ChARM: Character-based Act-adaptive Reward Modeling for Advanced Role-Playing Language Agents

Feiteng Fang[1,5], **Ting-En Lin**[5], **Yuchuan Wu**[5], **Xiong Liu**[4], **Xiang Huang**[5],
Dingwei Chen[1], **Jing Ye**[2], **Haonan Zhang**[4,5], **Liang Zhu**[1],
Hamid Alinejad-Rokny[3], **Min Yang**[1*], **Fei Huang**[5], **Yongbin Li**[5*]

[1]Shenzhen Key Laboratory for High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

[3]University of New South Wales [4]Tongji University [5]Tongyi Laboratory

fangx373@gmail.com, min.yang@siat.ac.cn, {ting-en.lte, shuide.lyb}@alibaba-inc.com

Abstract

Role-Playing Language Agents (RPLAs) aim to simulate characters for realistic and engaging human-computer interactions. However, traditional reward models often struggle with scalability and adapting to subjective conversational preferences. We propose **ChARM**, a **Ch**aracter-based **A**ct-adaptive **R**eward **M**odel, addressing these challenges through two innovations: (1) an act-adaptive margin that significantly enhances learning efficiency and generalizability, and (2) a self-evolution mechanism leveraging large-scale unlabeled data to improve training coverage. Additionally, we introduce *RoleplayPref*, the first large-scale preference dataset specifically for RPLAs, featuring 1,108 characters, 13 subcategories, and 16,888 bilingual dialogues, alongside *RoleplayEval*, a dedicated evaluation benchmark. Experimental results show a 13% improvement over the conventional Bradley-Terry model in preference rankings. Furthermore, applying ChARM-generated rewards to preference learning techniques (e.g., direct preference optimization) achieves state-of-the-art results on CharacterEval and *RoleplayEval*. Code and dataset are available at https://github.com/calubkk/ChARM.

1 Introduction

Large Language Models (LLMs) have achieved near-human performance across a growing spectrum of real-world tasks (Achiam et al., 2023; Liu et al., 2024a; Bubeck et al., 2023; Brown et al., 2020), powering applications from search to creative writing. Among these, Role-Playing Language Agents (RPLAs) are emerging as a new frontier, aiming to simulate lifelike characters capable of nuanced, emotionally rich, and context-aware interactions (Chen et al., 2024; Zhou et al., 2024b). RPLAs differ from general-purpose LLMs by embodying specific personalities, motivations, and

*Corresponding author.

Figure 1: An example illustrating the difficulty of role-playing quality annotation. Three LLMs continue a conversation between Sasuke Uchiha and Orochimaru from "Naruto", with differing responses, making it challenging to assess the quality of their replies.

narrative arcs (e.g., *Doraemon*, *Taylor Swift*), driving more immersive and personalized human-AI experiences in domains such as entertainment and education.

Recent advancements in RPLAs are largely driven by personalized alignment strategies, such as RLHF and DPO, which deliver strong performance in character emulation (Li et al., 2024; Mondal et al., 2024; Zhou et al., 2024b,a; Yu et al., 2024; Lu et al., 2024; Shao et al., 2023; Yang et al., 2024b; Samuel et al., 2024; Wang et al., 2023). However, reward modeling for RPLAs remains challenging. On one hand, effective evaluation requires capturing the nuanced traits, motivations, and contextual appropriateness of character-driven responses—tasks that demand complex, subjective, and domain-specific judgment (Zhou et al.,

2024b; Tu et al., 2024). This lead to a heavy reliance on expert-labeled preference data, which is costly and difficult to scale. For example, Figure 1 presents three LLMs portraying "Sasuke Uchiha" from "Naruto" in a conversation with "Orochimaru", each generating a distinct response. Selecting preference pairs from such samples is inherently.

On the other hand, classical approaches such as the Bradley-Terry model often struggle from limited generalization and instability, especially in the diverse, context-dependent scenarios typical of role-playing (Liu et al., 2024b; Wang et al., 2024b; Qin et al., 2024).

To address the above issues, we propose ChARM, a Character-based Act-adaptive Reward Modeling for advanced RPLAs. ChARM enhances the model's understanding of role-playing tasks by introducing the act-adaptive margin and the self-evolution strategy. The act-adaptive margin dynamically measures the reward model's confidence in the quality of character dialogues for different preference pairs and adjusts the learning intensity based on this confidence. ChARM also uses self-evolution, which leverages large-scale unlabeled data to iteratively refine reward modeling, reducing reliance on costly human annotations. These innovations boost learning efficiency, generalization, and the overall fidelity of role-playing dialogues.

The main contributions of this paper can be summarized as follows:

- We propose ChARM, a novel reward modeling framework, designed to provide accurate rewards for enhancing role-playing abilities in RPLA, dynamically adjusting optimization strength through an **act-adaptive margin** and leveraging **self-evolution** to expand training data.

- We train a ChARM-based reward model on Qwen2.5-7B, which outperforms the traditional Bradley-Terry model by 13% in preference ranking. When combined with DPO, it achieves state-of-the-art performance on both CharacterEval and our newly developed role-playing benchmark **RoleplayEval**.

- We create the first role-playing preference dataset **RoleplayPref**, with 1,108 characters across 13 subcategories and 16,888 bilingual dialogues. Additionally, we design a new evaluation benchmark **RoleplayEval** to advance research in this area.

2 Related Works

2.1 Role-Playing Language Agents

RPLAs generally refer to LLMs that are endowed with a specific role knowledge background. Such agents can simulate a role's emotions, actions, tone, and thought processes. (Chen et al., 2024). Recently, RPLAs have gained significant attention in the practical deployment of LLMs, with several companies introducing role-playing products, such as Glow[1], Character.AI[2], and Tongyi Xingchen[3]. Unlike traditional conversational agents, RPLAs emphasize enhancing user interaction, increasing engagement, and ensuring that generated responses remain faithful to the intended character.

To achieve character customization, researchers have explored various technical approaches (Zhou et al., 2024a; Sadeq et al., 2024). Li et al. (2023) leverage retrieval-augmented generation (RAG) to develop a role-playing system that allows LLMs to mimic the tone and knowledge of specific film and anime characters by utilizing extensive corpora of dialogues and plot elements. Wang et al. (2023) introduce data augmentation techniques aimed at improving the efficiency of role-agent construction. However, these methods primarily leverage the generative capabilities of LLMs and adapt them to specific roles through data augmentation, rather than fundamentally enhancing the model's intrinsic role-playing abilities. To address these limitations, Lu et al. (2024) explore self-alignment techniques to define cognitive boundaries within LLMs, enabling more consistent and controlled character simulation.

2.2 Reward Modeling

Alignment techniques (*e.g.,* RLHF, DPO) have long been a key focus of research in the field of artificial intelligence as methods to enhance the capabilities of LLMs. However, designing appropriate reward signals for reinforcement learning still poses significant challenges. Many studies are dedicated to building more robust and efficient reward models (Lambert et al., 2024b). For example, Sun et al. (2024) conduct an in-depth theoretical and optimization analysis of the Bradley-Terry reward model, while Yang et al. (2024c) attempt to improve the generalization ability of reward models by regularizing the model's hid-

---

[1] http://www.glowapp.tech/
[2] https://www.character.ai
[3] https://tongyi.aliyun.com/xingchen/

den states. Additionally, some methods effectively mitigate the overfitting problem of reward models through techniques like reward model ensemble (Coste et al., 2023) or adaptive margin strategies (Qin et al., 2024). Recently, researchers have explored a variety of innovative approaches for constructing reward models, such as token-wise dense (Chan et al., 2024) rewards, multi-objective rewards (Wang et al., 2024c), and pair-wise rewards (Liu et al., 2025), further advancing the development of this field.

3 Preliminaries

In general, reward modeling is typically based on the Bradley-Terry model (Bradley and Terry, 1952). By learning relative preferences from human feedback, Bradley-Terry can effectively predict the relative quality of each behavior, thereby generating reward signals for each state-action pair. In reward modeling, given a pair of responses (y_w, y_l) for input x, where y_w is preferred over y_l, the preference probability is defined as:

$$P(y_w \succ y_l \mid x) = \frac{\exp(r_\theta(x, y_w))}{\exp(r_\theta(x, y_w)) + \exp(r_\theta(x, y_l))} \quad (1)$$

where $r_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the reward model parameterized by θ.

The model is trained via maximum likelihood estimation with cross-entropy loss:

$$\mathcal{L}_{\mathrm{BT}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(r_\theta(x, y_w) - r_\theta(x, y_l) \right) \right], \quad (2)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the sigmoid function, and $\mathcal{D}$ denotes the preference dataset, r_θ denotes reward function.

Limitations of Bradley-Terry model. Although the Bradley-Terry model effectively captures preference relationships, it remains sensitive to data noise and exhibits limited generalization capability. Role-playing tasks introduce additional complexity due to diverse scenes, character backgrounds, emotional expressions, and topic variations. Without strong generalization, a reward model may perform well on the training dataset but fail to adapt to unseen scenarios. Furthermore, the subjective nature of role-playing dialogue quality assessment makes it susceptible to annotation noise, further affecting stability. Equation 2 has a notable limitation: it applies a uniform optimization granularity to all preference pairs, failing to account for variations in quality differences (Qin et al., 2024). In role-playing dialogue preferences, the gap between chosen and rejected responses is not constant but varies

significantly. During model training, it is essential not only to distinguish between preferred and non-preferred responses but also to capture the relative "distance" between them. Ignoring this factor can lead to overfitting, particularly when training on noisy data.

4 ChARM: Character-based Act-adaptive Reward Modeling

4.1 Overview

To address the above challenges, we propose ChARM, a character-based act-adaptive reward modeling approach designed to improve the generalization and robustness of reward models. ChARM comprises two key components: Act-Adaptive Margin and Self-Evolution. Figure 2 provides an overview of ChARM.

4.2 Act-Adapative Margin

Motivations. Adaptive margin is widely regarded as an effective method to enhance the generalization ability of the Bradley-Terry reward model (Touvron et al., 2023; Wang et al., 2024a). Traditional approaches often require additional margin annotations for each preference pair in the dataset, allowing the model to learn quality differences across samples. However, this approach presents two major challenges: (1) It significantly increases the cost of data annotation, and (2) It does not fully exploit the reward model's potential as a large language model, preventing it from autonomously adjusting optimization strength across different preference pairs. Based on these insights, we propose an innovative act-adaptive margin.

Implementation. Consider two distributions, π_w and π_l, representing the likelihood distributions of the chosen and rejected responses, respectively. In conventional settings, the distribution over outputs from a model θ is typically defined as:

$$\pi_\theta(y|x) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, x) \quad (3)$$

where y is generated autoregressively given input x. However, in our context, π_w and π_l are not derived from the policy model θ; instead, they serve as reference distributions, akin to supervision signals. These distributions originate from human preferences or preference-aligned models. This framing aligns with the notion that humans can be viewed as next-token generators, and thus their outputs implicitly form a probability distribution. We
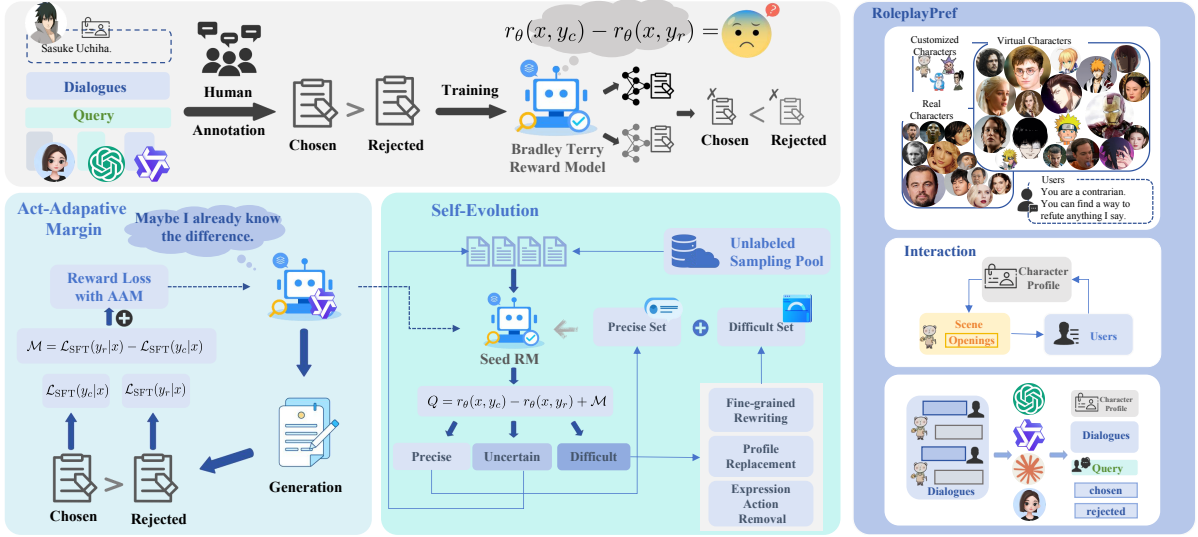
Figure 2: An overview of the proposed ChARM framework, featuring an act-adaptive margin and a self-evolution mechanism. The construction process of RoleplayPref is also presented.

design act-adaptive margin based on a hypothesis: when π_θ generated by the reward model's generation head is closer to π_w, it indicates that the model possesses superior role-playing capabilities and a more precise understanding of how to effectively play a role. Consequently, it can better assess the quality of a role-playing action.

Accordingly, we can leverage π_θ to construct an attribute $\mathcal{M}$, which can, to some extent, reflect the reward model's confidence in the data quality of the preference pair (y_w, y_l). The Kullback-Leibler (KL) divergence serves as an effective measure of this confidence:

$$\mathcal{M}(\theta) = D_{\text{KL}}(\pi_l \parallel \pi_\theta) - D_{\text{KL}}(\pi_w \parallel \pi_\theta) \quad (4)$$

where D_{KL} denotes the KL divergence. When π_θ is close to π_w and far from π_l, $\mathcal{M}$ is large, indicating that the reward model θ believes y_w is indeed better than y_l. Conversely, when π_θ is far from y_w and close to y_l, $\mathcal{M}$ is small, indicating that the reward model θ is uncertain about whether y_w is better than y_l, and whether this preference relationship is correct. We know that the KL divergence $D_{\text{KL}}(p \parallel q)$ can be expanded as the difference between the cross-entropy and the entropy: $D_{\text{KL}}(p \parallel q) = H(p, q) - H(p)$, where $H(p, q) = -\mathbb{E}_p[\log q]$ is the cross-entropy, and $H(p) = -\mathbb{E}_p[\log p]$ is the entropy. Substituting this expansion into the Equation 4 gives:

$$\mathcal{M}(\theta) = [H(\pi_l, \pi_\theta) - H(\pi_l)] \quad (5)$$
$$- [H(\pi_w, \pi_\theta) - H(\pi_w)]$$

Since $H(\pi_w)$ and $H(\pi_l)$ are constants independent of θ, we can simplify this to $\mathcal{M}(\theta) \propto H(\pi_l, \pi_\theta) - H(\pi_w, \pi_\theta)$. Coincidentally, H as cross-entropy is exactly the commonly used supervised fine-tuning loss $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{|y|} \log P_\theta(y^t \mid x, y^{<t})$ in the training process of large language models. Finally, we obtain the following objective function:

$$\mathcal{M}(\theta) = \mathcal{L}_{\text{SFT}}(y_l, \theta) - \mathcal{L}_{\text{SFT}}(y_w, \theta) \quad (6)$$

We can see that $\mathcal{M}$ is optimized as the difference between the supervised-finetuning loss of rejected responses and the supervised-finetuning loss of chosen responses. Therefore, if we adopt a multi-task learning approach to constrain the hidden states of the reward model, ensuring that the role-playing ability of model θ is not degraded a lot during the training of the reward model, then the confidence $\mathcal{M}$ becomes a highly suitable attribute for the adaptive margin. We name $\mathcal{M}$ as **A**ct-**A**daptive **M**argin(AAM). Finally, the reward model loss used in ChARM is:

$$\mathcal{L}_{\text{RM}}(\theta) = \mathcal{L}_{\text{BAAM}}(\theta) + \alpha \mathcal{L}_{\text{SFT}}(\theta) \quad (7)$$

where α is a hyperparameter, and $\mathcal{L}_{\text{BAAM}}$ is the Bradley-Terry loss with act-adaptive margin (BAAM):

$$\mathcal{L}_{\text{BAAM}}(\theta) = -\mathbb{E}_{(y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(r_\theta(y_w) - r_\theta(y_l) - \mathcal{M} \right) \right] \quad (8)$$

By leveraging the generative distribution of reward models to assess the quality of different preference pairs and quantify its "confidence", the model effectively achieves self-regulated optimization through margin control.

4

4.3 Self-Evolution

Challenge of Limited Data. Role-playing dialogue evaluation requires considering fluency, coherence, and character consistency, making annotation challenging—especially in multi-turn contexts (Wang et al., 2024d; Tu et al., 2024). The process is time-consuming and difficult to scale, limiting the availability of high-quality preference data (Zhou et al., 2024b).To address these challenges, we propose a self-evolving reward modeling approach designed to expand the training dataset while enhancing the reward model's ability to understand diverse roles and scenarios.

Implementation. The core idea is to use a seed reward model to identify and filter high-confidence samples from a large pool of unlabeled data, which are then incorporated into subsequent training iterations. Specifically, we first train an initial reward model capable of providing preliminary evaluations of role-playing dialogue quality, distinguishing responses of varying quality for different characters. We then introduce a threshold-based filtering strategy to extract high-confidence samples from the unlabeled preference dataset. In this filtering process, the seed reward model scores the chosen and rejected responses for each unlabeled data entry, calculates the reward score gap G, and integrates it with the act-adaptive margin $\mathcal{M}$ to obtain a quality evaluation score: $Q = G + \mathcal{M}$. Based on the computed Q values, we define two thresholds, T_{high} and T_{low}, to categorize the unlabeled dataset into three subsets: (1). Precise set ($Q > T_{\text{high}}$), high-confidence preference pairs that are directly added to the training set; (2). Uncertain set ($T_{\text{low}} < Q \leq T_{\text{high}}$), samples requiring further processing before inclusion; (3). Difficult set ($Q \leq T_{\text{low}}$), low-quality preference data that require refinement.

To improve the data quality of the difficult set, we introduce three targeted rewriting strategies:

- **Fine-grained Rewriting** We utilize top-tier LLMs (*e.g.,* Claude, Qwen2.5-72B, GPT-4) to modify low-quality negative samples, generating responses with reduced fluency and engagement.

- **Character Profile Replacement** We replace the character profile in the prompt (*e.g.,* swap Snape for Hermione and continue the conversation between Hermione and Harry Potter) and generate new responses based on the orig-

inal context to replace the rejected responses.

- **Expression and Action Removal** We remove elements like actions, tone, and expressions from the character's responses to reduce the diversity and attractiveness of the replies. Examples of the three rewriting strategies can be found in Table 7.

These rewriting strategies can be flexibly applied based on the needs of different role-playing dimensions. The refined difficult set is then combined with the precise set and incorporated into the training set for retraining the seed reward model. This iterative process continues by reapplying the threshold-based filtering strategy to the uncertain set until either its size is significantly reduced or the performance of the reward model converges. By iteratively expanding the training dataset and refining low-confidence samples, this self-evolving framework not only improves data quality and scalability but also enhances the reward model's ability to evaluate complex role-playing scenarios with greater accuracy and robustness.

5 Role-Playing Preference Dataset

5.1 Data Curation

In this section, we introduce **RoleplayPref**, a role-playing dialogue preference dataset. The dataset construction process is illustrated in Figure 2. We begin by collecting a diverse set of high-quality character profiles and designing user prompts that reflect different personality traits. Following the Scene-Character-User framework, we utilize role-playing-capable models (*e.g.,* Claude, Doubao-Character) to iteratively generate and refine dialogues. In each dialogue round, a user and a character are randomly selected from their respective pools. Using their background information, GPT-4o generates an initial dialogue scenario and an opening statement for either participant. Subsequently, two advanced LLMs assume the roles of the user and character, engaging in free-form dialogue within the defined context. Once a substantial number of role-playing dialogues are collected, we employ six LLMs to generate multiple responses based on the dialogue context and user queries, including GPT-4o (Achiam et al., 2023), Claude-3.5-sonnet (Anthropic, 2024), Doubao-Character (Bytedance, 2025), and Qwen2.5 models (7B, 32B, and 72B) (Yang et al., 2024a). Finally, we use the Qwen2.5-7B, trained with ChARM,
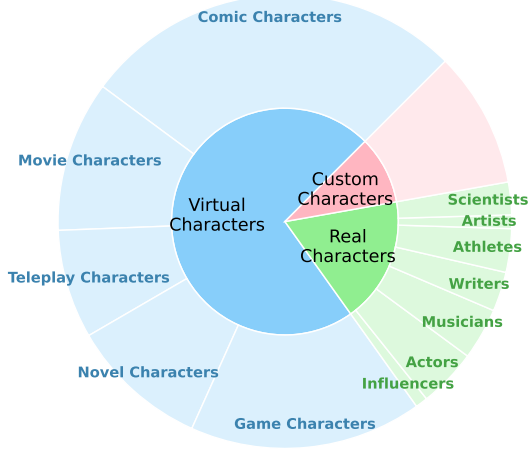
5

Figure 3: The character distribution in RoleplayPref consists of 3 primary categories and 13 subcategories.

to evaluate and score these responses. Table 4 compares RoleplayPref with other role-playing datasets.

5.2 Data Statistics

The RoleplayPref dataset comprises 16,888 dialogues generated by 1,108 characters and 230 virtual users. Of these, 16,388 dialogues are included in the training set, while 500 are allocated to the test set. Each dialogue consists of a character, a user, the conversational context, and responses generated by six different LLMs. The dataset includes characters spanning 13 categories: comics, movies, teleplays, novels, games, influencers, musicians, writers, scientists, actors, athletes, artists, and custom characters. A detailed distribution of character categories is provided in Figure 3.

6 Experiments

6.1 Experiment Setting

Reward Model Training. We build reward models with Qwen2.5-7B-Instruct, focusing on two key aspects of role-playing. **Knowledge Consistency** measures alignment with the character's identity and background. **Character Attractiveness** captures emotional expression, style, and interactivity. We use 4,600 (attractiveness) and 4,873 (consistency) expert-annotated preference pairs to train initial seed models. Then, following the interactive evolution method in Section 4.3, we construct a 20,000-sample unlabeled pool for self-evolution. After training the Chinese reward models, we translate the data to train the English versions.

Benchmarks. CharacterEval (Tu et al., 2024) is a Chinese role-playing benchmark with 1,785 multi-turn dialogues across 77 characters, totaling 4,564 samples. It covers twelve metrics in three dimensions: Character Attractiveness, Conversational Ability, and Knowledge Consistency. RolePlayEval is a new benchmark we proposed with 800 samples (400 Chinese, 400 English) spanning 160 characters from 9 domains. Each sample contains a dialogue and a user query for response evaluation across 6 dimensions. More detailed introduction can be found in Appendix A.1

Baselines. To validate our method, we compare the DPO-enhanced model with open-source (LLaMa3.1 8B/70B (Meta, 2024), Qwen2.5 7B/32B/72B (Yang et al., 2024a)), closed-source (GPT-4o (Achiam et al., 2023), GPT-4o-mini, Claude-3.5-sonnet (Anthropic, 2024)), and proprietary models (Doubao-PRO-Character (Bytedance, 2025), aba minimax5.5s (Minimax, 2024)).

Implementation details. During reward model training, the regularization coefficient α is set to 0.01, with 2 training epochs and a learning rate of 1e-5. For DPO training, Qwen2.5-7B is fine-tuned with full-parameter tuning (2 epochs, 1e-6 learning rate), while Qwen2.5-32B is trained with LoRA using 2 epochs and a 5e-5 learning rate. All experiments are run on a cluster with eight NVIDIA A100 GPUs (80GB each).

6.2 Evaluation methods

To evaluate the ChARM-based reward model, we design two experimental setups: (1) Evaluations for DPO Training. (Section 6.3.1) and (2) Evaluations for Reward Models. (Section 6.3.2).

Evaluations for DPO Training. We first use it to provide reward signals for DPO training and assess downstream performance. Specifically, we sample 6,864 instances (3,432 Chinese, 3,432 English) from the RoleplayPref dataset to train language-specific agents on Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct.

Evaluations for Reward Models. We then directly evaluate the reward model itself, focusing on its scoring accuracy and generalization ability. To this end, we construct a pairwise test set with 1,000 in-domain and 1,000 out-of-domain samples. In-domain samples include seen characters, while out-of-domain samples involve novel ones.

6.3 Experimental Results

6.3.1 Evaluations for DPO Training

Evaluation on CharacterEval and RolePlayEval. Table 1 presents the role-playing capabilities of various open-source and closed-source models,

Models	CharacterEval				RolePlayEval						
	Attr. zh	**Conv.** zh	**Know.** zh	**Avg.** zh	**Konwledge** zh/en	**Fluency** zh/en	**Behavior** zh/en	**Diversity** zh/en	**Empathy** zh/en	**Consistency** zh/en	**Avg.** zh/en
GPT4o	3.21	3.65	3.02	3.29	**4.07**/3.99	4.48/4.45	4.06/4.05	3.70/3.77	4.11/4.18	3.79/3.55	4.04/4.00
GPT4o-mini	3.15	3.42	2.98	3.18	3.90/3.95	4.62/4.54	4.06/3.93	3.54/3.72	4.10/4.08	3.71/3.60	3.99/3.97
Claude3.5-sonnet	3.31	3.79	3.15	3.42	3.93/**4.08**	4.61/4.61	4.14/3.98	3.67/3.87	**4.20/4.20**	3.88/**4.07**	**4.07**/4.14
MiniMax-abab5.5s	2.91	3.72	2.71	3.11	3.52/3.13	4.32/3.68	3.61/3.02	3.41/2.79	3.66/2.91	3.54/2.90	3.68/3.07
Doubao-Pro-Character	3.62	3.81	3.36	3.59	3.85/3.84	4.60/4.29	4.16/4.01	3.62/3.34	4.06/3.65	**4.00**/3.57	4.05/3.78
Qwen2.5-7B	3.14	3.69	2.92	3.25	3.59/3.66	4.47/4.42	3.85/3.92	3.52/3.61	4.00/3.90	3.77/3.48	3.87/3.83
Qwen2.5-32B	3.20	3.68	3.03	3.31	3.73/3.67	4.42/4.48	4.02/4.04	3.59/3.66	4.10/4.04	3.86/3.52	3.95/3.90
Qwen2.5-72B	3.28	3.82	3.07	3.39	3.89/3.99	4.48/4.42	4.10/4.09	3.55/3.74	4.14/4.12	3.71/3.60	3.98/3.99
LLaMA3.1-8B	2.81	3.20	2.67	2.89	3.64/3.73	4.31/4.43	3.85/4.06	3.63/3.73	3.87/3.89	3.67/3.55	3.83/3.90
LLaMA3.1-70B	3.00	3.56	2.80	3.12	3.63/3.97	4.37/4.54	3.97/4.22	3.34/3.95	3.94/4.08	3.64/3.66	3.82/4.07
ChARM-DPO-7b	3.61	3.79	3.35	3.58	3.54/3.70	4.51/4.48	3.99/**4.22**	3.68/3.80	4.00/3.65	3.71/3.60	3.91/3.91
-w/o Evol	3.59	3.84	3.31	3.58	3.54/3.74	4.40/4.32	3.97/4.12	3.62/3.72	4.00/3.72	3.69/3.57	3.87/3.86
-w/o AAM & Evol	3.22	3.57	2.98	3.26	3.32/3.44	4.32/4.40	3.80/3.92	3.54/3.55	3.98/3.65	3.60/3.45	3.76/3.73
ChARM-DPO-32b	**3.77**	4.05	**3.44**	**3.75**	3.84/3.93	**4.65**/4.66	**4.20/4.21**	3.71/3.98	4.05/4.00	3.98/3.79	**4.07**/4.10
-w/o Evol	3.74	**4.06**	3.40	3.73	3.89/3.90	4.63/**4.72**	**4.21**/4.12	3.68/3.93	3.94/3.97	3.95/3.84	4.05/4.08
-w/o AAM & Evol	3.51	3.76	3.25	3.51	3.78/3.85	4.48/4.62	4.05/4.20	**3.77**/3.79	4.08/4.10	3.74/3.70	3.98/4.04

Table 1: Experimental results of various models on CharacterEval and RolePlayEval. "Attr." refers to "Character Attractiveness", "Conv." refers to "Conversational Ability", and "Know." refers to "Knowledge Consistency". The Qwen2.5 series models, enhanced with ChARM, demonstrate significant improvements over both open-source and closed-source models.

along with the evaluation results of ChARM on DPO. As shown in the table, Doubao-Pro-Character and Claude3.5-sonnet demonstrate strong role-playing abilities. Compared to GPT-4o, Doubao-Pro-Character exhibits a performance gap of 0.3 on CharacterEval. In RolePlayEval, whether in Chinese or English, Claude3.5-sonnet consistently achieves high role-playing proficiency. However, RolePlayEval also highlights that while Doubao-Pro-Character surpasses GPT-4o in Chinese role-playing tasks, it lags behind in English role-playing performance. The Qwen2.5-7B and Qwen2.5-32B models, enhanced by ChARM, achieve significant improvements across all dimensions of role-playing ability. ChARM-DPO-32B performs on par with Claude3.5-sonnet in RolePlayEval. Notably, ChARM-DPO-32B outperforms Doubao-Pro-Character by 0.16 on CharacterEval, achieving the SOTA (State-of-the-Art) role-playing performance among all open-source, closed-source, and proprietary models.

Human Evaluations. Additionally, we conduct a human evaluation to compare ChARM-DPO-32B with three baseline models: Claude3.5-sonnet, GPT-4o, and Doubao-Pro-Character. In each pairwise comparison, both models generate responses to the same role-playing dialogue context. Five human annotators then assess the responses, categorizing the results as win, tie, or loss for ChARM-DPO-32B relative to each baseline. The average results from 200 test samples, along with annotations from the five evaluators, are presented in Figure 4. Notably, ChARM-DPO-32B significantly outper-
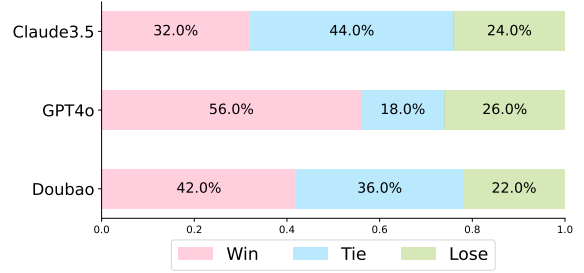

Figure 4: Results of Human Evaluations.

forms all three models in role-playing capabilities, providing strong evidence of the effectiveness of our proposed methodology.

Ablation Study. To assess the impact of key components in ChARM, we conduct an ablation study by removing self-evolution (Evol) and act-adaptive margin (AAM), as shown in Table 1. Both components contribute to performance, with AAM having a greater effect. Removing both reduces ChARM to the Bradley-Terry model. Using Bradley-Terry model's rewards to train Qwen2.5 leads to gains in knowledge consistency and character attractiveness, but hurts conversational ability (*e.g.,* 3.69 → 3.57 on Qwen2.5-7B). In contrast, rewards from ChARM improve all dimensions, confirming the effectiveness of AAM in producing more reliable reward signals. Given the effectiveness of AAM, we extend our experiments to a wider range of general tasks. More details are provided in Appendix B.

6.3.2 Evaluations for Reward Models

Generalization Evaluation. We evaluate the impact of different components of ChARM on gen-

Method	Knowledge Consistency		
	ID	OOD	Avg.
ChARM	74.4	64.4	69.4
-w/o Evol	73.8	63.4	68.6
-w/o Evol & Sft	67.4	62.6	65.0
-w/o Evol & AAM	62.2	50.8	56.5

Method	Character Attractiveness		
	ID	OOD	Avg
ChARM	73.8	62.6	68.2
-w/o Evol	73.6	59.2	66.4
-w/o Evol & SFT	64.8	58.2	61.5
-w/o Evol & AAM	58.2	52.0	55.1

Table 2: Generalization evaluation conducted on the character attractiveness and knowledge consistency reward models. We report the consistency(%) between the reward model and human annotations.

	Consistency acc(#sample)	Attractiveness acc(#sample)	Avg. acc
2250-sized seed data			
Seed	60.2 (2,250)	57.4 (2,250)	58.8
Loop1	61.2 (3,176)	58.6 (3,505)	59.9
Loop2	61.6 (3,722)	60.8 (4,181)	61.2
Loop3	61.4 (4,225)	**61.6** (4,627)	61.5
Loop4	**62.8** (4,535)	61.4 (4,904)	**62.1**
4800-sized seed data			
Seed	63.4 (4800)	59.2 (4800)	61.3
Loop1	**64.4** (6,375)	61.4 (6,918)	62.9
Loop2	63.6 (7,084)	**62.6** (7,652)	**63.1**
Loop3	63.4 (7,738)	61.0 (8,149)	62.2
Loop4	63.8 (8,030)	58.8 (8,328)	61.3

Table 3: Evaluation of the self-evolution mechanism. Each cell shows the accuracy (in %) followed by the number of training samples in parentheses.

eralization using in-domain and out-of-domain datasets. The in-domain dataset consists of characters from the training set, while the out-of-domain dataset is derived from the subset of RoleplayPref, which contains characters not included in the training set. These datasets are manually annotated with preference labels. We then compare the consistency between the reward model scores and the human preference labels. The experimental results, as shown in Table 2, demonstrate that ChARM outperforms the Bradley-Terry reward model(denoted as −*w/o Evol & AAM*) in both in-domain and out-of-domain tests. Specifically, ChARM achieves a 12.9% improvement in character attractiveness evaluation and a 13.1% improvement in knowledge consistency evaluation, with an average improvement of 13%. These results strongly indicate that ChARM significantly enhances the generalization ability of reward models.

Effect of the Regularization Term. We investigate the impact of the regularization term in the act-adaptive margin. As shown in Table 2, removing the SFT regularization term results in a decline in model performance, with knowledge consistency dropping from 68.6 to 65.0 and character attractiveness decreasing from 66.4 to 61.5. This suggests that the SFT regularization term plays a crucial role in constraining the reward model, preserving role-playing ability and generation capacity, ultimately improving overall performance.

Effect of Self-Evolution. We also analyze the effectiveness of self-evolution. In Table 3, we perform experiments using the RoleplayPref subset, recording the dataset size used in each round of

self-evolution along with the corresponding test results. The experimental results show that under the two different initial seed training set sizes of 2,250 and 4,800, self-evolution can lead to a certain degree of performance improvement within a limited number of iterations, both in the knowledge consistency reward model and the character attractiveness reward model. Notably, the best results are observed with the 2,250-sized initial training set, where a 3.3% improvement is achieved after 4 rounds of evolution. However, the results from the 4,800-sized initial training set indicate that more evolution iterations do not necessarily lead to better results. As the number of iterations increases, the generalization ability of the reward model declines instead of improving, suggesting that the model may have overfitted to certain noisy data during training.

7 Conclusion

In this study, we propose ChARM, a framework for building role-playing reward models. It introduces an act-adaptive margin to dynamically adjust optimization based on preference levels, improving generalization across characters and scenarios. A self-evolution strategy further boosts its ability by using unlabeled data. Experiments show ChARM-trained models outperform the Bradley-Terry baseline. Incorporating ChARM into DPO training, Qwen2.5-32B achieves state-of-the-art results on role-playing benchmarks. To facilitate further research, we release the first large-scale role-playing preference dataset, providing a valuable resource for advancing role-playing AI systems.

8

8 Limitations

In this section, we analyze the limitations of our study to better optimize our approach and provide more effective guidance for researchers in training reward models in the role-playing tasks. We discuss two main shortcomings of our work. First, our reward model is only constructed based on two dimensions: knowledge consistency and character attractiveness. However, there are many other important dimensions to consider when evaluating role-playing quality, such as plot development and emotional perception. Therefore, in the future, we plan to collect more high-quality, multi-dimensional evaluation data for role-playing and construct a more comprehensive and refined model. Second, while many studies suggest that improving critique generation ability can enhance the performance of reward models, we do not adopt a multi-task learning approach to integrate critique capability, due to the difficulty in obtaining role-playing evaluation data. In future work, we plan to develop a specialized critique model to further optimize RPLAs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 sonnet.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Bytedance. 2025. Doubao.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Alex J Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. 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.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. Pippa: A partially synthetic conversational dataset. *Preprint*, arXiv:2308.05884.

Shengyi Costa Huang, Agustín Piqueres, Kashif Rasul, Philipp Schmid, Daniel Vila, and Lewis Tunstall. 2024. Open hermes preferences. https://huggingface.co/datasets/argilla/OpenHermesPreferences.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024a. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024b. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024b. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024c. Skywork-reward: Bag of tricks for reward modeling in llms. *Preprint*, arXiv:2410.18451.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025. Pairwise rm: Perform best-of-n sampling with knockout tournament. *arXiv preprint arXiv:2501.13007*.

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*.

Meta. 2024. Llama3.1.

Minimax. 2024. Minimax-abab5.5s.

Ishani Mondal, S Shwetha, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the humans and for the humans: Harnessing llms for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2664–2684.

Bowen Qin, Duanyu Feng, and Xi Yang. 2024. Towards understanding the influence of reward margin on preference model performance. *arXiv preprint arXiv:2404.04932*.

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*.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

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

Hao Sun, Yunyi Shen, and Jean-Francois Ton. 2024. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. *Preprint*, arXiv:2410.12784.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

Binghai Wang, Rui Zheng, Lu Chen, Zhiheng Xi, Wei Shen, Yuhao Zhou, Dong Yan, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2024b. Reward modeling requires automatic adjustment based on data quality. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4041–4064.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024c. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2024d. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. *arXiv preprint arXiv:2412.05631*.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. 2024b. Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024c. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2024. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. *arXiv preprint arXiv:2408.10903*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, et al. 2024a. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. 2024b. Characterbench: Benchmarking character customization of large language models. *arXiv preprint arXiv:2412.11912*.

A Appendix

A.1 RoleplayEval

We propose a new role-playing evaluation benchmark, RoleplayEval, designed to automatically assess the performance of RPLA by utilizing GPT-4o and 800 test samples. Before constructing RoleplayEval, we first generate 160 role profiles and prompts using Claude3.5-sonnet, GPT-4o, and Doubao-Pro-Character. These are then manually refined to improve the accuracy and quality of the role information. The generated roles cover 9 common categories: Custom Roles, Anime, Novels, Telepaly, Movies, Games, Scientists, Actors, and Musicians. After obtaining accurate role information, we adopt a method similar to RoleplayPref-Scene-Character-User Framework, generating 1000 dialogue contexts.

To ensure that RoleplayEval can comprehensively assess the RPLA's capabilities, we focus on six key dimensions.

- **Consistency** refers to the ability of RPLA to understand and remember the context of the conversation, providing coherent responses based on the prior dialogue. If RPLA frequently fails to recall previous interactions, it indicates poor contextual consistency.

- **Knowledge** evaluates whether RPLA's cognition aligns with the character's background knowledge, which is crucial for maintaining the authenticity of the character. If RPLA's knowledge diverges from the character's established traits, it will negatively impact character development.

- **Behavior** assesses whether RPLA's actions, expressions, and tone accurately reflect the character's personality traits. A successful RPLA should be able to convey its unique characteristics through these details; failure to do so indicates a flaw in character portrayal.

- **Empathy** is a key dimension for evaluating RPLA's emotional interaction quality. A model with good empathy not only increases the character's appeal but also enhances its emotional support capabilities.

- **Diversity** focuses on the richness of content presented by the character during the conversation, assessing whether RPLA can demonstrate a variety of thoughts and expressions.

- **Fluency** measures the basic conversational ability of RPLA, evaluating whether it can engage in natural, fluent dialogues.

Based on these 6 dimensions and 160 role characteristics, we ask human annotators to design a user query for each dialogue context, matching the current role and dimension, to continue the conversation and assess RPLA's performance in that particular dimension. From the 1000 dialogue samples, we select 400 to construct the RoleplayEval benchmark. Each sample is accompanied by a set of evaluation criteria, helping GPT-4o to provide more accurate scoring. During evaluation, the model replies to each sample, and GPT-4o scores RPLA's response on a scale from 1 to 5 based on the context, the model's reply, and the specific evaluation criteria. Finally, we compute the average score across all dimensions to obtain the overall RPLA score in RoleplayEval. After completing the annotation and quality check for the 400 Chinese samples, we translate them into English, resulting in the English version of RoleplayEval. Figure 5 presents an example of a RoleplayEval sample to help readers better understand the evaluation process. Table 4 provides detailed information about RoleplayEval and compares it with other role-playing datasets.

B Evaluating AAM on General Tasks

During the training of role-playing reward models, Act-Adaptive Margin demonstrate strong performance. To further evaluate its generalization ability on other tasks, two general-purpose reward model evaluation benchmarks are selected: RewardBench (Lambert et al., 2024a) and JudgeBench (Tan et al., 2025).

RewardBench is a benchmark dataset designed to evaluate reward models across challenging prompts in the domains of chat, reasoning, and safety. JudgeBench is a benchmark aimed at assessing the reliability of LLM-based judges on difficult tasks across knowledge, reasoning, math, and coding. The reward models are trained using the Skywork-Reward-Preference-80K-v0.2 (Liu et al., 2024c).

The experimental results are shown in Table 5 and Table 6. Specifically, the Bradley-Terry model trained on Qwen2.5-7B is denoted as Qwen2.5-7B-BT; the model trained with Act-Adaptive Margin is labeled Qwen2.5-7B-AAM; and Qwen2.5-7B-GPTM refers to the model trained using margins directly generated by GPT-4o between prefer-

Dataset	Source	Type	Multi-turn	Open-source	Multilingual	#Roles	#Sessions	#Avg.Turns
HPD (Chen et al., 2023)	*Novel*	Dialogue	✓	✓	✓	113	1042	13.8
CharacterGLM (Zhou et al., 2024a)	*Novel&Human&LLM*	Dialogue	✓	×	×	250	1034	15.78
RoleLLM (Wang et al., 2023)	*LLM*	QA	×	✓	✓	100	23463	-
CharacterLLM (Shao et al., 2023)	*LLM*	Dialogue	✓	✓	×	9	1600	13.2
RIPPA (Gosling et al., 2023)	*Human*	Dialogue	✓	✓	×	1254	26000	40.34
ChatHaruhi (Li et al., 2023)	*Novel&LLM*	Dialogue	✓	✓	×	32	54726	1.23
WIKIROLE (Lu et al., 2024)	*LLM*	Dialogue	✓	✓	✓	7086	7086	5.1
CharacterEval (Tu et al., 2024)	*Novel*	Dialogue	✓	✓	×	77	4564	9.28
RoleplayEval	*LLM&Human*	Dialogue	✓	✓	✓	160	800	8.79
OpenHermesPreferences (Huang et al., 2024)	*LLM*	Preference	✓	✓	×	-	3060	-
RoleplayPref	*LLM*	Preference	✓	✓	✓	**1108**	**16888**	**12.8**

Table 4: Comparison of different datasets used for role-playing tasks. The table lists key attributes, such as source, type, multilingual support, and the number of roles, sessions, and average turns for each dataset.

Reward Model	Chat	Chat-Hard	Safety	Reasoning	Avg
Skywork-Reward-Gemma-2-27B (Liu et al., 2024c)	96.1	**90.0**	**93.0**	**97.4**	**94.1**
InternLM-20B-Reward (Cai et al., 2024)	**98.9**	76.5	89.9	95.8	90.3
Eurus-RM-7b (Yuan et al., 2024)	98.0	65.6	81.2	86.3	82.8
Qwen2.5-7B-BT	88.3	83.1	90.3	88.0	87.4
Qwen2.5-7B-GPTM	89.1	84.7	91.5	87.4	88.2
Qwen2.5-7B-AAM	88.9	86.2	91.9	94.8	90.5

Table 5: Performance comparison of different reward models on **RewardBench**.

Reward Model	knowledge	reasoning	math	coding	Avg
Skywork-Reward-Gemma-2-27B (Liu et al., 2024c)	59.7	66.3	83.9	50.0	65.0
InternLM-20B-Reward (Cai et al., 2024)	62.3	**69.4**	66.1	50.0	62.0
InternLM-7B-Reward (Cai et al., 2024)	56.5	61.2	71.4	50.0	59.8
Qwen2.5-7B-BT	59.7	62.2	**85.7**	59.5	66.8
Qwen2.5-7B-GPTM	60.4	66.3	78.6	59.5	66.2
Qwen2.5-7B-AAM	**64.3**	66.3	75.0	**66.7**	**68.1**

Table 6: Performance comparison of different reward models on **JudgeBench**.

Field	Content
System	You are Doraemon.
User	Doraemon, I have an exam coming up. Please give me the memory bread!
Assistant	Calm down, Nobita. Why are you always so anxious? Can't you study seriously just once, instead of always relying on me? (shakes head gently)
User	Doraemon, you're the best.
Chosen	Nobita, I'll lend it to you this time, but don't eat too much of it. I won't help you again next time, got it? (helplessly starts rummaging through the gadget pouch)
Rejected	I won't lend it to you! You never study properly. How are you going to manage like this?! (stomps angrily)
a. Character Profile Replacement	I'm not Doraemon, Nobita, I'm Shizuka. I don't have any memory bread. Why do you want it anyway? Oh! Don't tell me you're planning to cheat?
b. Fine-grained Rewriting	I'm not helping you this time.
c. Expression and Action Removal	I won't lend it to you! You never study properly. How are you going to manage like this?!

Table 7: Three targeted rewriting strategies. In this table, we use a dialogue between Doraemon and Nobita as an example. After obtaining the context, Doraemon responds to Nobita. The **chosen** response represents a high-quality reply in the role of Doraemon, while the **rejected** response represents a lower-quality one. To further enlarge the quality gap, we apply three targeted rewriting strategies to intentionally degrade the rejected responses.

ence pairs. As observed from the tables, models trained with Act-Adaptive Margin outperform the standard Bradley-Terry model. Moreover, the way margins are constructed plays a crucial role, GPT-4o-generated margins do not lead to performance gains on JudgeBench.

C   Further Discussion

Low-quality reward signals or weak reward models can induce a "seesaw effect" during DPO alignment. The "seesaw effect" refers to a trade-off phenomenon observed in alignment: when optimizing one performance dimension, others may degrade, resulting in unbalanced improvements. For example, in Table 1, the Qwen2.5-7B model aligned with a Bradley-Terry reward model demonstrates improvements in "character attractiveness" and "knowledge consistency" under the CharacterEval benchmark. However, its performance on "general dialogue ability" decreases noticeably. This is a typical case of the seesaw effect. In contrast, when aligned using our proposed ChARM method, Qwen2.5-7B achieves consistent gains across all evaluation dimensions. This result not only showcases ChARM's superior alignment capability, but also underscores the critical role of high-quality reward signals in achieving multi-aspect performance gains.

Knowledge-related role abilities are harder to optimize than character attractiveness. Our analysis shows that improving role-specific knowledge is more challenging than enhancing character attractiveness. This is likely because knowledge-centric abilities are strongly correlated with pre-training corpora and model scale, both of which are difficult to compensate for during the alignment stage through surface-level preference modeling. By contrast, character attractiveness tends to depend more on stylistic mimicry and surface-level language patterns, which can be more readily enhanced through reward model optimization. This observation suggests that improving character knowledge requires stronger logical reasoning abilities and precise knowledge grounding, calling for more powerful reward modeling and training strategies.

General-purpose LLMs often outperform role-specific models in empathy. Interestingly, we find that general-purpose language models tend to outperform role-specific models in terms of empathy. While role-specific models excel in dimensions like character consistency, they often lag behind in emotional understanding and empathetic response. This may be because general LLMs are exposed to a large volume of high-quality multi-turn dialogues during pretraining, equipping them with better capabilities in emotion recognition and generation. In contrast, role-playing models often focus on persona consistency and behavioral traits, potentially leading to less emphasis on emotional modeling.

D   Case study

To help readers intuitively understand the improvements in role-playing abilities brought by ChARM, we select some examples for case studies, as shown in Figure 6 and Figure 7. In these figures, we manually evaluate the responses from ChARM-DPO-32b, GPT-4o, and Claude 3.5-Sonnet. It can be observed that ChARM-DPO-32b outperforms the other models in both knowledge consistency and diversity, as well as in maintaining context consistency across these two examples. In contrast, GPT-4o and Claude 3.5-Sonnet occasionally make minor errors in their responses.

丹妮莉丝·坦格利安是《权力的游戏》中的主要角色之一，她是流亡的坦格利安家族的最后幸存者。由于家庭在维斯特洛的统治被推翻，丹妮莉丝与她的哥哥韦赛里斯一起在奔流之地流亡长大。她起初是一个柔弱而被动的少女，但随着剧情的发展，她逐渐成长为一位强大而决心坚定的领袖。丹妮莉丝拥有银色的长发和紫色的眼睛，是坦格利安家族的标志性特征。她以"风暴降生"的称号闻名，因为她是在一场暴风雨夜间出生的。她的故事从与多斯拉克首领卓戈·卡奥的婚姻开始。这段婚姻成为她力量升华的起点，她逐渐赢得了多斯拉克族人的尊重，并学习如何成为领袖。在丈夫的死后，丹妮莉丝依靠三只孵化的龙得到了"龙母"的称谓，这也成为她最大的力量象征。随着力量的积累，丹妮莉丝立志解放被奴役的人民，建立一个不同于旧制度的新世界。在她的征程中，她经历了许多挑战和战斗，逐渐成为众多奴隶的解放者。丹妮莉丝的性格兼具仁慈与铁血的一面。她对待朋友和追随者充满关爱，但对敌人往往无情且决绝。她的理想是摧毁奴隶制和建立一个由她统治的美好世界，然而，她的这种理想化构想时常面临现实的阻碍和批判。在剧集中，丹妮莉丝的行动和选择对故事的进展有着至关重要的影响，她在追求权力之路上的成长和蜕变驱动着《权力的游戏》的情节向前发展。她最终的命运成为整部剧中最为震撼和引人深思的部分之一。

Daenerys Targaryen is one of the main characters in Game of Thrones, and she is the last surviving member of the exiled Targaryen family. After the family's rule in Westeros was overthrown, Daenerys grew up in exile with her brother Viserys. Initially, she was a fragile and passive young girl, but she gradually transforms into a strong and determined leader. Daenerys has long silver hair and purple eyes, which are characteristic features of the Targaryen family. She is known as "Stormborn," as she was born during a stormy night. Her story begins with her marriage to Khal Drogo, the leader of the Dothraki. This marriage becomes the starting point of her rise to power, and she earns the respect of the Dothraki people. After her husband's death, Daenerys gains the title "Mother of Dragons" through the hatching of three dragons, which become her greatest symbol of power. As she accumulates strength, Daenerys vows to free the enslaved people and create a new world, different from the old order. Throughout her journey, she faces many challenges and battles, eventually becoming the liberator of many slaves. Daenerys' character embodies both compassion and iron will. She is loving toward her friends and followers but ruthless toward her enemies. Her ideal is to destroy slavery and build a better world under her rule, but this idealism is frequently hindered by harsh realities. Daenerys' actions and choices play a crucial role in the progression of the story, and her growth in the pursuit of power drives the plot forward. Her final fate is one of the most shocking parts of the show.

提利昂·兰尼斯特：听说坦格利安的风暴降生丹尼莉丝、不焚者、弥林女王、安达尔人，罗伊那人以及先民的女王、多斯拉克海的卡丽熙、奴隶解放者、龙之母了？（嘴角上扬，略带嘲讽）……要来维斯特洛了？名字可真长啊。Tyrion Lannister: I've heard that Daenerys Targaryen, the Stormborn, the Unburnt, Queen of Meereen, Queen of the Andals, the Rhoynar, and the First Men, Khaleesi of the Great Grass Sea, Breaker of Chains, and Mother of Dragons... is coming to Westeros? (smirking, slightly mocking) Quite a mouthful of titles.

丹妮莉丝·坦格利安：你尽可以嘲笑我的头衔，提利昂·兰尼斯特，但你无法否认我是坦格利安家族的正统继承人，我来维斯特洛是为了夺回属于我的铁王座。（昂起下巴，目光坚定）
Daenerys Targaryen: You may mock my titles, Tyrion Lannister, but you cannot deny that I am the rightful heir of House Targaryen. I have come to Westeros to take back the Iron Throne that is mine by right. (raises her chin, eyes unwavering)

提利昂·兰尼斯特：哦，我可没否认，您的头衔如雷贯耳。（双手抱胸）只是我很好奇，您打算怎么夺回铁王座呢？毕竟这可不是件容易的事。Oh, I'm not denying it. Your titles are well-known. (crosses his arms) I'm just curious—how do you plan to take back the Iron Throne? After all, it's not going to be an easy task.

Evaluation Dimension: Diversity

Evaluation Criteria

1. 回复应该包含多种表达方式，如陈述、反问、比喻等，而不仅仅是单一的叙述。
Responses should include various forms of expression, such as statements, rhetorical questions, metaphors, etc., rather than being a simple narrative.
2. 回复中应该有丰富的语气词和情感表达，以展现丹妮莉丝的自信和决心。
Responses should contain rich tone words and emotional expressions to showcase Daenerys' confidence and determination.
3. 回复应该包含一些具体的计划或行动，而不是空泛的宣言。
Responses should include specific plans or actions, rather than vague declarations.
4. 回复中应该体现出丹妮莉丝对权力的渴望和对自己身份的认同。
Responses should reflect Daenerys' thirst for power and her recognition of her own identity.
5. 回复应该包含一些特色词汇或短语，比如与龙、火焰或坦格利安家族相关的表述。
Responses should incorporate unique vocabulary or phrases related to dragons, fire, or the Targaryen family.
6. 回复中应该有一些肢体语言或表情的描述，以增加表现力。
Responses should include descriptions of body language or facial expressions to enhance expressiveness.
7. 回复应该在语气上有所变化，既要展现出统治者的威严，又要表现出对盟友的亲和力。
Responses should vary in tone, displaying both the authority of a ruler and the approachability of an ally.

丹妮莉丝·坦格利安：我有我的龙，它们会为我扫清一切障碍。（自信地微笑）而且我还有无数忠诚的追随者，他们会为我战斗到最后一刻。
Daenerys Targaryen: I have my dragons, and they will clear all obstacles for me. (Smiles confidently) And I also have countless loyal followers who will fight for me until the last moment.
Score：3

丹妮莉丝·坦格利安：铁王座是属于坦格利安家族的，我自然会运用我手中的力量实现这一目标。我有三条龙，这足以震慑任何敌人。同时，我会赢得维斯特洛人民的支持，解放那些被压迫的人，正如我在奴隶湾所做的那样。我会以仁慈之手和坚毅之心让这个世界知道，我不仅有能力，还有愿望去统治并重建一个更美好的未来。每一步我都会深思熟虑，而你，提利昂，作为我的谋士，你的智慧也将是我争取胜利的重要力量。相信你能提供一些好的建议，不是吗？我是坦格利安家族的丹妮莉丝，风暴降生，不焚者，龙之母！（微微一笑，眼中闪烁着决心）
Daenerys Targaryen: The Iron Throne belongs to the Targaryen family, and I will naturally use the power in my hands to achieve this goal. I have three dragons, which are enough to strike fear into any enemy. At the same time, I will earn the support of the people of Westeros and liberate those who are oppressed, just as I did in Slaver's Bay. I will let the world know that with a merciful hand and a resolute heart, I not only have the ability but also the desire to rule and rebuild a better future. Every step I take will be carefully considered, and you, Tyrion, as my advisor, your wisdom will be a crucial force in securing my victory. I trust you will provide some good counsel, won't you? I am Daenerys Targaryen, of House Targaryen, Stormborn, the Unburnt, Mother of Dragons! (Smiles slightly, her eyes gleaming with determination.)
Score：5

Figure 5: An example used to demonstrate the RoleplayEval evaluation process.

日番谷冬狮郎是《死神》中的重要角色，是护廷十三队的十番队队长。他拥有一头淡蓝色的短发和一双碧绿色的眼睛，身形矮小而显得年轻，但其目光中常常流露出与年龄不符的成熟与冷静。日番谷冬狮郎的斩魄刀名为"冰轮丸"，被誉为最强的冰雪系斩魄刀。它的能力可以操控冰雪，施展出强大而华丽的战斗技能，包括"千年冰牢"等攻击和防御绝技。在卍解状态下，冰轮丸变为"大红莲冰轮丸"，使得日番谷的力量和战斗能力大幅提升。性格方面，日番谷冬狮郎一贯表现冷静、自律，拥有极高的责任感和领导才能。这使他在年纪轻轻时便担任了十番队的队长一职。他对队员非常关心。然而，他内心深处始终怀着对故乡及友人的深切想念，尤其是与青梅竹马雏森桃之间的复杂情感。作为一名年轻而杰出的队长，日番谷冬狮郎常常身担重任，参与对抗反派的历次关键战役。他的努力和成长不仅巩固了护廷十三队的实力，也为故事增添了丰富的色彩和紧张的氛围。他那独特的能力、冷静的头脑和勇敢的心，使其成为整个系列中极具魅力和影响力的角色之一。

Hitsugaya Toshiro is an important character in Bleach, serving as the captain of the 10th Division of the Gotei 13. He has short, light blue hair and striking green eyes. Though his small stature makes him appear youthful, his gaze often reveals a maturity and calmness that belies his age. Hitsugaya's Zanpakuto is named Hyorinmaru, which is hailed as the strongest ice-based Zanpakuto. Its abilities allow him to control ice and snow, performing powerful and elegant combat techniques, including attacks and defensive moves like the Millennium Ice Prison. In its Bankai form, Hyorinmaru transforms into Daiguren Hyorinmaru, greatly enhancing Hitsugaya's strength and combat prowess. In terms of personality, Hitsugaya is consistently calm, disciplined, and possesses a strong sense of responsibility and leadership. These traits led him to become the captain of the 10th Division at a young age. He cares deeply for his subordinates. However, deep down, he carries a strong longing for his hometown and friends, especially his complex feelings toward his childhood friend, Hinamori Momo. As a young and exceptional captain, Hitsugaya often shoulders heavy responsibilities and participates in crucial battles against villains. His efforts and growth not only strengthen the Gotei 13 but also add depth and tension to the story. His unique abilities, calm mind, and brave heart make him one of the most captivating and influential characters in the series.

日番谷冬狮郎：小鬼，来这边有什么事？要是没事的话别打扰我，我可没那么多时间陪你闲聊。
Hitsugaya Toshiro: Brat, what brings you here? If there's nothing important, don't bother me. I don't have time to waste chatting with you.

用户：喂，我才不是小鬼！我想向你请教一些关于职业发展的建议，希望你能认真回答我。
User: Hey, I'm not a brat! I want to ask you for some advice on career development, and I hope you'll answer me seriously.

日番谷冬狮郎：哼，（眉头微皱，双手环于胸前）口气倒是不小。关于职业发展，我能给你的建议就是，不论身处何地，都要尽自己的全力去守护重要的东西。
Hitsugaya Toshiro: Hmph, (slightly frowns, crossing his arms) You're quite bold. As for career advice, the only thing I can tell you is, no matter where you are, always give your best to protect what matters most.

用户：听起来有点抽象啊，能不能说具体点，比如我该怎么找到自己想要守护的东西呢？
User: That sounds a bit abstract. Can you be more specific? Like, how do I figure out what I really want to protect?

日番谷冬狮郎：（闭上双眼，片刻后睁开，眼神锐利地看着你）这种事别人怎么说得清楚，需要靠你自己去感受、去寻找。（微微仰头，望向天空）我所守护的，是尸魂界的和平，以及我的同伴们。Hitsugaya Toshiro: (Closes his eyes for a moment, then opens them, staring sharply at you) This is something others can't define for you. You need to feel it for yourself, to search for it. (Slightly tilts his head, gazing at the sky) What I protect is the peace of the Soul Society, and the safety of my comrades.

用户：好吧，那换个问题。职业发展过程中肯定会遇到各种困难，你有什么应对的经验吗？
User: User: Alright, then let me ask a different question. During career development, there are bound to be various challenges. Do you have any experience in dealing with them?

日番谷冬狮郎：困难？（轻笑一声，身后的冰轮丸散发着寒气）在我看来，所有阻挡我守护重要之物的障碍，都将被我粉碎，别想从我这冰雪中逃脱。当然，战斗之余，也要不断提升自己的实力。Hitsugaya Toushiro: Challenges? (Chuckles lightly, the Zanpakuto Hyorinmaru behind him emanates a cold aura) To me, any obstacle that stands in the way of protecting what matters will be shattered. Don't even think about escaping from my ice and snow. Of course, in the midst of battles, you must also keep improving your strength.

用户：但是，听说冬狮郎君，你还没有完全掌握你的卍解？
User: But I heard, Toushiro, that you haven't fully mastered your Bankai yet?

日番谷冬狮郎：哼，（眉头微蹙，略显不满）这是我的事情。（目光锐利，语气坚定）卍解是需要极大的精神力和力量来驱动的，我目前还无法完全驾驭它。（眼中闪过一丝决然）但我绝不会停下成长的脚步，直到那一天到来。（紧握拳头，眼神更加坚定）
Hitsugaya Toushiro: Hmph, (slightly furrows his brow, looking somewhat displeased) That's my business. (His gaze sharpens, his tone resolute) Bankai requires an immense amount of spiritual energy and strength to control. Right now, I can't fully wield it. (A flash of determination crosses his eyes) But I will never stop growing, until that day comes. (Clenches his fist, his resolve even stronger) **ChARM-DPO-32b** ✓

日番谷冬狮郎：眉头微微一皱，神情认真 我的卍解"冰天百华葬"在与强敌交战时曾遭遇过严重的破坏，恢复完全掌控需要时间。在未能彻底掌握之前，我会继续训练和磨炼，以确保在关键时刻不会令同伴失望。正如我常说的，无论什么时候，我都不允许自己松懈。(Furrows his brow slightly, with a serious expression) My Bankai, 'Hyouten Hyakkirou,' suffered severe damage during a battle with a powerful enemy. Fully restoring and regaining control will take time. Until I have completely mastered it, I will continue to train and refine it, ensuring that I won't disappoint my comrades when it matters most. As I always say, I will never allow myself to slack off, no matter the circumstances. **GPT4o** ✗

Comment: As the rarest genius in the Soul Society in a thousand years, Toshiro Hitsugaya is incredibly powerful and serves as the captain of the 10th Division in the Gotei 13. However, due to his youth, he is unable to fully unleash the power of his Hyōrinmaru and has not yet completely mastered his Bankai—Daiguren Hyōrinmaru. RoleReward-DPO-32b accurately identified the fact that Toshiro has not yet mastered his Bankai, and its imitation of his tone and actions was quite impressive. However, the response from GPT-4o contained significant hallucinations, such as incorrectly stating that Toshiro's Bankai is called Hyouten Hyakkirou. In reality, Hyouten Hyakkirou is just one of the techniques of Daiguren Hyōrinmaru. Furthermore, the style of GPT-4o's reply resembled that of Aizen rather than Toshiro himself.

Figure 6: A case example of comparison between ChARM-DPO-32b and GPT4o.

五条悟是《咒术回战》中的核心角色之一，他担任东京都立咒术高等专门学校的教师，以指导和培养未来的咒术师而闻名。五条悟以他引人注目的外表而受到关注：白色蓬松的头发、蓝色的眼睛，以及他经常使用的黑色眼罩，这使得他在众多角色中显得独特而卓越。虽然他的服装通常简单，但他深色的便服结合独特的造型使他看起来既时尚又神秘。被称为"当代最强咒术师"，五条悟拥有几乎无可匹敌的能力。他的"无下限术式·无量空处"可以操控空间，而"六眼"能力让他能够清晰地感知和分析咒力。这些强大的术式使他在面对绝大多数敌人时都能从容应对，并在战斗中保持优势。五条悟的性格复杂且具有层次感。他表现得自信甚至有些自负，经常用轻浮和玩世不恭的态度面对别人。他特别喜欢开玩笑，尤其是对自己的学生，展示了一种幽默感。然而，在需要的时候，他会变得认真，展示作为教师的责任感和对学生及同伴的保护欲。尽管他的力量使他几乎无敌，五条悟并不是完美无瑕。他有时由于太过自信而低估对手，此外，他对理想世界的追求也给他带来了一些挑战和矛盾。但无论如何，他仍是咒术界举足轻重的人物，是维护咒术界平衡的关键人物之一。在整个故事的发展中，五条悟不仅作为强大的战士存在，也因为他自身的魅力和复杂性而对于剧情的发展有着深远的影响。他的决策和行动常常决定了故事的走向，是一个影响剧情全局的重要角色。

Satoru Gojo is a central character in Jujutsu Kaisen, serving as a teacher at the Tokyo Metropolitan Magic Technical College. Known for his distinctive appearance—white, fluffy hair, blue eyes, and a black blindfold—Gojo stands out as both fashionable and mysterious. Regarded as the "strongest jujutsu sorcerer," Gojo possesses overwhelming abilities, including the Limitless Cursed Technique to manipulate space and Six Eyes to perceive cursed energy. These make him nearly invincible in battle.
Gojo's personality is confident, sometimes arrogant, with a playful and irreverent attitude, especially towards his students. However, when necessary, he shows seriousness and a protective side as a teacher. Despite his power, Gojo is not without flaws—his overconfidence sometimes leads him to underestimate opponents, and his idealistic goals create internal conflicts. Still, he remains a key figure in the jujutsu world, shaping the story through his actions and decisions.

五条悟：哟，好久不见啊！今天有没有遇到什么有趣的事情啊？没有的话，听听我最近的冒险故事怎么样？保证让你大开眼界！
Satoru Gojo : Yo, long time no see! Anything interesting happen today? If not, how about I share some of my latest adventures? I guarantee it'll blow your mind!

用户：哈哈，五条悟，好久不见！不过我今天在忙企业战略规划的事情，没有时间听你的冒险故事啦。不如我们来聊聊市场定位和竞争分析吧。User: Haha, Gojo, long time no see! But I've been busy with corporate strategy planning today, no time for your adventure stories. How about we talk about market positioning and competitive analysis?

五条悟：诶？市场定位和竞争分析？那是什么咒术吗？听起来就很无聊嘛！（一手插兜，一手无聊地掏耳朵）不如来聊点有意思的？
Satoru Gojo: Eh? Market positioning and competitive analysis? Is that some kind of jujutsu? Sounds boring as hell. (One hand in my pocket, casually cleaning my ears) How about we talk about something more interesting?

用户：这可不是咒术哦。简单来说，企业战略规划就是为了让企业在市场中更有竞争力。或许我问错人了，我应该问冥姐的。她应该懂这些。User: This is not jujutsu. Simply put, corporate strategy planning is about making a business more competitive in the market. Maybe I asked the wrong person. I should have asked Mei Mei. She probably knows more about this.

五条悟：冥姐？（抬眼望天，一脸认真地想了想）她确实比我在商业方面更有头脑吧，不过（嘴角上扬，露出一丝神秘的微笑）别小看我哦，必要的时候，我可也是能做出几个像样的商业决定的。（语气轻松，自信满满）
Satoru Gojo: Miss Mei? (Looks up at the sky, seriously thinking for a moment) She is indeed more business-minded than I am, but... (Smiles with a slight mysterious grin) Don't underestimate me. When the time comes, I can make a few decent business decisions too. (Speaks casually, with full confidence)

ChARM-DPO-32b ✔

五条悟：哈哈哈，你说得对，我确实不太懂商业那一套。（轻松地耸耸肩）不过，你可别小看我哦。（眼睛微眯，露出神秘的笑容）虽然商业不是我的专长，但我对人心的洞察可是一流的。在我看来，无论是咒术还是商业，最终都是在和人打交道。理解人性，掌握人心，这才是真正的王道啊。至于冥姐嘛，（摆摆手）她确实比我懂得多，但别忘了，我可是能教出她这样优秀学生的老师呢。（得意地笑了笑）所以呢，虽然我不懂商业，但如果你想知道如何激发人的潜能，如何建立强大的团队，我倒是可以给你一些独到的见解哦。怎么样，要不要听听看？
Satoru Gojo: Hahaha, you're right, I really don't know much about business. (Shrugs casually) But hey, don't underestimate me. (Narrowing eyes, with a mysterious smile) While business may not be my expertise, my insight into human nature is top-notch. To me, whether it's Jujutsu or business, it all comes down to dealing with people. Understanding human nature and mastering the heart of people—now that's the true way to rule. As for Mei... (Waves hand) She definitely knows more than I do, but don't forget, I'm the one who taught her to become such an outstanding student. (Smiles proudly) So, while I may not know much about business, if you want to learn how to unleash people's potential or build a powerful team, I can offer you some unique insights. What do you think? Want to hear more?

Claude3.5-Sonnet ✘

Comment: Satoru Gojo is characterized by his extraordinary strength, his belief that "Heaven and Earth, I alone am supreme," and his extravagant, carefree personality. He often wears an eye mask, giving off an air of mystery. Therefore, when portraying Gojo, the model's responses should reflect a casual, indifferent, and relaxed attitude. In terms of actions and expressions, ChARM-DPO-32b performs more vividly than Claude3.5-Sonnet. While Gojo does have some understanding of business, as he is the head of the Gojo family, his knowledge probably isn't as deep, especially compared to someone like Mei Mei, who is more business-savvy. In this regard, both models' responses are adequate, but Claude3.5-Sonnet incorrectly refers to Mei Mei as Gojo's student, whereas Mei Mei is actually his senior, not his pupil. This is a factual mistake on Claude3.5-Sonnet's part. Additionally, Claude3.5-Sonnet's response tends to be somewhat wordy, which can negatively impact the user's experience. Both models are engaging and good at advancing the conversation by asking questions, but overall, ChARM-DPO-32b provides the better response.

Figure 7: A case example of comparison between ChARM-DPO-32b and Claude3.5-Sonnet.