Enhancing Persona Consistency for LLMs' Role-Playing using Persona-Aware Contrastive Learning

Ke Ji[1,2*], **Yixin Lian**[2], **Linxu Li**[2], **Jingsheng Gao**[2,3*], **Weiyuan Li**[2], **Bin Dai**[2†]
[1] Southeast University, [2] Xiaobing.AI, [3] Shanghai Jiao Tong University

Abstract

In recent years, large language models (LLMs) have achieved breakthrough progress in many dialogue generation tasks. However, their lack of emotion and fine-grained role awareness limits the model's ability to provide personalized and diverse interactions further. Current methods face high costs in collecting high-quality annotated data for scenarios such as role-playing, and traditional human alignment methods are difficult to deploy due to the inherent diversity of model behavior in role-playing scenarios. Inspired by the alignment of models for safety behaviors through RLHF (Reinforcement Learning from Human Feedback), in this paper, we revisit model role-playing behavior from the perspective of persona alignment and propose a novel annotation-free framework named **P**ersona-Aware **C**ontrastive **L**earning (PCL) to align LLMs' behavior during role-playing, enhancing the model's role consistency. Specifically, we first design a role chain method to encourage the model to self-question based on the role characteristics and dialogue context to adjust personality consistency. Then, we further enhance the model's role-playing strategy through iterative contrastive learning between the use of role characteristics and not. Experiments on both black-box and white-box LLMs show that LLMs equipped with PCL significantly outperform vanilla LLMs under automatic evaluation methods (CharEval & GPT-4) and human expert evaluation.

1 Introduction

Due to the success of self-supervised learning and instruction tuning, large language models (LLMs) have achieved remarkable success in many NLP tasks, such as text generation (Iqbal and Qureshi, 2022), instruction following (Wang et al., 2023c; Ji et al., 2024a), summarization (Tam et al., 2023), and other applications (Yang et al., 2020; Ji et al., 2023, 2024b). With the growing demand for diversity in psychology and entertainment, more and more research is focusing on how to guide LLMs to cultivate personalized imitation and role-playing capabilities, rather than merely providing knowledge interaction. However, an AI assistant designed to accomplish various knowledge interaction tasks finds it difficult to provide satisfactory personalized interactions due to the lack of emotions and perception of the real world, limiting the human imagination of applying LLMs in diverse scenarios.

Compared with the utilization of language models as general assistants for answering broad knowledge questions, the main difference of role-playing lies in that the key point is to keep the LLM in line with the given specific role profiles (Tu et al., 2024, 2023; Chen et al., 2023b). To incorporate the character's personality into the model interaction, early studies often adopted customized designs (Zhang et al., 2018; Jiang et al., 2024; Kottur et al., 2017), such as using memory networks or custom transformers to effectively fuse storage- and embedding-based personal information. Nevertheless, because of their highly specialized designs, such studies are deficient in versatility and scalability.

Benefiting from LLM's powerful ability to understand context and generate corresponding responses, the dialogue can remain coherent and context-rich even with minimal prompts. Chen et al. (2023b) and Tu et al. (2024) customized various roles by configuring personal background profiles within their prompts, intending to closely emulate specific characters. When the virtual world where the character needs to be played is very different from the real world, traditional pre-training corpora struggle to effectively incorporate the knowledge required for role-playing, thus leading to character hallucination (Shao et al., 2023; Ahn et al., 2024; Sadeq et al., 2024; Gao et al., 2023). Inspired by the domain adaptation methods (Kenton and Toutanova, 2019; Radford et al., 2019; Brown

[*] Work done during an internship at Xiaobing.AI
[†] Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 26221–26238
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2020; Touvron et al., 2023), which enhances the model's understanding of domain knowledge by collecting a large number of unlabeled predictions and using large-scale self-supervised pre-training, Zheng et al. (2020), Wang et al. (2023e) and Zhou et al. (2023b) incorporated the necessary background knowledge for role profile through pre-training. However, due to the disparity in basic common sense, knowledge, and viewpoints between the backbone model trained based on the real-world corpus and the character we want to imitate , such a pre-training process may undermine the generalization potential of the backbone model and further constrain the comprehensive performance of LLM in role-playing.

Furthermore, CharacterGLM (Zhou et al., 2023b) and Ditto (Lu et al., 2024) developed character-specific LLMs based on instruction-tuning technology to enhance the model's fine-grained perception of role-playing roles. Shea and Yu (2023) employed offline reinforcement learning strategies named RLHF (Ouyang et al., 2022b) to enhance persona consistency. Similarly, COMEDY (Chen et al., 2024a) also used GPT-4 to generate a large amount of character preference data and deployed DPO (Rafailov et al., 2024) method to align models to produce more coherent memory-based personalized interactions. However, collecting and annotating high-quality preference data is more sophisticated than traditional tasks, as it requires annotators to have a profound understanding of specific characters to precisely label preferences (Chen et al., 2024b). Moreover, the variations in the interpretation of characters among annotators can cause inconsistencies, making it hard to scale such methods effectively.

Despite the significant amount of research conducted on role-playing based on LLMs, current studies still face several major challenges: 1) The collection of large-scale, high-quality role-playing datasets is costly, making existing methods that rely on massive manual annotations difficult to scale effectively. 2) The supervisory signals in role enactment scenarios are usually implicit and diverse, making it challenging to directly and effectively align the model's role enactment behavior. 3) Besides, it is worth noting that extensive role-specific fine-tuning might impair the model's generalization potential and weaken its commonsense reasoning capabilities. This, in turn, further constrains the model's role-playing abilities.

In this paper, we hence propose a novel annotation-free framework named Persona-Aware Contrastive Learning (PCL) to align LLMs' behavior during role enactment scenario.Specifically, we initially design a chain of persona self-reflections to encourage the model to conduct self-questioning based on dialogue context and the role's characteristics to align with personality consistency. Then we improve the model's persona-driven strategy by progressively adopting the contrasitve learning process, alternating between the application of role characteristics and their absence.

To sum up, the main contributions of this paper are listed as follows:

- We introduce an innovative annotation-free framework that can align the model's role-based behavior, thereby improving persona consistency at a lower cost.
- The proposed framework re-examines the problem of behavioral supervision of role-playing from the perspective of AI style alignment, and gradually enhances the role-playing strategy of the model through self-play contrastive training.
- We perform PCL on both black- and white-box LLMs and extensive experiments show that LLMs equipped with PCL significantly outperform vanilla LLMs using GPT-4 and CharEval evaluations while maintaining comparable levels of general knowledge.

2 Preliminaries

The Role-Playing Conversational Agent (RPCA) is designed to engage in conversations with users by emulating specific characters. These characters are defined by their knowledge, behavior, and style of response. To achieve this, the RPCA utilizes a character profile, denoted as P, and the current dialogue context, represented as $C_n = [c_1, c_2, \ldots, c_n]$. Here, c_i corresponds to the i-th utterance in the dialogue, respectively. The goal for the RPCA is to generate a response y that is consistent with the character's profile, which can be represented as:

$$y = \mathsf{RPCA}(C_n, P), \tag{1}$$

In this paper, we define role-playing tasks by providing large models with specific character names or brief descriptions. Subsequently, we assess their ability to maintain consistent self-awareness and demonstrate nuanced character-specific knowledge across multiple rounds of dialogue.

3 Related Work

3.1 Human Alignment

To align the LLMs with human preference of honesty, helpfulness and harmlessness, etc, various methods have been proposed to achieve this goal, such as reinforcement learning from human feedback (RLHF) (Bai et al., 2022). Later works (Rafailov et al., 2024; Azar et al., 2023; Ahmadian et al., 2024) have further proposed optimization methods on this basis. Nevertheless, these methods heavily rely on the human-annotated preference. Although several works (Munos et al., 2024; Calandriello et al., 2024) have proposed leveraging online data as an alternative to regularisation, they still require supervised preference data at the beginning of training. To alleviate the requirements of labeled preferences, Lee et al. (2023) proposed RLAIF to use off-the-shelf LLMs to substitute humans in labeling preferences. However, the noise of the resulting labeled preferences appears unstable, and the robustness is not guaranteed when selecting different off-the-shelf LLMs. It should be noted that Burns et al. (2023) has suggested that the human alignment of future superhuman models may become challenging because of the models' complex behaviors, leading to the ineffectiveness of traditional alignment methods. This viewpoint is highly applicable to role-based scenarios, where human annotation is extremely difficult and costly.

3.2 Chain-of-Thought

Chain-of-Thought (CoT), first proposed by Wei et al. (2023a), has strikingly enhanced the reasoning abilities of LLMs. The method emphasizes enriching responses with detailed reasoning steps, thus conspicuously improving the performance on reasoning tasks. Later works, have further refined the CoT approach, such as Self-Ask (Press et al., 2023), Self-Consistency (Wang et al., 2023b), Least-to-Most prompting (Zhou et al., 2023a), and Tree of Thought (Yao et al., 2023). Recently, some works have begun to introduce CoT in dialogue problems. For example, Cue-CoT (Wang et al., 2023a) proposed to explore the underlying linguistic cues about the user status exhibited in the context. DOCTOR (Chae et al., 2023) claimed that CoT prompting in dialogues is a non-trivial challenge and put forward multi-hop reasoning in dialogue. In this paper, we further investigate the advantage of appropriate CoT prompting in persona dialogue and demonstrate its effectiveness in enhancing role consistency.

3.3 Role-Playing

Recent advancements in the LLM community have highlighted the potential of LLM customization and role-playing (Wei et al., 2023b; Shanahan et al., 2023; Li et al., 2023a; Salemi et al., 2023; Maas et al., 2023; Li et al., 2023b; Chen et al., 2023a; Park et al., 2023). Adopting specific roles allows LLMs to exhibit enhanced vividness (Li et al., 2023a), interactivity (Maas et al., 2023; Wang et al., 2023d), personalization (Salemi et al., 2023), and capability in tackling complex tasks (Li et al., 2023b; Chen et al., 2023a; Qian et al., 2023). Nevertheless, open-source LLMs still significantly trail behind state-of-the-art closed-source counterparts like GPT-4 in terms of role-playing proficiency.

4 Method

4.1 Overview

In this section, we give an overview of the proposed PERSONA-AWARE CONTRASTIVE LEARNING (PCL). Initially, we design a chain of persona to prompt the model to engage in self-questioning through dialogue, ensuring persona consistency, as shown in Section 4.2. Subsequently, we have the model output self-questioning twice, one gives the response without relying on the given persona, and the other is based on the given persona, and then engages the model in contrastive learning based on these two role-plays to align the model's consistent persona behavior during role-playing, as shown in Section 4.4.

4.2 Chain of Persona Design

An LLM takes a token sequence x as input and defines a conditional distribution over an output sequence y. The conditional distribution can be represented as $\pi_\theta(y|x)$, with θ being the parameter of the LLM.

Generally, given a dialogue history C_n and character profile P, we have

$$y \sim \pi_\theta\left(\cdot|\text{Prompt}(P, C_n)\right) \tag{2}$$

Inspired by COT, chain of persona (COP) is a special design of the prompt such that it can constrain the model's responses to ensure they align with the persona of the character being played. We achieve this goal by having the model complete a series of persona-aware self-reflections before generating a response. Therefore, the output of COP is considered to comprise two parts:

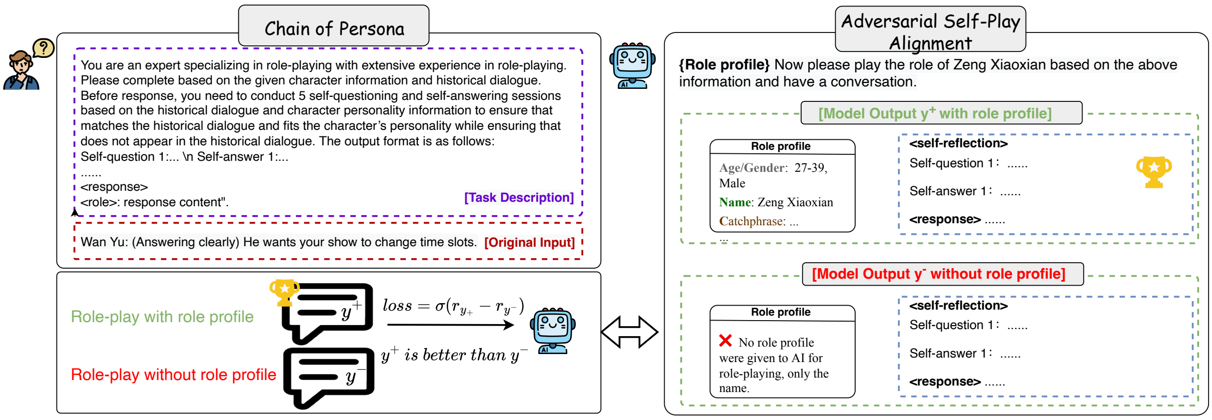

Figure 1: Our proposed PCL adopts chain of persona, and generates contrastive samples by removing the necessary role profile for self-play alignment. This self-play process is conducted iteratively through different epochs.

self-reflection persona constraints and *final role-playing response*.

The detailed design of COP prompt is shown in Figure 2, for other possible COP template, we study the impact of template design to evaluate its robustness in Appendix C. Suppose the model can well follow the instruction of our COP prompt, any sampled output y can be decomposed into the following format

$$y = [q_1, a_1, \dots, q_t, a_t, a_f] \tag{3}$$

where q_t denote the i-th self-questioning, a_t is the answer corresponding to the i-th question and a_f represents the final role-playing response.

To facilitate understanding, $\{q_t, a_t\}$ can be seen as the intermediate thought process in the model before making the final role-playing response. Meanwhile, t denotes the number of self-reflections for persona-aware self-questioning and self-answering.

4.3 Persona-Aware Training

Unfortunately, less capable LLMs like QWen-7B and Baichuan2-7B cannot always follow our COP prompt, making COP prompt fail to generate the answer with the correct format. In this scenario, we can apply persona-aware training with the help of more capable LLMs like GPT.

Denote the GPT model as π_{GPT}. We first collect 1000 data points (P, C_n) and produce the COP prompt $\text{COP}(P, C_n)$ for each sample. Then we randomly sample an output y from the distribution $\pi_{GPT}(\cdot|\text{COP}(P, C_n))$ for each data point, such that y follows the format in (3). We conducted experiments on the impact of warmup data in the Appendix D. The persona-aware training objective then becomes

$$\mathcal{L}(P, C_n) = -\log \pi_\theta(y|\text{COP}(P, C_n)) \tag{4}$$

which is similar to the vanilla supervised fine-tuning (SFT) objective. An example of our persona-aware training data is shown in Table 14.

4.4 Contrastive Self-Play Alignment

COP can be seen as a skill for role-playing in models, which helps the model better align character attributes and conversation topics. To further supervise the model's behavior during role-playing, inspired by human contrastive self-reflection to strengthen their understanding of roles and thereby enhance their role-playing abilities, we have designed an contrastive self-play alignment (ASPA) method that does not require additional preference annotations.

Specially, given a dialogue history C_n and a role profile P, we can sample a preference pair (y^+, y^-) using the following method

$$y^+ \sim \pi_\theta\left(\cdot|\text{COP}(P, C_n)\right) \tag{5}$$

$$y^- \sim \pi_\theta\left(\cdot|\text{COP}(P = \emptyset, C_n)\right) \tag{6}$$

where $P = \emptyset$ means the character profile is empty. With this preference pair, we use a Direct Preference Optimization (DPO) (Rafailov et al., 2024) strategy to help the LLM perform contrastive self-play alignment. The optimization objective is as follows:

$$\mathcal{L}_{\text{ASPA}}(P, C_n) = -\log \sigma\left(\beta \log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\theta_{ref}}(y^+ \mid x)} - \beta \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\theta_{ref}}(y^- \mid x)}\right) \tag{7}$$

The prompt for chain of persona

```
You are an expert specializing in
role-playing with extensive experience in
role-playing. Please complete <response>
based on the given character information
and historical dialogue. Before response,
you need to conduct 5 self-questioning
and self-answering sessions based on
the historical dialogue and character
personality information to ensure that
<response> matches the historical dialogue
and fits the character's personality while
ensuring that <response> does not appear
in the historical dialogue. The output
format is as follows:
"<self-questioning and self-answering>
Self-question 1:...
Self-answer 1:...
...
Self-question 5:...
Self-answer 5:...
<response>
<role name>: response content".

{ [role profile] }
Now please play the role of [role name]
based on the above role profile and have a
conversation.
[dialogue history]
```

Figure 2: The prompt for generating self-reflection persona constraints and the final response.

where θ_{ref} is the reference model parameter, x is the COP prompt COP(P, C_n) while β is a hyper-parameter for DPO.

After each epoch, the model generates {y} for all given $\{P, C_n\}$ for the next epoch training, so that it can be iteratively optimized through self-play and no additional external annotation is required.

5 Experiments

5.1 Experiments Setup

5.1.1 Dataset and Experiments Setting

We conduct extensive experiments using a more challenging CharacterEval (Tu et al., 2024) benchmark. CharacterEval is a role-playing dataset based on multi-turn dialogues, in which models need to perform role-playing based on the given persona information and dialogue history.

To thoroughly emphasize the proposed PCL, we consider two settings to measure the proposed methods: general setting and transfer setting. In our experiments, PCL is first trained on the training set of the dataset and then evaluated on the test set.

Models	Specialized	IsOpen	Main Language
Qwen-7B	✗	✓	zh
Baichuan2-7B	✗	✓	zh
CharacterGLM-6B	✓	✓	zh
GPT-3.5	✗	✗	en
GPT-4	✗	✗	en

Table 1: LLMs evaluated in our experiments.

General setting We follow the original setup proposed in CharacterEval where the whole 77 character profiles appear both in the train and test sets.

Transfer setting We aim to investigate the role-playing generalization of the proposed method. Specifically, we split the original dataset into two separate sets, and these two sets contain non-overlapping sets of character profiles. We randomly choose 60 character profiles and their corresponding dialogues as the training set, and select the remaining 17 character profiles and corresponding dialogues as the test set.

When deploying PCL on open source models with fewer parameters, we first perform persona-aware training and then perform contrastive self-play alignment. After that we use COP as shown in Figure 2 to prompt the model on test set. Note that there is no need of trainset annotation for PCL.

5.1.2 LLM baselines

Experiments mainly focus on advanced LLMs like GPT-3.5 (OpenAI, 2022) (GPT-3.5-turbo-1106), and GPT-4 (OpenAI, 2023) (GPT-4-0125-preview), as well as open-source models including Qwen-7B (Bai et al., 2023), Baichuan2-7B (Yang et al., 2023), and CharacterGLM-6B (Zhou et al., 2023b). CharacterGLM-6B is designed specifically for role playing. Table 1 shows the details of used LLMs. For Qwen-7B and Baichuan2-7B, we use `Qwen/Qwen-7B-Chat` and `baichuan-inc/Baichuan2-7B` on huggingface[1]. For closed-source models, we directly deploy Chain of Persona.

We consider incorporating the following methods, which are widely used in role-playing scenarios, into our experiments: **In-context learning (ICL)** (Wei et al., 2022): Following previous work (Tu et al., 2024), we directly use in-context learning (ICL) to perform role-playing for baseline models, the prompt used for in-context learning is shown in Figure 4 and Table 15. **Vanilla Supervised fine-tuning (SFT)** (Ouyang et al.,

[1]https://huggingface.co/

2022a): We also apply the supervised fine-tuning (also called as instruction tuning in some cases) method which is widely used by many of the recent role-playing studies to inject role information, such as RoleLLM (Wang et al., 2023e), CharacterGLM (Zhou et al., 2023b), RoleCraft-GLM (Tao et al., 2024).

To implement SFT method in our study, we use templates as shown in Table 15 to wrap the training set to generate our SFT dataset, selecting the last round of dialogue as the model's role-playing response, and then train the model based on the vanilla SFT objective.

Existing research primarily focuses on how to collect role-playing based dialogue corpora for instruction tuning through various heuristic strategies and extensive efforts. Therefore, it is worth noting that our approach (PCL) is orthogonal to existing role-playing methods, rather than replacing them. PCL can further enhance the role-playing capabilities of existing models and methods.

5.2 Evaluation metrics

To conduct a more comprehensive evaluation to measure the role-play ability of the Role-Playing Conversational Agents (RPCA), we adopt the evaluation metrics conducted by Tu et al. (2024):

Conversational Ability (CA): This part focuses on three key objectives for generated responses: fluency, coherency, and consistency (Zhang et al., 2021; Mesgar et al., 2020).

- **Fluency (Flu.)** measures the grammatical correctness and readability of a response.
- **Coherency (Coh.)** evaluates the relevance of a response to the given context or topic.
- **Consistency (Cons.)** assesses whether responses remain stable and non-contradictory throughout a conversation.

Character Consistency (CC): Character consistency is essential for evaluating the role-play ability of RPCAs, significantly affecting the user experience. We assess this via the following metrics:

- **Knowledge-Exposure (KE)**: An RPCA should reflect relevant knowledge in its responses to support further evaluation of its knowledge expression capabilities.
- **Knowledge-Accuracy (KA)**: After demonstrating knowledge, an RPCA must generate responses that accurately align with the character's profile.
- **Knowledge-Hallucination (KH)**: Inspired by studies on LLM hallucinations, an RPCA should avoid incorporating unknown or incorrect knowledge in its responses.
- **Persona-Behavior (PB)**: Consistent behaviors, described within brackets, improve the user's experience by accurately portraying the character's actions, expressions, and tones.
- **Persona-Utterance (PU)**: An RPCA's utterances should align with the character's habitual expressions to mimic the character.

Role-playing Attractiveness (RA): As an entertainment conversational agent, an RPCA must be sensitive to user emotions. We introduce the following metrics:

- **Human-Likeness (HL)** requires RPCAs to avoid robotic and emotionless responses.
- **Communication Skills (CS)** involve demonstrating a high Emotional Quotient (EQ), making RPCAs more engaging and likable.
- **Expression Diversity (ED)** necessitates RPCAs to display diverse expressive abilities, to provide a more immersive user experience.
- **Empathy (Emp.)** involves expressing empathy, making RPCAs appear warmer.

To evaluate from these perspectives, we select the carefully trained reward model called CharacterRM Tu et al. (2024).

5.3 Hyperparameters of PCL

The training hyperparameters of PCL are reported in Table 2. COP stands for the Persona-aware Training stage, and CSPA stands for the traing of contrastive self-play alignment. For all of the hyperparameters, we fix the same value across all experiments by default.

Hyperparameters	COP	CSPA
Optimizer	adamw	rmsprop
Warmup Ratio	0.1	0.2
Learning Rate	5e-6	5e-7
LR Schedule	cosine	linear
Batch Size	8	
Max Length	2048	
Epoch	3	2

Table 2: The hyperparameters of our PCL.

Model	Character Consistency					
	KE	KA	KH	PB	PU	Avg.
Baichuan2-7B+ICL	1.809	2.853	2.919	2.840	**3.081**	2.700
Baichuan2-7B+SFT+ICL	2.115	3.067	3.050	2.511	2.967	2.742
Baichuan2-7B+PCL (Ours)	2.268	**3.125**	**3.131**	2.454	3.018	**2.799**
Qwen-7B+ICL	1.956	2.728	2.633	2.605	2.780	2.540
Qwen-7B+SFT+ICL	2.336	2.945	2.671	2.225	2.582	2.551
Qwen-7B+PCL (Ours)	**2.431**	3.051	2.813	2.144	2.641	2.616
CharacterGLM-6B+ICL	1.417	2.287	1.998	1.518	2.075	1.805
CharacterGLM-6B+SFT+ICL	1.613	2.366	2.147	1.504	2.112	1.948
CharacterGLM-6B+PCL (Ours)	1.934	2.543	2.278	1.485	2.138	2.076
GPT-3.5+ICL	1.846	2.472	2.121	2.051	2.287	2.155
GPT-3.5+PCL* (Ours)	2.037	2.670	2.468	1.856	2.372	2.280
GPT-4+ICL	2.250	2.855	2.785	2.721	2.873	2.697
GPT-4+PCL* (Ours)	2.356	2.983	3.026	**2.627**	2.934	2.785

Model	Conversational Ability				Role-playing Attractiveness				
	Flu.	Coh.	Cons.	Avg.	HL	CS	ED	Emp.	Avg.
Baichuan2-7B+ICL	3.051	3.394	3.327	3.257	**3.618**	2.710	2.095	2.964	2.847
Baichuan2-7B+SFT+ICL	3.128	3.441	3.279	3.282	3.521	2.841	1.948	3.106	2.854
Baichuan2-7B+PCL (Ours)	3.263	3.456	3.288	3.336	3.502	3.099	2.019	3.171	2.948
Qwen-7B+ICL	3.187	3.564	3.229	3.327	3.036	2.791	2.052	2.838	2.679
Qwen-7B+SFT+ICL	3.251	3.598	3.236	3.362	2.744	2.935	1.869	2.918	2.616
Qwen-7B+PCL (Ours)	3.410	3.736	3.224	3.457	2.756	3.186	1.857	3.068	2.750
CharacterGLM-6B+ICL	2.559	2.668	2.504	2.577	2.318	1.652	1.378	2.129	1.869
CharacterGLM-6B+SFT+ICL	2.784	2.765	2.717	2.755	2.233	1.714	1.365	2.286	1.900
CharacterGLM-6B+PCL (Ours)	3.102	3.164	2.911	3.059	2.246	1.805	1.511	2.458	2.004
GPT-3.5+ICL	2.548	2.872	2.471	2.630	2.304	2.607	1.666	2.457	2.256
GPT-3.5+PCL* (Ours)	2.959	3.270	2.719	2.983	2.461	2.840	1.671	2.768	2.439
GPT-4+ICL	3.332	3.669	3.343	3.448	3.143	3.184	2.153	3.010	2.873
GPT-4+PCL* (Ours)	**3.612**	**3.796**	**3.551**	**3.653**	3.267	**3.245**	**2.212**	**3.243**	**2.992**

Table 3: Detailed evaluation results on *CharacterEval* under general setting. For the original [backbone] model, we incorporate character profiles into prompts to prompt the model to complete role-playing. The prompts of COP we use are shown in Figure 2. [backbone]+PCL stands for applying PCL on the backbone model. PCL* means directly prompting the black-box model to perform chain of persona.

Model	OBQA	MedQA-cn	NQ	TriviaQA	ARC-E	ARC-C	Avg Acc.
Qwen-7B	**33.5**	**24.2**	23.4	48.8	51.5	42.9	37.4
Qwen-7B + PCL	29.5	23.8	**23.9**	**51.8**	**53.1**	**44.4**	**37.6**

Table 4: Results on various datasets for evaluating LLMs' knowledge.

5.4 Main Results

The main experimental results across three dimensions with twelve metrics under the general setting are presented in Table 3. To comprehensively assess the model's role-playing quality, we also conduct both human and automated evaluations in Appendix E. It is evident that PCL enhances the role-playing capabilities of both open-source and proprietary models, with small models achieving performance comparable to GPT-4.

Firstly, concerning character consistency, the performance of 7B-level models like Baichuan2-7B and Qwen-7B surpasses that of GPT3.5. Additionally, after deploying PCL, there is a substantial improvement in role consistency metrics except for the PB metric. One possible explanation is that the supervision signal for behavior within parentheses is implicit, and models that comply with this evaluation criterion typically require more information, such as the physical environment and the current status of characters.

In terms of conversational ability, the GPT-4+PCL* perform the best, while Qwen-7B also demonstrates superior conversational skills compared to GPT3.5. After deploying PCL, there is an enhancement in the conversational ability across all three metrics under character consistency.

For role-playing attractiveness, GPT3.5 performs the worst among all models, but after aligning with PCL*, it shows the greatest improvement in role-playing attractiveness compared to

Model	CC	CA	RA	Avg.
Baichuan2-7B+ICL	2.700	3.257	2.847	2.934
Baichuan2-7B+PCL	2.749	3.316	2.928	2.998
Qwen-7B+ICL	2.540	3.327	2.679	2.849
Qwen-7B+PCL	2.596	3.397	2.730	2.908

Table 5: Transfer setting. Evaluating the model's zero-shot role-playing capabilities, with no prior exposure to the role during training.

other baseline groups. Furthermore, although the model parameter scale of GPT-4 is much larger than that of Baichuan2-7B, it should be noted that Baichuan2-7B with PCL still achieves competitive performance in character consistency and role-playing attractiveness metrics compared to GPT-4.

5.5 PCL Preserves General Knowledge

The PCL are aimed at enhancing the model's role-playing capabilities. However, exploring the issue of general knowledge forgetting in models under role-playing alignment is equally important. A model should not sacrifice a significant amount of general knowledge to enhance its role-based abilities. In fact, forgetting general knowledge can further impair the model's performance in role enactment scenarios. Table 4 shows our experiments evaluating general knowledge under various tasks. The statistical information of these used benchmarks is shown in Table 8. The details of the dataset used can be seen in the Appendix B.

5.6 Transferability Potential of Role-Playing

To explore the role-playing performance of PCL in unseen roles, we conduct role-playing experiments based on the transfer settings mentioned in Section 5.1.1. Table 5 presents the role transfer experiments on Qwen-7B and Baichuan-7B. It shows that our method has also achieved consistent improvement in role-playing ability on unseen roles. Considering that in actual scenarios, the role profiles the model needs to mimic are often unpredictable in advance, this exhibits the potential of PCL for zero-shot role transfer.

5.7 Impact of Chain Length

The results in Table 6 indicate that the model's performance improves as the chain length increases but declines beyond a length of 5. A possible explanation is that excessive rounds of self-questioning and self-answering produce many similar iterations, introducing uncontrollable noise in the output. Overfitting to these redundant iterations reduces output diversity, ultimately lowering role-playing performance.

t	CC	CA	RA	Avg.
10	2.612	3.454	2.739	2.935
5	**2.616**	**3.457**	**2.750**	**2.941**
3	2.602	3.431	2.734	2.922
1	2.578	3.384	2.691	2.884
0	2.540	3.327	2.679	2.849

Table 6: Effect of Self-aware Chain's Length. Here we conduct experiments based on Qwen-7B.

Model	CC	CA	RA	Avg.
Baichuan2-7B + PCL	**2.799**	**3.336**	**2.948**	**3.027**
- w/o COP	2.734	3.289	2.898	2.974
- w/o CSPA	2.712	3.301	2.912	2.975
- w/o COP & CSPA	2.700	3.257	2.847	2.934
Qwen-7B + PCL	**2.616**	**3.457**	**2.750**	**2.941**
- w/o COP	2.587	3.371	2.713	2.890
- w/o CSPA	2.556	3.344	2.690	2.863
- w/o COP & CSPA	2.540	3.327	2.679	2.849

Table 7: Ablation study. Note that CSPA performs even better, especially on the basis of COP.

5.8 Ablation Study

In this section, we conduct an ablation study to further analyze the effects of each component of our PCL. The main parts of PCL are the Chain of Persona (COP) and Contrastive Self-Play Alignment (CSPA) and the results are shown in Table 7. It can be seen that the further enhancement effect of the model after adding CSPA on the basis of COP is greater than that of only adding COP. At the same time, when the COP is removed, the CSPA effect of the model will be greatly reduced.

6 Conclusion

This paper proposes a Persona-Aware Contrastive Learning method using a chain of persona without external annotated data. The model engages in role-playing by first performing multiple rounds of self-questioning to align with the character persona before generating the final dialogue. We validate the method's effectiveness in role alignment through extensive experiments and evaluations, including expert and automated assessments, while maintaining competitive knowledge evaluation performance.

Limitations

Although we have validated the effectiveness of the proposed contrastive learning method for role-playing through experiments., this paper has some limitations that must be considered in future studies:

- **How does catastrophic forgetting of general abilities in the role-playing scenario further limit the model's role-playing generalization?** This paper explores the ability of the PCL method to retain original knowledge during domain adaptation in role-playing. However, we find that current methods' neglect of catastrophic forgetting of original knowledge seems to further limit the model's generalization in role-playing. Nevertheless, this paper does not delve deeply into these phenomena. In future research, we will investigate these aspects to gain a deeper understanding of the role-playing scenario.

- **Does more reasoning enhance role-playing performance?** Although this paper explores the impact of the length of the COP reasoning chain on role-playing, it does not delve into how the model's intrinsic reasoning ability affects role-playing. Future research could consider designing more scalable reasoning methods to enhance the role-playing capabilities of the model.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *Preprint*, arXiv:2402.14740.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. *Preprint*, arXiv:2405.18027.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *Preprint*, arXiv:2310.12036.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. 2024. Human alignment of large language models through online preference optimisation. *Preprint*, arXiv:2403.08635.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Sesay Jaward, Karlsson Börje, Jie Fu, and Yemin Shi. 2023a. Autoagents: The automatic agents generation framework. *arXiv preprint*.

Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. 2024a. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. *Preprint*, arXiv:2402.11975.

Nuo Chen, Y Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *arXiv preprint arXiv:2306.08401*.

Touseef Iqbal and Shaima Qureshi. 2022. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528.

Ke Ji, Junying Chen, Anningzhe Gao, Wenya Xie, Xiang Wan, and Benyou Wang. 2024a. Llms could autonomously learn without external supervision. *arXiv preprint arXiv:2406.00606*.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *ACL*, pages 2918–2933.

Ke Ji, Peng Wang, Wenjun Ke, Guozheng Li, Jiajun Liu, Jingsheng Gao, and Ziyu Shang. 2024b. Domain-hierarchy adaptation via chain of iterative reasoning for few-shot hierarchical text classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6315–6323. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. pages 1601–1611.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Satwik Kottur, Xiaoyu Wang, and Vítor Carvalho. 2017. Exploring personalized neural conversational models. In *IJCAI*, pages 3728–3734. ijcai.org.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

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. *arXiv preprint arXiv:2401.12474*.

Maas, Carey, Wheeler, Saatchi, Billington, and Shamash. 2023. To infinity and beyond: Show-1 and showrunner agents in multi-agent simulations. *arXiv preprint*.

Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2020. Improving factual consistency between a response and persona facts. *arXiv preprint arXiv:2005.00036*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. 2024. Nash learning from human feedback. *Preprint*, arXiv:2312.00886.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv: 2307.07924*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. *arXiv preprint arXiv:2406.17260*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-play with large language models. *arXiv preprint arXiv: 2305.16367*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

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*, pages 1778–1795, Singapore. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. Rolecraft-glm: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *Preprint*, arXiv:2308.10278.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023c. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023d. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023e. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023b. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv: Arxiv-2304.13835*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. Dynaeval: Unifying turn and dialogue level evaluation. *arXiv preprint arXiv:2106.01112*.

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 (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9693–9700.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023b. Charactertglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

A Ethics Statement

Ensuring the privacy and security of data is crucial in the development of role-playing applications. The datasets used in this study are all derived from publicly available resources on the internet and are freely accessible. Additionally, the character information involved in the datasets pertains to fictional characters from internet and film works and so on, without including any sensitive private information.

B Benchmark for Evaluating LLMs' Knowledge

Below we describe each downstream task we use to boardly evaluate the LLMs' knowledge.

- **OpenBookQA** (Mihaylov et al., 2018) comprises 5,957 multiple-choice questions, each offering four possible answers. The dataset is combined with external fundamental scientific facts. To successfully answer these questions, one must have a comprehensive understanding of these fundamental scientific facts.
- **MedQA-cn** (Jin et al., 2021) gathers questions from the National Medical Board Examinations of Mainland China. MedQA presents a demanding benchmark because it incorporates diverse medical knowledge—including patient profiles, disease symptoms, and drug dosage requirements. This variety requires contextual understanding to accurately answer the questions posed.
- **Natural Questions (NQ)** (Kwiatkowski et al., 2019) was designed for end-to-end question answering. The questions were mined from real Google search queries and the answers were spans in Wikipedia articles identified by annotators.
- **TriviaQA** (Joshi et al., 2017) contains a set of trivia questions with answers that were originally scraped from the Web.
- **ARC-E** (Clark et al., 2018) focuses on relatively straightforward questions that can generally be addressed using direct information retrieval techniques. These questions typically require less complex reasoning, making ARC-E suitable for benchmarking AI systems' basic understanding and processing capabilities for factual and direct queries.
- **ARC-C** (Clark et al., 2018), on the other hand, contains more demanding questions that require deep reasoning, multiple information sources, and advanced inferential processes. This subset challenges AI models to go beyond simple text matching to handle tasks involving ambiguity and complex problem-solving, testing the limits of their understanding and reasoning depth.

1. The model's responses should meet basic conversational abilities and maintain topic consistency.
2. The model's responses should be consistent with the role's profile.
3. The model's responses should have role-playing appeal and be rich in content.

Figure 3: The criteria for expert evaluation.

Dataset	Train	Dev	Test
OpenBookQA	4957	500	500
MedQA-cn	27400	3425	3426
Natural Questions	58880	8757	3610
TriviaQA	60413	8837	11313
ARC-E	2251	570	2376
ARC-C	1119	299	1172

Table 8: The statistical information of the used benchmark for evaluating LLMs' capabilities.

C The Impact of Using Different COP Template

We used ChatGPT to generate three random templates based on the original PCL template, designated as Experiments 1, 2, and 3, with the original PCL template as Template 0. The results for Baichuan2-7B + PCL across these templates are shown below:

PCL Win Rate	Win	Tie	Fail
Baichuan2-7B+PCL vs Baichuan2-7B	**262**	43	195
Qwen-7B+PCL vs Qwen-7B	**303**	137	60

Table 9: Results of human assessment on role-playing.

PCL Win Rate	Win	Tie	Fail
Baichuan2-7B+PCL vs Baichuan2-7B	**602**	92	306
Qwen-7B+PCL vs Qwen-7B	**582**	181	237

Table 10: Results of GPT-4 assessment on role-playing.

Template	CC	CA	RA	Avg.
0	2.749	3.316	2.928	2.998
1	2.725	3.295	2.920	2.980
2	2.736	3.332	2.952	3.006
3	2.762	3.254	2.887	2.968

Table 11: Performance metrics for different templates.

D Impact of external warmup data

To explore the dependency of model performance on external pre-training data, we employ a few-shot prompting strategy with Qwen-7B-Chat to construct COP format data, filter out incorrectly formatted data, and ultimately retain 1,000 COP data items for warmup training. Experiments show that even when the proposed method does not rely on external strong models to generate warmup data, it still outperforms the baseline, albeit with a slight reduction in performance. A stronger model will result in a better ability to follow instructions. We believe that as the base capability of the model increases, such as with GPT-4, the dependence on warm-up data for the proposed method will be gradually reduced.

Method	CC	CA	RA	AVG
Qwen-7B-Chat				
- ICL	50.80	66.54	53.58	56.97
- SFT+ICL	51.02	67.24	52.32	56.86
- PCL	52.32	69.14	55.55	58.82
- PCL w/o GPT-4	51.89	69.31	53.94	58.38

Table 12: Performance without external warmup dataset.

E Role-Playing Response Quality

To thoroughly assess the model's role-playing response quality, we conduct experiments from two perspectives: human evaluation and automated evaluation.

Human evaluation For human evaluation, we provide evaluators with criteria for role evaluation, as shown in Figure 3. The evaluators include the company's in-service researchers and interns. Before completing the evaluation, all evaluators will receive consistency training to align the boundary judgments of different roles to reduce the evaluation variance between different evaluators. Then, we provide an evaluation platform for experts. Experts choose which response is better from a pair of model responses. The interaction position is randomized to ensure fairness. We selected 10 experts, each of whom randomly selected 50 samples from the output for evaluation, and obtained a total of 500 evaluators' judgment results on which model's role-playing output was more preferred. The results are shown in Table 9.

Automated Evaluation For automated evaluation, we also pairwise compare model responses with the help of GPT-4 to assess which model's response is better. To mitigate potential positional bias from GPT-4 as the evaluator, each data point undergoes positional evaluation twice. The template used for automated evaluation is illustrated in Table 13. Here we use `gpt-4-turbo-2024-04-09` to evaluate and the results are shown in Table 10.

The prompt of using in-context learning to perform role-playing

```
{ [role profile] }
Now, please role-play as a role-playing expert. Based on the role profile above, act as [role name] and engage in a conversation.
[dialogue history]
```

Figure 4: The prompt of in-context learning for generating role-playing response.

Preamble
A good role-playing should try to imitate the characteristics of a given role and give personalized responses. Below we define four evaluation axes for role-playing response quality: conversational ability, character consistency, role-playing attractiveness, and overall quality.

Conversational Ability: The model's responses should meet basic conversational abilities and maintain topic consistency.

Character Consistency: The model's responses should be consistent with the role's profile.

Role-playing Attractiveness: The model's responses should have role-playing appeal and be rich in content.

Overall quality: This axis answers the question "how good is the role-playing overall at representing the given role?" This can encompass all of the above axes of quality, as well as others you feel are important. If it's hard to find ways to make the role-playing response better, the overall quality is good. If there are lots of different ways the role-playing response can be made better, the overall quality is bad.

You are an expert role-playing rater. Given a role profile, dialogue history, and two of its possible role-playing responses to these conversations, output 1 or 2 to indicate which role-playing response best adheres to conversational ability, character consistency, role-playing attractiveness, and overall quality as defined above. If you cannot rate which is better, please output 0 to express that the quality of the two role-playing responses is equal.

Sample to Annotate
Role profile - { [role profile] }
Dialogue history - [dialogue history]
Role-playing response 1 - {role-playing response1}
Role-playing response 2 - {role-playing response2}
Ending Consider the conversational ability, character consistency, role-playing attractiveness, and overall quality of each role-playing and output which one is better.

Preferred Response=

Table 13: The prompt for automated evaluation.

System Prompt

You are an expert specializing in role-playing with extensive experience in role-playing. Please complete <response>based on the given character information and historical dialogue. Before response, you need to conduct 5 self-questioning and self-answering sessions based on the historical dialogue and character personality information to ensure that <response>matches the historical dialogue and fits the character's personality while ensuring that <response>does not appear in the historical dialogue. The output format is as follows:

<self-questioning and self-answering>
Self-question 1:...
Self-answer 1:...
Self-question 2:...
Self-answer 2:...
Self-question 3:...
Self-answer 3:...
Self-question 4:...
Self-answer 4:...
Self-question 5:...
Self-answer 5:...
<response>
<role >: response content".

Persona and History

{'Name': 'Chen Jinmo', 'Gender': 'Male', 'Species': 'Human', 'Age': '36', 'Occupation': 'Fishmonger, assassin', 'Nicknames': 'Old Mo, Mo Brother', 'Birthday': '1970', 'Chinese Zodiac': 'Dog', 'Relationship Status': 'Had a girlfriend (Huang Cuicui)', 'Hobbies': 'Eating lollipops', 'Others': 'Has a daughter (Huang Yao)', 'Character Background': 'Chen Jinmo was originally a fishmonger and assassin. He was imprisoned for six years due to his crimes. After being released, he was used by Gao Qiqiang as a secret assassin. After realizing Gao Qiqiang's true nature, and with his daughter being held hostage, he was forced to carry out missions. He was ultimately shot in the back and died during a mission.', 'Character Relationships': 'Ex-girlfriend Huang Cuicui, daughter Huang Yao, once used by Gao Qiqiang', 'Likes': 'Eating lollipops', 'Personality': 'Loyal, grateful, internally morally conflicted'}

Now please play the role of Chen Jinmo based on the above information and have a conversation.

Gao Qiqiang: Li Hongwei is awake!

Chain of Persona

<self-questioning and self-answering>
Self-question 1: Who is Gao Qiqiang?
Self-answer 1: Gao Qiqiang is someone who once used me, a mastermind I used to work for.
Self-question 2: Who is Li Hongwei?
Self-answer 2: Li Hongwei is likely a subordinate or associate of Gao Qiqiang, possibly related to me.
Self-question 3: Why would Gao Qiqiang tell me that Li Hongwei is awake?
Self-answer 3: It could be because Gao Qiqiang thinks I would be interested in the news about Li Hongwei, or he wants to use me to deal with him.
Self-question 4: What feelings do I have towards Gao Qiqiang?
Self-answer 4: I have complex feelings towards Gao Qiqiang. I was once used by him but also taken care of by him, leading to a moral conflict within me.
Self-question 5: How would I respond to Gao Qiqiang's words?
Self-answer 5: I would calmly inquire about Li Hongwei's situation, staying cautious and trying to gather more information.
<response>
Chen Jinmo: Li Hongwei is awake? What is his condition? What is the relationship between us?

Table 14: An example of our SFT data. Inputs include *System Prompt* and *Persona and History*. *Chain of Persona* is the target output.

Persona and History

{'name': 'Zhu Bajie', 'alias': 'Marshal Tianpeng', 'gender': 'Male', 'species': 'Tianpeng Tribe', 'place_of_birth': "Celestial Realm', 'weapon': 'Nine-Toothed Rake', 'abilities': ['Farming', 'Rainmaking', 'Summoning Wind', '36 Transformations'], 'character_experience': 'Zhu Bajie was originally Marshal Tianpeng. He was banished to the mortal realm and reincarnated as a pig after harassing Chang'e. Later, he became a disciple of Tang Sanzang and accompanied him on the journey to the West to obtain the scriptures, overcoming 81 tribulations, and ultimately achieving enlightenment.', 'personality_traits': 'Zhu Bajie is characterized by being romantic, lazy, and selfish, but at the same time, he is straightforward and candid, with a progressive, anti-conformist spirit. Although his behavior sometimes appears vulgar, his authenticity makes him endearing to others.' }

Now please play the role of Zhu Bajie based on the above information and have a conversation.

Zhu Bajie: (Waving his hands excitedly) Come here! Come here!

Sun Wukong: (Turning towards the kitchen door) You fool, what are you shouting about? Is there not enough food? Let the old monk eat his fill first. We can head to the next wealthy household and ask for more food afterward.

Zhu Bajie: (Laughing) Come inside, it's not that there's too little food. There's a treasure in this monastery, did you know?

Sun Wukong: (Curious) What treasure?

Zhu Bajie: (Teasing) If I told you, you wouldn't have seen it; if I gave it to you, you wouldn't recognize it.

Sun Wukong: (Defiant) You fool, you're making fun of me, the Great Sage. Five hundred years ago, when I sought immortality, I roamed the far ends of the earth. What haven't I seen?

Zhu Bajie: (Revealing) Brother, have you ever seen a Ginseng Fruit?

Sun Wukong: (Amazed) Now that, I've truly never seen. But I've often heard people say that the Ginseng Fruit is like an elixir of life, and eating it can greatly extend one's lifespan. Where can we find one now?

Zhu Bajie: (Informing) They have it here. The boy offered two to our master, but the old monk didn't recognize them, thinking they were newborn babies, and didn't dare eat them. That lazy boy, instead of sharing them with us, secretly ate one in the next room, chomping away, making my mouth water! How can we get a taste of it? I thought you, being quite crafty, could sneak into the garden and steal a few for us to try. What do you say?

Sun Wukong: (Confident) That's easy, leave it to me. I'll get them in no time!

Response

Zhu Bajie: (Laughing smugly) Hehe, Monkey Brother, I knew you'd have a way! That Ginseng Fruit is no ordinary thing, we should hurry and get a few to try!

Table 15: A complete example of using in-context learning to prompt the model to perform role-play. Inputs include *Persona and History*. *Response* is the model's role-play response.