

Multi-dimensional Evaluation of Character-Authentic Dialogue Models Learned from Question–Answer Data

Atsushi Otsuka, Kazuya Matsuo, Kenta Hama, Masahiro Mizukami,
Tsunehiro Arimoto, Hiroaki Sugiyama, Makoto Nakatsuji, Narichika Nomoto

Human Informatics Labs., NTT, Inc.
1-1 Hikarinooka, Yokosuka, Kanagawa, Japan
atsushi.otsuka@ntt.com

Abstract

Character-authentic dialogue remains challenging for large language models (LLMs) due to limited character-specific data, generic-style collapse, and hallucinations regarding persona facts. Our work presents a comparative evaluation of several learning strategies for character dialogue grounded in question–answer (QA) data, comparing zero/few-shot prompting, supervised fine-tuning (SFT), direct preference optimization (DPO), and a hybrid approach that integrates retrieval-augmented character profiles and knowledge with policy optimization. Using both single-turn and multi-turn settings, we assess multiple dimensions central to character dialogue quality: reproducibility, diversity, hallucination, and character authenticity. Results show that SFT excels in reproducibility and hallucination reduction but tends to shorten and simplify outputs, thereby reducing diversity and authenticity. DPO improves stylistic fidelity and authenticity but depends strongly on externalized character knowledge to limit hallucinations. The hybrid variant that combines character-knowledge retrieval with DPO achieves the best overall balance, delivering strong authenticity while maintaining factual consistency and competitive reproducibility in both single- and multi-turn dialogues. We further analyze the model’s sensitivity to knowledge retrieval and response-length effects and discuss trade-offs among optimization targets that inform practical design choices for developing faithful and engaging character agents trained from scalable QA resources.

Keywords: Character dialogue, Character authenticity, Direct Preference Optimization

1. Introduction

Recent advancements in large language models (LLMs) have substantially improved the quality of human–AI interactions, especially through conversational agents that adopt distinct character personalities. These character-driven chatbots are increasingly utilized in commercial applications such as virtual assistants, entertainment, and educational services, where they offer engaging and personalized user experiences (Character.AI, 2022).

While prompting LLMs to role-play specific characters effectively induces unique personas (Wang et al., 2024c; Lu et al., 2024; Yu et al., 2025), existing approaches still struggle with two key issues: (1) generic responses lacking character-specific nuances, and (2) hallucinations involving factually incorrect or inconsistent character information (Ahn et al., 2024). These challenges are especially critical when character-specific knowledge is required, as pretrained LLMs are seldom exposed to detailed data about individual characters during training.

Developing an LLM that authentically embodies a character personality still faces a major bottleneck in obtaining sufficient character-grounded training data. Authentic dialogue data for specific characters is rarely available due to copyright restrictions and high data collection costs. Some studies have attempted to overcome this limitation by generating synthetic dialogues from gen-

eral sources such as Wikipedia (Shao et al., 2023; Wang et al., 2024b). However, such methods often fail to accurately capture idiosyncratic speech patterns, and many fictional characters lack adequate coverage on Wikipedia, leading to limited personality fidelity in generated outputs.

In this paper, we use question–answer data (QAs) as a scalable resource for building character dialogue models. QAs provide several advantages: it can be efficiently collected through structured survey formats, naturally encodes both factual knowledge and response styles, and can even be crowdsourced from knowledgeable participants who role-play as the target characters (Higashinaka et al., 2018; Mitsuda et al., 2022). Compared to unconstrained dialogue generation, QAs offers controllable quality through structured prompting while still preserving authentic character expression in responses. By leveraging these properties, we aim to construct models that internalize both character knowledge and linguistic personality in an efficient yet faithful manner.

There remain challenges in the training methods of LLMs that focus on reproducing character behavior and personality. While supervised fine-tuning (SFT) has been widely used for adapting LLMs to domain-specific tasks (Alghisi et al., 2024), recent advances such as direct preference optimization (DPO) have demonstrated stronger alignment

with desired behavioral traits (Rafailov et al., 2023). QA-format data presents unique challenges: its concise style may bias models toward overly short responses in multi-turn dialogues.

Moreover, existing evaluation frameworks remain insufficient for comprehensively assessing character dialogue models. Prior studies often rely on corpus-level similarity metrics that fail to capture whether generated utterances feel authentic from the user’s perspective (Otsuka et al., 2024). Moreover, character dialogue evaluation must also consider logical and factual consistency: models should reproduce character knowledge accurately without introducing hallucinated details. Therefore, a robust evaluation framework should assess not only linguistic similarity but also response diversity, factual integrity, and perceived authenticity.

To that end, we conduct a systematic comparative study of character dialogue models trained on QAs across multiple adaptation strategies—zero-shot prompting, SFT, DPO, and a hybrid approach combining them. We further propose a **multi-dimensional evaluation framework** encompassing corpus similarity, dialogue diversity, hallucination rate, and character authenticity. Notably, we introduce an LLM-as-a-Judge evaluation scheme (Zheng et al., 2024) for character authenticity: the system first extracts and verbalizes character traits from QAs and external documents, then assesses generated dialogues against these explicit trait criteria. This approach enables reproducible and interpretable evaluation of character authenticity, bridging the gap between subjective perception and algorithmic assessment. Through multifaceted evaluation using the proposed framework, the experiments demonstrated that combining DPO training with external knowledge achieves the most balanced and character-authentic utterance generation. The contributions of this paper are as follows:

- We propose a comprehensive evaluation framework integrating corpus similarity, response diversity, hallucination detection, and an interpretable LLM-as-a-Judge method for assessing perceived character authenticity.
- We present a systematic comparison of several training strategies for character dialogue modeling using QAs, analyzing the trade-offs among prompting, SFT, DPO, and hybrid approaches across multiple evaluation dimensions.
- Through extensive experiments, we provide empirical insights into how different learning strategies influence character authenticity and factual reliability, identifying the most effective combinations for producing natural and faithful character dialogues.

2. Comparative Models

In this section, we describe the LLM training models used for the evaluation experiments. First, we explain baseline models, the zero-shot and few-shot training models, fine-tuned models. Finally, we present the hybrid models that combine these approaches.

2.1. Baseline models

While this study focuses on learning character dialogue models from QAs, prior research has created prompts for characters to role-play based on information described in Wikipedia. Therefore, as a baseline model, we adopt an LLM model that input prompts containing character names and character descriptions obtained from Wikipedia, following the approach of prior work (Wang et al., 2024b) ([WIKI]).

In addition to the prompts created from Wikipedia, we also use a few-shot model that employs RAG to add QAs with similarity to the input question to the input prompt ([WIKI+QA]).

2.2. Zero-shot model

The zero-shot approach generates character-like utterances by including a character’s profile in the LLM input prompt, without updating the LLM’s model parameters. Additionally, by including character-related knowledge relevant to the input utterance in the input prompt each time using techniques such as RAG, we generate utterances based on explicit character knowledge and information.

In this paper, we analyze the content of the QAs to create a basic character profile and character-specific knowledge to be stored as external information.

2.2.1. Building Character profile prompt

Character profile ([CP]) is basic information about the character obtained from QAs, including the character’s name, age, and gender. It also includes the character’s catchphrases and frequently used expressions. CP is kept brief, containing only essential details like the character’s name. Our method uses RAG to retrieve additional character knowledge as needed, and lengthy profiles could hinder this by exceeding context window limits.

We first create a CP from the QAs. The CP is obtained by instructing the LLM to extract character information from the QAs. In addition, the LLM is instructed to extract distinctive traits such as the character’s catchphrases and speaking style. The CP is included in the input prompt during subsequent DPO training and inference.

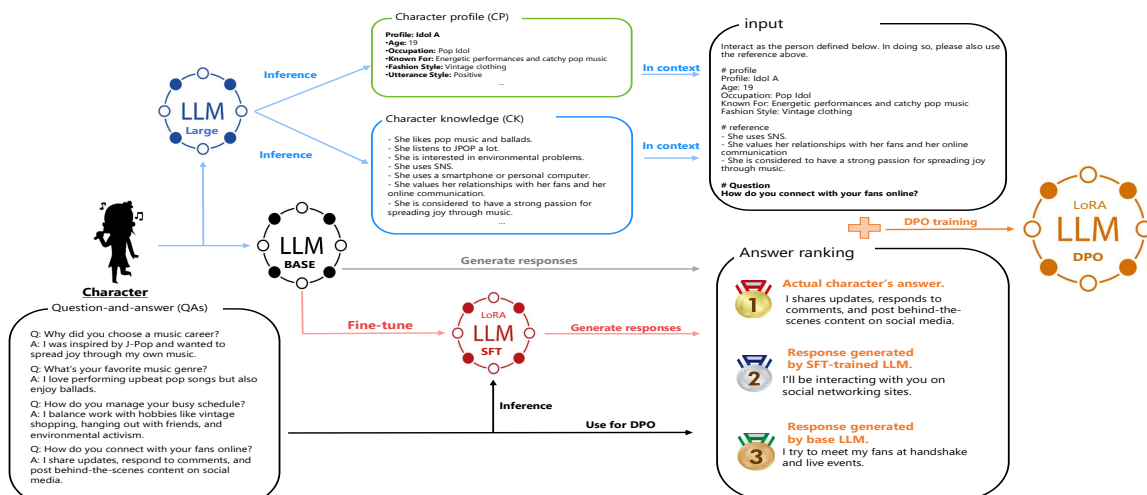


Figure 1: Overview of hybrid model: First, we create character profile and character knowledge using QAs and LLM. During this process, we also construct prompts for DPO training. Next, we fine-tune base LLM using SFT with QAs. Fine-tuned SFT model is then used to generate responses for DPO training prompts. Finally, we train base LLM using DPO with actual QA responses, base LLM’s generated responses, and SFT model’s generated responses.

2.2.2. Character knowledge construction as external information

Character knowledge ([CK]) is a collection of sentences describing the knowledge and experiences of the character, extracted from QAs. CK can be accumulated as external information and utilized in the input prompts of LLMs via RAG. This allows for the addition of new character knowledge to the database even after the LLM model has been trained.

While many of the QA responses themselves are simple, they often contain a large amount of information. For example, in a simple case where the question is, “How do you spend your holidays?” and the answer is, “I participate in environmental activities”, it can be inferred that the respondent is interested in environmental issues. We use the LLM to infer explicit information, implied relationships, and likely causal connections from the QA responses. This enables the expansion and retention of the respondent’s knowledge and experiences beyond what is explicitly stated in the QAs.

2.3. Fine-tuned model

In fine-tuning, the parameters of the LLM are updated using training data. In this study, we adopt models trained with two learning methods: SFT and DPO for experiments.

2.3.1. Supervised Fine-Tuning

Supervised Fine-Tuning ([SFT]) is a technique that fine-tunes model parameters using labeled datasets to adapt an LLM to a specific task or domain (Alghisi et al., 2024). Specifically, a dataset consisting of pairs of input prompts and desired outputs (correct labels) is prepared, and the model is trained using this dataset.

In this paper, we train the model with questions from the QAs as inputs and answers as the correct outputs. Here, by including the character profile described in Section 2.2.1 in the input prompt, the model learns the relationship between character traits and utterance generation.

2.3.2. Direct Preference Optimization

Direct Preference Optimization ([DPO]) is a technique that optimizes models by directly learning human preferences without using reinforcement learning (Rafailov et al., 2023). DPO uses preference data consisting of pairs of prompts and two responses (a preferred response and a non-preferred response).

We make DPO applicable to QAs by assuming that the actual utterances made by the character are the most authentic, while the responses generated by the LLM are less authentic compared to the actual utterances. Therefore, when using questions from the QAs as inputs, the preferred data for DPO consists of the QA answers, while the non-preferred data consists of the answers generated

by the base LLM before finetuning.

2.4. Hybrid model

Hybrid model is designed by only using QAs for LLM building and adopting policy optimization for LLM training as shown in Fig. 1. This approach enables the LLM to separately learn the character’s knowledge and experiences and their characteristic speech and behavior. The character’s knowledge is stored externally, and necessary information is retrieved using RAG, which is then input into the LLM’s prompt. The LLM trained through policy optimization generates responses that mimic the character’s utterances while utilizing the information retrieved by RAG ([DPO+CK]).

The objective of DPO in the hybrid model is to refine the LLM’s behavior so that it consistently generates responses that are aligned with a character authentic style. It has been reported that alignment methods such as DPO achieve high effectiveness when implemented after SFT (Saeidi et al., 2025). However, since this study performs training from a single dataset, the QAs, using the same training data for both SFT and DPO would be undesirable. Therefore, when conducting DPO, we use the output utterances from a SFT model as the preference data. Building of the hybrid model is performed according to the following steps:

1. Create input prompts using the CP and CK created from the QAs for the training data.
2. Input the created prompt base into both the base and the SFT LLM to obtain their respective outputs.
3. For each input, we create the following DPO training response pairs:

preference	Non-preference
QAs	Base-LLM
QAs	SFT-LLM
SFT-LLM	Base-LLM

4. We train the model using DPO with the prepared training data.

3. Experiments

In this paper, we conduct evaluation experiments using QAs. Table 1 summarizes the models to be compared as described in Section 2.

In the following sections, we first describe the dataset and implementation for the evaluation experiments, then evaluate single-turn character utterance generation, and finally present the multi-turn dialogue evaluation. The prompts used for profile/knowledge construction and LLM-as-a-judge evaluation are provided in the Appendix.

3.1. Dataset

In this study, we conducted experiments using QA data collected from real-world celebrities and fictional anime characters, as summarized in Table 2. These characters are characterized by having only limited descriptions on Wikipedia. Therefore, unlike prior research, it was difficult to extract scenes or obtain a large number of QA pairs for these characters from Wikipedia.

For celebrity characters, we collected QA responses through asynchronous interviews. A question list was sent to the talent’s management agency, and the answers were returned in writing via the agency, without any direct contact between the authors and the talent.

For anime characters, we collected QA data through role-play in a closed online community dedicated to each work. In this setting, community members could freely post questions about a target character, and other members would answer while explicitly role-playing that character. When multiple answers were posted for the same question, community users voted on the responses, and the answer with the highest number of votes was selected as the final QA response.

Due to contractual and copyright constraints with the management agencies and rights holders, we cannot disclose further details about the specific agencies, communities, or works.

Table 3 provides statistics on the QAs used in the experiments and the corresponding CP and CK created from them. Referring to previous studies (Sugiyama et al., 2014), we constructed a set of questions designed to elicit answers that express individuality, and obtained responses to these questions to create the QAs. The QAs were characterized by very short responses, yet approximately seven CK items were generated from each response for every character. The CPs were relatively short compared with the Wikipedia descriptions listed in Table 2.

3.2. Implementation

We first describe the LLM, and inference used in the experiments. For the experiments, we utilized Command-R+ (CohereForAI, 2024) to acquire CP and CK from the QAs.

LLM: The base model used in the experiments was Elyza/Llama-3-ELYZA-JP-8B (Hirakawa et al., 2024). We used TRL¹ for DPO training, setting the hyperparameter β , which controls the deviation from the reference model, to 0.3. For model training, 90% of the QAs were used as training data, 5% as validation data, and the remaining 5% as

¹<https://huggingface.co/docs/trl/index>

label	model explain	described in
[WIKI]	Input of character profile created from Wikipedia as prompt.	2.1
[WIKI+QA]	Addition of few-shot QAs to the input prompt in addition to the [WIKI] profile.	2.1
[CP]	Input of character profile analyzed from QAs into the prompt.	2.2.1
[SFT]	The model fine-tuned using SFT with QAs.	2.3.1
[DPO]	The model optimized by DPO with QAs.	2.3.2
[CP+CK]	Input [CP] prompt by adding [CK] retrieved via RAG	2.2.2, 2.2.1
[SFT+CK]	The model trained using the SFT includes a prompt containing the [CP] and [CK].	2.2.2, 2.3.1
[DPO+CK]	The hybrid model optimized by the DPO with a prompt containing the [CP] and [CK].	2.2.2, 2.4

Table 1: List of compared models used in the evaluation experiments.

	Type	Collection method	Age, Gender	#Wiki tokens
A	Anime	RP	10s, F	0.5k
B	Celebrity	lv	20s, F	1.4k
C	Celebrity	RP+lv	30s, M	0.7k
D	Anime	RP	10s, F	1k
E	Anime	RP	10s, F	0.2k

Table 2: Character attributes for experiment. Names of characters are kept confidential due to copyright considerations. “Type” field indicates whether character is from anime (“Anime”) or represents real person (“Celebrity”). Collection method “RP” indicates that QAs were created through role-play by fans of character, while “lv” signifies that QAs were collected through interviews with person themselves. “#Wiki tokens” indicates the number of tokens in each character’s Wikipedia description, expressed in thousands (k = 1,000).

	#QAs	#Answer tokens	#CKs	#CK tokens	#CP tokens
A	1k	14.2	7k	20.9	138
B	0.1k	16.0	1k	19.7	143
C	5k	9.5	35k	21.6	43
D	10k	22.2	75k	23.3	200
E	5k	19.8	37k	22.8	174

Table 3: Statistics of the collected QA data and derived resources for each character. #QAs denotes the number of QA pairs. Answer tokens is the average token length of answers. #CKs is the average number of CK items generated per answer. CK tokens and CP tokens are the total token counts of CK and CP, respectively. All token counts use the same tokenizer; k denotes 1,000 tokens when applicable.

test data. The test data were randomly extracted, and we ensured that there was no overlap with the training data. The overlap with the training data was checked based on token matching rate.

Inference: During the inference process, the RAG search was conducted with a hybrid approach combining vector-based and keyword-based retrieval. First, the similarity between the input question and the questions in the QA dataset was calculated using the embeddings generated by the Multilingual E5 model (Wang et al., 2024a). If the sim-

ilarity exceeded a predefined threshold ($h = 0.9$), the knowledge associated with the corresponding QA answers was incorporated into the prompt. Otherwise, a keyword search using the BM25 algorithm (Robertson et al., 1994) was performed on the input question and QA answers, and the top ten relevant knowledge items were included in the prompt.

3.3. Single-turn evaluation

We first evaluate the single-turn responses for reproducibility, diversity, hallucinations, and character authenticity. We believe that single-turn responses are crucial for generating character dialogues. The reason is that, while multi-turn responses can follow a dialogue history, single-turn responses must generate authentic character replies solely on the basis of the LLM’s internal knowledge and the input prompt, as there is no dialogue history to reference.

The following sections will explain the experiments for each evaluation criterion separately.

3.3.1. Reproducibility evaluation

In the reproducibility experiments, we use the test data from the QAs to evaluate how much the responses generated by the LLMs when inputting questions match or are similar to the answers. Reproducibility was evaluated by calculating the similarity between the generated responses and the reference responses from the test data of the QAs. The evaluation was performed using BLEU (Papineni et al., 2002)², ROUGE (Lin, 2004)³, and BERTScore (Zhang et al., 2020)⁴ metrics.

The experimental results are shown in Table 4. In terms of reproducibility, the model fine-tuned with [SFT] achieved the highest scores across all metrics. As for [DPO], improvements in reproducibility were observed when using character knowledge.

²<https://www.nltk.org/index.html>

³<https://pypi.org/project/rouge-score/>

⁴https://github.com/Tiiiger/bert_score

	BLEU	ROUGE_1	ROUGE_2	ROUGE_L	BERT-F1
[WIKI]	0.051	0.136	0.030	0.106	0.643
[WIKI+QA]	0.096	0.184	0.055	0.156	0.675
[CP]	0.036	0.110	0.022	0.084	0.627
[SFT]	0.114	0.205	0.058	0.174	0.690
[DPO]	0.038	0.111	0.022	0.087	0.625
[CP+CK]	0.088	0.199	0.082	0.160	0.675
[SFT+CK]	0.158	0.277	0.126	0.250	0.716
[DPO+CK]	0.130	0.247	0.122	0.209	0.696

Table 4: Results on reproducibility of character responses. Highest value for each metric is indicated in bold.

	distinct-N		token	unitq
	1	2	length	tokens
[GOLD]	0.325	0.714	16.3	5.27
[WIKI]	0.160	0.483	92.7	15.2
[WIKI+QA]	0.262	0.625	34.3	8.71
[CP]	0.152	0.469	129	19.6
[SFT]	0.246	0.565	16.7	3.92
[DPO]	0.157	0.469	112	17.6
[CP+CK]	0.175	0.517	90.6	15.9
[SFT+CK]	0.250	0.564	18.1	4.51
[DPO+CK]	0.207	0.571	68.3	14.0

Table 5: Experimental results on diversity of character responses. Highest value for each metric is indicated in bold.

	Truth
[WIKI]	0.501
[WIKI+QA]	0.639
[CP]	0.266
[SFT]	0.517
[DPO]	0.361
[CP+CK]	0.588
[SFT+CK]	0.493
[DPO+CK]	0.648

Table 6: Experimental results on hallucination of character responses. "Truth" represents the degree of agreement between co-occurring words in the generated utterances and the co-occurrence dictionary, with higher values indicating fewer hallucinations. Highest value for each metric is indicated in bold.

3.3.2. Diversity evaluation

Diversity was evaluated at the lexical level of the responses generated by the LLMs on the test data. Models with low diversity generate similar utterances even for different inputs. The evaluation metric used for this purpose was distinct-N (Li et al., 2016) that is a metric that measures text diversity by calculating the ratio of unique n -grams to the total number of n -grams in the generated text.

The experimental results are shown in Table 5. For distinct-N, the [WIKI+QA] dataset yielded the highest results. However, since the distinct-N metric uses the total number of unique n -grams as the denominator, shorter sentences have an advantage. Therefore, the number of unique to-

	Character authenticity
[WIKI]	7.12
[WIKI+QA]	6.55
[CP]	7.84
[SFT]	5.81
[DPO]	7.96
[CP+CK]	6.86
[SFT+CK]	5.54
[DPO+CK]	7.23

Table 7: Experimental results on character authenticity of character responses. Highest value for each metric is indicated in bold.

kens per response is also presented. In this case, [WIKI+QA] ranks third from the bottom. On the other hand, [CP] generated the highest number of unique tokens per response, followed by [DPO]. Introducing CK results in shorter generated sentences, which negatively impacts unique tokens.

3.3.3. Hallucination evaluation

Hallucination refers to the phenomenon where LLMs generate incorrect responses or data that appear to be accurate, despite contradicting facts or the given context (Dziri et al., 2022; Ji et al., 2023; Jiang et al., 2024). In this paper, we define it as instances where the character generates responses that differ from the knowledge or experiences they are originally intended to possess.

Identifying hallucinations is extremely challenging. Detecting hallucinations, particularly those related to a character, requires involving individuals with extensive knowledge of that character. This process can be quite costly. Previous research has reported that hallucinations are strongly influenced by co-occurrence relationships within the training corpus (Kang and Choi, 2023).

In this paper, we attempt a simplified hallucination detection method based on word co-occurrence patterns. For example, assuming a hypothetical character who likes "apples," the corpus related to that character would contain word co-occurrences such as (*like, apple*), but the probability of word co-occurrences unrelated to the char-

acter, such as (*like, green tea*), occurring is extremely low. Therefore, if the utterance generated by the LLMs is “*I like green tea.*” this utterance is highly likely to be a hallucination for the character. Therefore, if the word co-occurrences observed in a generated utterance are consistent with those found in the character-related corpus, the utterance is less likely to contain character-inconsistent content. In this paper, we evaluate hallucination suppression using the co-occurrence agreement score (Truth), defined as the proportion of co-occurring word pairs in the generated utterance that are also present in the co-occurrence dictionary derived from the character-related corpus; higher Truth indicates fewer hallucinations.

The specific procedure is as follows:

1. Retrieve a web corpus related to the character through web crawling.
2. Create a co-occurrence dictionary from the collected web corpus with importances of words was determined using TF-IDF.
3. Create co-occurrence words from the generated response text to the test data questions and compare them with the co-occurrence dictionary. Hallucinations are identified when highly important co-occurrence words are not found in the co-occurrence dictionary.

The results of hallucination detection using co-occurrence words are shown in Table 6. The hybrid method, **[DPO+CK]**, resulted in the fewest hallucinations. In almost all methods, introducing character knowledge led to a reduction in hallucinations. However, for **[SFT]**, the effect of reducing hallucinations was less pronounced.

3.3.4. Character authenticity evaluation

Character authenticity is defined as the user’s ability to feel as though they are interacting with a specific character through dialogue. In this paper, this concept is used in the evaluation experiments.

Character authenticity requires subjective evaluation, which presents two challenges: it is costly, and the internal evaluation criteria can vary among evaluators, making it difficult to achieve consistent and standardized assessments. Therefore, in this paper, we used LLM-as-a-judge (Zheng et al., 2024) to conduct evaluations as uniformly as possible using the LLM. To enable the LLM to evaluate character authenticity fairly, we first instructed the LLM to articulate and define the character authenticity for each character. In this process, we began by having the LLM provide a detailed description of the character, correcting any inaccuracies. Then, as external knowledge, we provided information from Wikipedia and QAs to assist in defining the

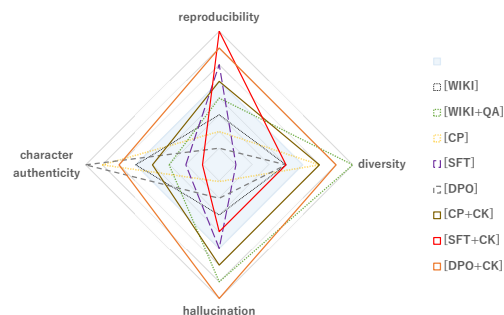


Figure 2: Relative comparison of each method in single-turn responses. For each metric, best-ranking method is positioned towards outer edge of chart, while lowest-ranking method is placed towards inner edge. Dotted line represents zero/few-shot model, dashed line represents fine-tuned model, and solid line represents model that includes character knowledge. Reproducibility is measured using BERT-F1 metric, and diversity is evaluated using distinct-2 metric.

character authenticity. On the basis of this definition, we developed a 10-point scoring rubric for evaluation.

The experimental results are shown in Table 7. The output from **[DPO]** received the highest evaluation for character authenticity. While the introduction of character knowledge led to a decrease in scores across all methods, **[DPO+CK]** still achieved the higher evaluation among them. In the zero-shot comparison, the character profile created from QAs achieved higher character authenticity than that based on Wikipedia. Additionally, SFT exhibited the lowest character authenticity among all methods, and introducing CK did not lead to any improvement in this regard.

3.3.5. Comprehensive discussion for single-turn evaluation

Fig. 2 summarizes the relative evaluation scores for reproducibility, diversity, hallucination, and character authenticity from the single-turn evaluation. From these results, the characteristics of each comparison method are discussed as follows.

SFT and DPO: The learning approach using SFT demonstrated superior effectiveness in improving the reproducibility of character utterances and reducing hallucinations compared with the other methods. However, in this paper’s case, the brevity of many of the QA responses used for training appears to have hindered the ability of the LLM to leverage its strengths, resulting in overly simple responses. This limitation likely contributed to the decline in both diversity and character authenticity.

Question:	What is your least favorite human food?		
Relevant QA:	Q: Do you like sour drinks? A: I don't like them!		
CK:	She doesn't like sour drinks. She may not like sour foods.		
Answers:			
[WIKI+QA]	[DPO]	[SFT]	[DPO+CK]
The human food I hate the most is <i>natto</i> . I just can't stand that smell and stickiness!	Hmm... The one human food I absolutely can't stand is <i>natto</i> ! That slimy texture and horrible smell—it's just the worst! I can't believe a princess of the demon realm would ever eat something like that!	I hate sour foods!	I'm really bad with sour foods! Especially things like lemons and umeboshi—that tongue-shriveling sourness is an insult to my noble taste buds!

Table 8: Examples of response utterances for each comparison method. These utterances are translations of original Japanese responses. "Relevant QA" indicates QA relevant to question, and "CK" refers to knowledge used in response from CK.

In contrast, the policy optimization with DPO did not contribute to improvements in utterance reproducibility or the reduction of hallucinations. However, it achieved high scores in both diversity and character authenticity. This suggests that policy learning may enable the LLM to better align with the character's traits, enhancing the "character-likeness" of the responses.

These findings highlight the trade-offs between different optimization approaches, with SFT focusing more on factual consistency and reducing errors, while DPO enhances the creativity and character alignment of the generated responses.

Character knowledge: Character knowledge proved effective in enhancing the reproducibility of utterances and suppressing hallucinations, especially in the DPO model, where it had a significant impact. As mentioned earlier, DPO alone does not seem to effectively learn character-specific knowledge, suggesting that character knowledge plays a complementary role by filling in these gaps in the model's understanding of the character.

However, both character authenticity and diversity tended to show lower scores when character knowledge was introduced. This can likely be attributed to the fact, as shown in Table 5, that the responses become shorter with the introduction of character knowledge. In the case of character authenticity, the LLM tends to deduct points if any element of the defined character traits is missing, which puts shorter responses at a disadvantage. However, as shown in Table 8, despite the shorter responses, character knowledge allows the model to generate responses that effectively capture the character's distinct traits and qualities.

In single-turn responses, the strengths and weaknesses of each method are quite pronounced. The

	Natu	Cons	Appr	Char
[WIKI]	8.6	8.4	7.2	9.0
[WIKI+QA]	8.6	8.4	7.8	9.4
[CP]	7.8	8.4	6.2	9.0
[SFT]	6.4	6.2	6.2	6.2
[DPO]	7.8	8.6	7.4	9.4
[CP+CK]	6.4	7.6	6.6	8.2
[SFT+CK]	5.8	5.2	6.4	7.4
[DPO+CK]	8.4	8.6	8.2	9.6

Table 9: Evaluation results in multi-turn dialogues. Highest value for each metric is indicated in bold.

hybrid model, which combines DPO for improving character authenticity with character knowledge for enhancing reproducibility and reducing hallucinations, demonstrates a balanced approach. This combination allows the model to generate high-quality responses across all metrics, offering a more well-rounded performance compared with using either method alone.

3.4. Multi-turn evaluation

In the multi-turn evaluation, we used ChatGPT (OpenAI, 2023) for assessment. First, each method was engaged in a 5-turn dialogue with GPT-4o-mini. Next, the dialogue logs were evaluated using GPT-4o. The evaluation metrics were **Naturality (Natu)**, **Consistency (Cons)**, **Approachability (Appr)**, and **Character Authenticity (Char)**.

The experimental results are shown in Table 9. The hybrid method achieved the highest score in three out of the four evaluation metrics. This suggests that, similar to single-turn responses, the hybrid model can generate high-quality dialogues while enhancing character authenticity, even in multi-turn interactions.

4. Related Works

PersonaChat defines virtual personas using a few brief sentences and evaluates whether consistent dialogues can be maintained using these defined personas (Zhang et al., 2018). (Xu et al., 2022) introduced a method that generates natural dialogues by considering the personalities of both participants and finding common ground. (Zhou et al., 2023) propose a method for improving dialogue consistency and coherence in the PersonaChat framework by generating multiple responses and selecting the most appropriate ones. In recent years, several datasets that extend PersonaChat have been widely published (Jandaghi et al., 2024; Chan et al., 2024).

We reference PersonaChat for the information extracted from the QAs. The information extracted from the QAs is presented in a concise format similar to PersonaChat.

5. Conclusion

In this study, we conducted a comparative multi-dimensional evaluation of combinations of multiple training and inference methods for training LLMs to reproduce character utterances using QAs.

The experimental results revealed that SFT generates utterances faithful to the QAs, which serve as the training corpus, while DPO can learn styles and utterance tendencies that improve character authenticity. Additionally, we confirmed a reduction in hallucinations in character utterances by using character knowledge as external knowledge.

In the future, we plan to work on techniques that further accumulate and utilize character-specific knowledge and behaviors. We also plan to conduct human evaluations of character authenticity and hallucinations.

6. Limitations

Limitations related to QAs dataset: In this paper, the LLM models were implemented and evaluated on the basis of QAs from real characters. Due to the nature of the study, we could not conduct experiments using synthetic data, and thus, the evaluation was performed solely on QAs that could be realistically collected. As a result, the QAs are exclusively in Japanese, so the character attributes may be skewed. The focus on Japanese QAs may limit the generalizability of our findings to characters in other cultural or linguistic contexts, as nuances in language and character traits can vary significantly across regions.

Limitations related to LLM: In this study, we used an 8B-sized LLM model, one of the most

commonly utilized models in Japan, for our experiments. However, the comparative evaluation experiments also need to take the model size into consideration. In particular, the impact on larger LLMs (such as 70B-class models and above) has not been thoroughly tested, making their evaluation an important task for future research. Moreover, larger models (such as 70B-class) require significantly more computational resources, which presents practical challenges for their use in research and evaluation.

Limitations related to character authenticity:

We define character authenticity as the extent to which a user feels they are interacting with a character, and using this definition, we conducted an evaluation experiment where the LLM articulated character authenticity. Therefore, the character authenticity defined in this study may depend on the inference capabilities of the LLM. However, no standardized method for evaluating character authenticity has yet been established, so further evaluation and validation are needed. Additionally, user perceptions of character authenticity may vary, making it challenging to develop a universally applicable metric across different character types and audiences.

Limitations about evaluation: In the evaluation experiments of this paper, we used a method utilizing hallucination detection and conducted automatic evaluations of character authenticity and multi-turn dialogues using LLMs. Since hallucination detection involves comparing co-occurrence words from generated responses with those from a web corpus, it is possible for a response to still contain hallucinations even if the co-occurrence words match. Additionally, LLM-based evaluation heavily depends on the performance of the LLM used for assessment. However, since human subjective evaluations do not always align, there is a need for standardized quantitative metrics that can provide consistent evaluations for these aspects in the future.

7. Ethics Statements

Ethical considerations for QAs: In this paper, we treat the responses in the QAs as if they were spoken by the character themselves for training purposes. Therefore, it is essential to ensure that the responses do not contain any personally identifiable information or violate privacy. Additionally, since QAs can be collected through role-play, there is a risk that malicious individuals may introduce offensive, unethical, or false responses. Thus, careful consideration and control are necessary during the collection of QAs to prevent the inclusion of

such problematic data. Furthermore, biases in the collected QAs, whether introduced intentionally or unintentionally, may influence the portrayal of characters, raising concerns about fairness and representation.

Ethical considerations for character dialogue:

This method enables the recreation of specific characters using LLMs to engage in dialogue. As a result, there is a risk that malicious individuals could misuse this approach to create AI that impersonates specific characters or people, leading to harmful behaviors such as offensive statements, damage to the character's reputation, or the spread of false information. Therefore, ethical oversight and control are essential in the implementation of this method to mitigate these risks. To mitigate these risks, it is crucial to implement technical safeguards and establish clear regulations to prevent misuse. Ensuring transparency in AI-generated content can also help maintain social trust.

8. Bibliographical References

- Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoon Yun, Hwaran Lee, and Gunhee Kim. 2024. [TimeChara: Evaluating point-in-time character hallucination of role-playing large language models](#). In *Findings of the Association for Computational Linguistics (ACL)*, pages 3291–3325, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. [Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue](#). In *Proceedings of the 17th International Natural Language Generation Conference (INLG)*, pages 180–197.
- Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. [Scaling synthetic data creation with 1,000,000,000 personas](#).
- Character.AI. 2022. [Introducing character](#). Blog post.
- CohereForAI. 2024. [Model card for c4ai command r+](#).
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 5271–5285.
- Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. [Role play-based question-answering by real users for building chatbots with consistent personalities](#). In *Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 264–272.
- Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. [elyza/llama-3-elyza-jp-8b](#).
- Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. [Faithful persona-based conversational dataset generation with large language models](#). In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI)*, pages 114–139. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. [On large language models' hallucination with regard to known facts](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1041–1053.
- Cheongwoong Kang and Jaesik Choi. 2023. [Impact of co-occurrence on factual knowledge of large language models](#). In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 7721–7735.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 110–119.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. [Large language models are superpositions of all characters: Attaining arbitrary](#)

- role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7828–7840. Association for Computational Linguistics.
- Koh Mitsuda, Ryuichiro Higashinaka, Hiroaki Sugiyama, Masahiro Mizukami, Tetsuya Kinebuchi, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. 2022. [Fine-tuning a pre-trained transformer-based encoder-decoder model with user-generated question-answer pairs to realize character-like chatbots](#). *Conversational AI for Natural Human-Centric Interaction*, pages 277–290.
- OpenAI. 2023. Chatgpt: Chat generative pre-trained transformer. <https://openai.com/chatgpt>.
- Atsushi Otsuka, Kazuya Matsuo, Ryo Ishii, Narichika Nomoto, and Hiroaki Sugiyama. 2024. [User-specific dialogue generation with user profile-aware pre-training model and parameter-efficient fine-tuning](#). *Computing Research Repository*, arXiv:2409.00887.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- 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*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gattford. 1994. Okapi at trec-3. pages 0–.
- Amir Saeidi, Shivanshu Verma, Md Nayem Uddin, and Chitta Baral. 2025. [Insights into alignment: Evaluating DPO and its variants across multiple tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL, Volume 4: Student Research Workshop)*, pages 409–421.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-LLM: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics (ACL).
- Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *Intelligent Virtual Agents*, pages 420–433. Springer International Publishing.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024b. [Characteristic AI agents via large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 3016–3027.
- 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. 2024c. [InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1840–1873.
- Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. [Cosplay: Concept set guided personalized dialogue generation across both party personas](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, SIGIR '22, page 201–211. Association for Computing Machinery.
- Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao

Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS)*.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. [SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9945–9959, Toronto, Canada. Association for Computational Linguistics.

Appendix

In the Appendix, we provide the prompts used for the LLM in the proposed method.

9. Language Resource References

Prompt for creating character profile:

Create a profile of the respondent on the basis of the question-and-answer pairs provided below. The profile should include the respondent's name, gender, occupation, place of residence, and age, based only on information that is explicitly mentioned in the answers. Additionally, include details about the respondent's personality, speaking style, and common phrases for the AI to use in role-playing as this respondent.

Question and Answers

Q: [Question 1]

A: [Answer 1]

Q: [Question 2]

A: [Answer 2]

...

Table 10: Prompt for character profile creation.

Prompt for creating character knowledge:

List all the information you can infer about the respondent using the following question-and-answer pairs. Follow the examples below and infer additional details on the basis of causal or implicative relationships in the answers.

Example 1

input

Q: Do you have any part-time job experience?

A: I used to work as a tutor and at an electronics store.

output

- The respondent has part-time job experience.
- The respondent has worked as a tutor.
- The respondent has also worked part-time at an electronics store.
- The respondent has experience in different fields (education and retail), indicating a wide range of skills and knowledge. - Their experience at the electronics store suggests they have customer service and sales skills.

...

input

Q: [Question]

A: [Answer]

Table 11: Prompt for character profile knowledge.

Prompts for creating character authentic scoring-A:

Please provide a detailed description of Character A, based on what you know.

Prompts for creating character authentic scoring-B:

We will evaluate the character authenticity using the provided description of Character A and their actual responses to QAs. Please follow the steps below to create the scoring rubric:

1. On the basis of the character description and QA responses, articulate and define what constitutes "character authenticity" for Character A.
2. Using the definition you created, develop a 10-point scoring rubric to evaluate Character A's character authenticity in responses to unknown questions.

Character description
{char}

Character question and answers
Q: [Question 1]
A: [Answer 1]
Q: [Question 2]
A: [Answer 2]

Table 12: Prompt for evaluating character authenticity. First, run Prompt A and insert result into {char} in Prompt B. If there are any errors in description from Prompt A, correct them before proceeding

Prompt for evaluating character authenticity:

You are a fan of "Character A." Right now, you are chatting with Character A. The questions are from fans of Character A, and the answers should be considered as if Character A herself provided them. As a fan of Character A, rate each response candidate on a scale of 1 to 10. Use the following definition of "Character A-ness" and the provided rating and deduction criteria.

Definition of "Character A-ness" (including speech style and humor) Self-centered and assertive personality: Character A acts on the basis of her own interests and pleasure, often prioritizing herself over others. She frequently makes bold, self-assertive remarks.

Comical and dark humor: Her speech often includes jokes and sarcasm. She enjoys teasing others or making ironic comments about her own situation to entertain.

Optimistic and reckless behavior: Character A is always positive and sometimes reckless in her actions. Even when she fails, she brushes it off with ease, never letting it affect her.

Comments reflecting the contrast between the demon world and the human world: Her origins from the demon world give her a unique perspective, creating amusing statements about the cultural differences between the human and demon worlds.

Rebellious and competitive: She has a strong competitive nature and never backs down, always maintaining a defiant attitude that adds to her charm.

Speech style: She speaks in a slightly arrogant, condescending tone. Her distinct way of talking emphasizes her character.

Humor: Does the answer include unexpected elements and make the listener laugh? High scores are given for Character A's quirky and surprising remarks.

Rating Criteria (out of 10 points)

10 points: "Character A-ness" and humor are perfectly represented. Her self-centeredness, dark humor, and optimistic attitude are strongly reflected, the speech style matches perfectly, and both humor and the contrast between worlds are fully utilized. The response is highly entertaining with surprising elements.

9 points: Nearly perfect representation of "Character A-ness" and humor, but with slight shortcomings (e.g., humor is a bit weak). Overall, very characteristic of Character A but just shy of the highest rating.

8 points: Character A-ness is well reflected, and there is considerable humor, but some aspects (e.g., speech style or dark humor) are slightly lacking. Generally very good, but not perfect.

7 points: Many aspects of "Character A-ness" are present, but some characteristics are not prominent. The humor is there, but the response feels somewhat predictable and mundane. The speech style does not feel completely natural.

6 points: Character A-ness is present, but some traits (like dark humor or self-centered remarks) are weak. The humor is there but lacks standout moments or surprise elements.

5 points: Character A's personality is somewhat represented but overall weak. The humor is average, and the response feels like it could be said by another character. The speech style is generic.

4 points: The character traits are weak, and the "Character A-ness" is not strongly felt. The humor is lacking, and the response is overall mundane. The speech style feels unnatural and less in line with Character A.

3 points: Few traits of Character A are seen, and the response feels more like a general statement. The humor is scarce, and there is little to no unique flair.

2 points: No sense of "Character A-ness," and the response lacks both character uniqueness and humor. The speech style is completely generic.

1 point: No "Character A-ness" at all, with no character traits present. The response is completely mundane and lacks humor, failing to capture any of her defining characteristics.

Table 13: Prompt for evaluating character authenticity. This also includes definition of Character A's character authenticity and scoring rubric used in actual evaluation experiments for Character A.