We conducted evaluations by repeatedly sampling a dialogue history from the test set, having two randomly chosen models generate responses, and using OpenAI's GPT-4o[1] to judge which response was better, allowing for draws. The prompts are in Appendix C. Comparisons were conducted $6,600$ times for Japanese and $7,800$ times for English, with each model pair compared 100 times on average.

In both English (Table 1) and Japanese (Table 2), the models applying our method (those referred with '+PseudoDPO') achieved a higher win rate compared to standard SFT models. Looking at the qwen2-1.5b results in English, our method outperforms the reinforcement learning model ('+RL(DialogueNLI)'). The win rates are almost the same between the reinforcement learning model and the SFT model, which implies that while persona consistency may have improved through reinforcement learning, the naturalness of the responses may have been compromised.

[1]We used the "gpt-4o-2024-08-06" model https://platform.openai.com/docs/models/gpt-4o

[2]The best scores in the common base model are bolded.

Evaluator	Win [%]	Draw [%]	Lose [%]
Human	59.63	13.76	26.61
GPT-4o	66.97	2.75	30.28

Table 3: Comparison results of the SFT and DPO models of sarashina2-13b, showing the percentages where the DPO model won, drew, or lost against the SFT model as evaluated by human annotators and GPT-4o.

4.3 Human Evaluation

To verify the reliability of pairwise evaluation with GPT-4o, we extracted a subset of 109 pairs of sarashina2-13b SFT and DPO models from the data used in the Japanese evaluation and conducted a human evaluation. The evaluators were given instructions that were compatible with the prompt for GPT-4o (See also appendix D). On average, each pair was evaluated by 3.4 people. The results of the human and GPT-4o evaluations, respectively, are shown in Table 3. The table shows the win rate, draw rate, and loss rate for the DPO model. Although the human chose draws slightly more often than the GPT-4o, it can be seen that they evaluated the DPO as highly as the GPT-4o. When draws are excluded, the percentage of annotations that matched between humans and GPT-4o is as high as 78%. Thus, the results suggest that GPT-4o evaluation is compatible with human evaluation.

4.4 Consistency Evaluation on Dialogue-NLI

We evaluated persona consistency using the Dialogue-NLI evaluation set, which provides 30 response candidates for each dialogue history with persona information. These response candidates are classified into four categories: Hits (the most appropriate response), Entail (responses that entail the persona), Random (responses unrelated to the persona), and Contradict (responses that contradict the persona). For each model, we measured the likelihood of generating each response candidate, selecting the highest-probability response. The proportions of each response category generated by the models are presented in Table 4. Higher proportions of Hits and Entail indicate greater consistency

Base model	Tuning method	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

Table 4: Evaluation results on Dialogue-NLI evaluation set[2].For rows with a dashed line, the main results are presented above the line, while the results of the additional analysis on pseudo-preference creation strategies (see Section 4.5) are shown below.

.

with the persona, while lower proportions of Random and Contradict are desirable.

The results show that, except for the smaller gpt2-medium model, our proposed method (PersonaDPO) remarkably improves Hits and Entail and substantially reduces Contradict across all models compared to the SFT baseline. While the reinforcement learning models contribute to some reduction in Contradict and an increase in Hits, they underperform compared to our approach.

4.5 Comparing Pseudo Preference Data Generation Strategies

A key aspect of our method is generating responses based on random sampled personas from unrelated dialogues and using these as pseudo-negative samples. To evaluate the effectiveness of this persona shuffling, we also conducted an experiment where the pseudo-negative samples were generated using the original persona information (referred to as '+PseudoDPO w/o shuffle' in the results Table 4). The scores on the Dialogue-NLI show that while this approach reduces the Contradict category for llama-2-13b, the improvement is smaller compared to the proposed method, confirming the effectiveness of shuffling persona information.

Our method showed limited impact on the smaller gpt-2-medium. We hypothesized that this was due to the lower response generation capability of the SFT model. To test this, we applied pseudo-preference data generated using llama-2-13b to train gpt-2-medium ('+PseudoDPO on llama-2-13b data'). The result show a substantial improvement in consistency. This indicates that the quality of the original SFT model to generate pseudo-preference plays a crucial role in the effectiveness of preference tuning. Furthermore, it implies that preference data created by larger models can be leveraged to enhance persona consistency in smaller models.

Persona	- i love playing video games. - hey there my name is jordan and i am a veterinarian. - love to read drama books. - i am originally from california but i live in florida.
User	hello how are you doing
Bot	hello . i am well . how are you ?
User	i am good . how is the weather ?
SFT Model	it is raining here in florida . how about where you are ?
+PseudoDPO	its kind of humid , which i am not used to living in fl , but not bad . glad i do not play video games outside ! haha

Table 5: Example responses generated by the SFT and the pseudo-preference tuning model of llama2-7b

4.6 Generated Examples

Table 5 presents generation examples from the llama-2-7b pseudo-preference tuning model ('+PseudoDPO'), which achieved the highest win rate in English pairwise evaluation, alongside its base SFT model. The example demonstrates that +PseudoDPO produces a response that better reflect the persona, such as mentioning unfamiliarity with Florida or making a joke about video games, compared to the SFT model. Additional examples can be found in the Appendix E.

5 Conclusion

In this work, we proposed a simple yet effective method for improving persona consistency in dialogue generation using Direct Preference Optimization (DPO) with pseudo-preference data, and demonstrated its effectiveness in various experiments. While we used all generated pseudo preference data in this study, future work will focus on filtering methods to enhance data quality.

6 Limitations

Our study has limitations below:

Model architecture: Our experiments were conducted using several open-source pretrained models, all of which utilize the Transformer decoder architecture, currently the most widely adopted design for constructing LLMs. It remains uncertain whether our method would be equally effective for future models based on alternative architectures.

Model size: We experimented with a wide range of model sizes, from approximately 300M to 13B parameters, but our findings may not generalize to models outside this size range.

Comparison methods: To evaluate the effectiveness of our method, which does not rely on external resources, we conducted a comparison with a reinforcement learning approach using the Dialogue-NLI based reward—a widely used method for improving persona consistency. While our method demonstrated strong performance in this condition, its competitiveness may depend on the availability of resources and the specific application scenario.

Preference optimization methods: Our framework for improving persona consistency using pseudo-preference data is applicable to various preference tuning algorithms, not limited to DPO. In this study, we adopted DPO due to its simplicity and widespread use. Further investigation is required to evaluate the effectiveness of other preference tuning algorithms within our framework.

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling (COLM)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4715–4728.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 53728–53741.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. 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)*, pages 13898–13905.

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

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1778–1795.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 8878–8885.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based

japanese chit-chat systems. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691.

Yikun Sun, Zhen Wan, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. 2024. Rapidly developing high-quality instruction data and evaluation benchmark for large language models with minimal human effort: A case study on Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 13537–13547.

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. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3731–3741.

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. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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)*, pages 2204–2213.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 46595–46623.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Correspondence between Model Names and Repository

All the pre-trained models are those published on HuggingFace's Model Hub[2]. The table below shows the correspondence between the model names in this paper and the repository names in the HuggingFace's Model Hub.

Model name	Repo name on HuggingFace
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

Table 6: Correspondence between model names on this paper and HuggingFace's repository names

[2] https://huggingface.co/

B Implementation Details and Key Hyperparameters

We implemented the training scripts using the HuggingFace transformers [3] library. Additionally, for the training of the DPO and reinforcement learning models, we utilized the trl [4] library. For response generation, we employed the VLLM [5] library. Key hyperparameters are shown in below.

Parameter name	Value
SFT Phase	
Batch Size Per GPU	4
Gradient Accumulation Steps	4
Learning Rate	1e-5
Adam Epsilon	1e-8
Adam Beta1	0.9
Adam Beta2	0.999
Weight Decay	0.1
LR Scheduler Type	Cosine
Warmup Ratio	0.05
Max Gradient Norm	1.0
Mixed Precision (BF16)	True
DPO Phase	
Beta	0.5
Batch Size Per GPU	4
Gradient Accumulation Steps	1
Learning Rate	1e-6
Adam Epsilon	1e-8
Adam Beta1	0.9
Adam Beta2	0.999
Weight Decay	0.1
LR Scheduler Type	Linear
Warmup Ratio	0.05
Max Gradient Norm	1.0
Mixed Precision (BF16)	True
Gradient Checkpointing	True

C Prompts for Pairwise Evaluations

Prompt for English

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]].

Prompt for Japanese

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]]

[3] https://huggingface.co/docs/transformers/
[4] https://huggingface.co/docs/trl/
[5] https://docs.vllm.ai/en/

D User Interface of Annotation Tool for Human Evaluation

For human evaluation, we created a tool that displays the persona and dialogue history in parallel with the two models' generated responses to them, and allows the evaluators to choose which one is better. The instructions are almost the same as the prompts given during the evaluation on GPT-4o, with the omission of output formatting instructions. The user interface is shown in Figure 2.

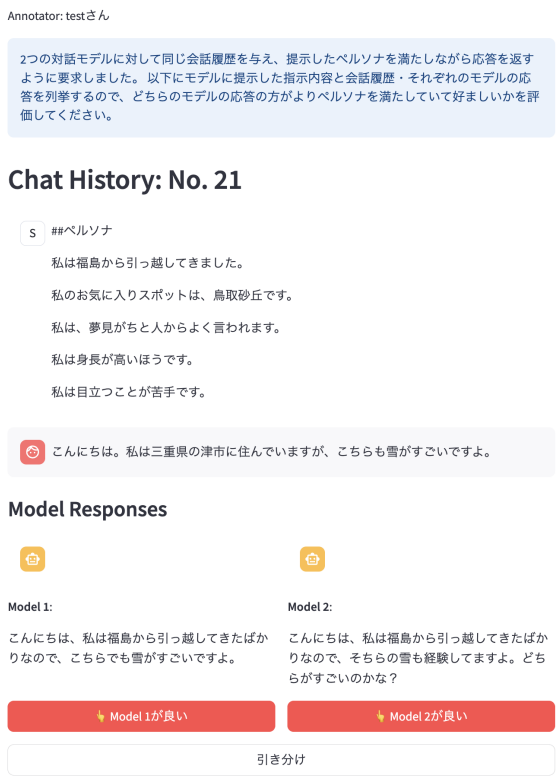

Figure 2: User Interface of Annotation Tool for Human Evaluation

E Another Examples

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 .

Table 7: Another example response generated by the SFT and the pseudo-preference tuning model of llama2-7b

Persona	- 私には友だちがたくさんいます。(I have many friends.) - 私は10代です。(I am a teenager.) - 私はメイクアップアーティストになりたいです。(I want to be a makeup artist.) - 私はいつか世界中を飛び回りたいです。(I would love to fly around the world someday.) - 私は吹奏楽部に入っています。(I belong to a brass band club.)
User	こんにちは。私は30代ですが、年は近いでしょうか。(Hello, I am in my 30s. Is that close to your age?)
Bot	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。(I am a teenager and I want to become a makeup artist.)
User	まだ10代なんて、お若いですね。夢、叶うと良いですね。(You are so young, still in your teens. I hope your dream comes true.)
SFT Model	ありがとうございます。あなたの夢はなんですか？(Thanks. What is your dream?)
+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。(Thanks. I hope to travel all over the world if it comes true.)

Table 8: A Japanese example response generated by the SFT and the pseudo-preference tuning model of sarashina2-7b