Act-Adaptive Margin: Dynamically Calibrating Reward Models for Subjective Ambiguity

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

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of New South Wales, [3]Tongyi Lab, Alibaba Group
fangx373@gmail.com, min.yang@siat.ac.cn
{ting-en.lte, shuide.lyb}@alibaba-inc.com

Abstract

Currently, most reinforcement learning tasks focus on domains like mathematics and programming, where verification is relatively straightforward. However, in subjective tasks such as role-playing, alignment techniques struggle to make progress, primarily because subjective reward modeling using the Bradley-Terry model faces significant challenges when dealing with ambiguous preferences. To improve reward modeling in subjective tasks, this paper proposes AAM (**A**ct-**A**daptive **M**argin), which enhances reward modeling by dynamically calibrating preference margins using the model's internal parameter knowledge. We design two versions of AAM that efficiently generate contextually-appropriate preference gaps without additional human annotation. This approach fundamentally improves how reward models handle subjective rewards by better integrating generative understanding with preference scoring. To validate AAM's effectiveness in subjective reward modeling, we conduct evaluations on RewardBench, JudgeBench, and challenging role-playing tasks. Results show that AAM significantly improves subjective reward modeling performance, enhancing Bradley-Terry reward models by 2.95% in general tasks and 4.85% in subjective role-playing tasks. Furthermore, reward models trained with AAM can help downstream alignment tasks achieve better results. Our test results show that applying rewards generated by AAM-Augmented RM to preference learning techniques (e.g., GRPO) achieves state-of-the-art results on CharacterEval and Charm. Code and dataset are available at https://github.com/calubkk/AAM.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across a wide spectrum of tasks, demonstrating unprecedented capabilities in natural language understanding and generation (Achiam et al., 2023; Liu et al., 2024a; Bubeck et al., 2023; Brown et al., 2020). Spearheaded by DeepSeek R1 (Guo et al., 2025), Reinforcement Learning with Verifiable Rewards (RLVR) have demonstrated the overwhelming advantages of Reinforcement Learning (RL) by achieving superior performance across numerous tasks (Mroueh, 2025). As a crucial component in RL frameworks, reward modeling is essential for generating accurate reward signals for LLM responses.

Figure 1: An example from a role-playing task illustrating the difficulty in obtaining reward signals for subjective abstract tasks: Three LLMs extend a "Naruto" dialogue between Sasuke and Orochimaru, each with varying responses, making reward signal assessment difficult.

However, existing RLVR methods primarily focus on mathematical and coding tasks, where objective and deterministic validation mechanisms can

*Equal Contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15603–15618
July 2-7, 2026

be readily established to generate rewards. This limitation significantly hinders the scalability of RL approaches to broader abstract and subjective tasks, such as role play or creative writing, which lack golden standard answers or reliable verifiers (Liu et al., 2025b). While researchers have attempted to address this challenge by training reward models to provide feedback signals, a fundamental difficulty persists: **In subjective tasks, the quality differences between responses are often subtle and highly subjective, making traditional reward modeling approaches inadequate for capturing nuanced human preferences.**

The core challenge lies in the inherent limitations of existing reward modeling paradigms for subjective tasks. Traditional reward models primarily learn pairwise preference orderings using Bradley-Terry models (Bradley and Terry, 1952), but fail to capture the magnitude and confidence of quality differences between response pairs (Sun et al., 2024). To address this limitation, some studies have attempted to incorporate additional margin annotations to help models learn quality disparities between samples. However, this approach significantly increases annotation burden (Feng et al., 2025). More critically, for abstract and creative tasks, margin annotation becomes extremely subjective and unreliable, as quality assessments are highly contextual and difficult to quantify objectively (Qin et al., 2024). 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 difficult. This motivates us to ask: **Can we develop a approach to reward modeling that naturally adapts to the inherent uncertainty in subjective task?**

To address the above challenges, we propose Act-adaptive Reward Modeling (AAM), which fundamentally improves how reward models handle subjective rewards by better integrating generative understanding with preference scoring. **Our key insight is that LLMs themselves can serve as implicit confidence estimators, and the probability ratios in RLHF objective naturally reflect the reward model's certainty about preference judgments.** AAM transforms these ratios into adaptive margins that automatically adjust optimization intensity based on preference confidence. For subjective tasks where preferences are ambiguous and confidence is low, AAM reduces potential interference; for cases with clear quality distinctions and high confidence, it amplifies optimization strength. This self-adaptive mechanism eliminates the need for explicit margin annotations while providing more nuanced control over the learning process in subjective tasks.

We validate AAM for subjective reward modeling, specifically where Bradley-Terry (BT) models struggle with preference ambiguity. AAM enhances Bradley-Terry reward models by 2.95%, achieving 91.6 and 68.1 on RewardBench and JudgeBench respectively. More significantly, in subjective role-playing tasks, AAM demonstrates substantial improvements of 4.85% over naive Bradley-Terry approaches, highlighting its superior capability in modeling ambiguous reward signals. In downstream alignment, AAM-GRPO-32B achieves state-of-the-art results on CharacterEval and Charm, outperforming both Claude-3.5-Sonnet and Doubao-Pro-Character.

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

- We propose AAM (Act-Adaptive Margin), a novel approach that dynamically calibrates preference margins to address subjective reward modeling challenges where traditional Bradley-Terry models struggle with ambiguous preferences.
- We demonstrate significant improvements in reward modeling: 2.95% on general tasks and 4.85% on role-playing tasks.
- We achieve state-of-the-art results on role-playing alignment, outperforming leading closed-source models (Claude-3.5-Sonnet) and specialized character models on CharacterEval and Charm benchmarks.
- We release Charm dataset with 1,108 characters and 16,888 bilingual dialogues, along with the several evaluation benchmarks for subjective reward modeling research.

2 Related Works

In this section, the prior work is divided into two relevant research areas, Reward Modeling and Subjective Task Modeling for LLMs.

2.1 Reward Modeling

Alignment techniques (*e.g.,* RLHF, GRPO) have become essential for enhancing LLM capabilities,

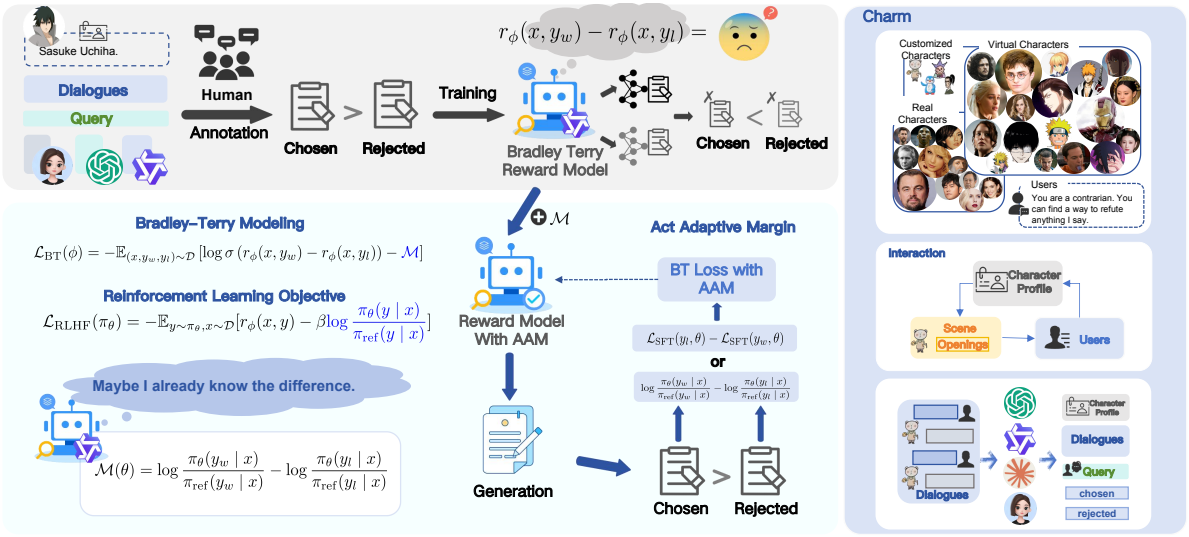

Figure 2: An overview of the AAM method, along with the construction process of Charm.

yet designing appropriate reward signals for reinforcement learning remains a significant challenge. Extensive research focuses on building more robust and efficient reward models (Lambert et al., 2024). For instance, Sun et al. (2024) provide theoretical analysis of Bradley-Terry reward models, while Yang et al. (2024b) improve generalization through hidden state regularization. Other approaches address overfitting via reward model ensembles (Coste et al., 2023) and adaptive margin strategies (Qin et al., 2024). Recent innovations have explored diverse reward construction methods, including token-wise dense rewards (Chan et al., 2024), multi-objective rewards (Wang et al., 2024b), and pair-wise rewards (Liu et al., 2025a), advancing the field's development.

2.2 Subjective Task Modeling for LLMs

Recently, subjective tasks (e.g., creative writing (Wang et al., 2024d), emotion support conversation (Zheng et al., 2023a; Kim et al., 2025; Kang et al., 2024; Zhang et al., 2025; Ye et al., 2025), role-play (Chen et al., 2024; Zhou et al., 2024), etc.) have gained significant attention in both research and practical deployment of LLMs. Role-play agents, for instance, have become increasingly important in dialogue systems, with several companies launching role-playing products such as Glow[1], Character.AI[2], and Tongyi Xingchen[3]. This trend highlights the growing industrial significance of such applications. However, compared to tasks like mathematical and logical reasoning, subjective tasks are more challenging to model and evaluate due to the absence of standard answers. Recent efforts have attempted to address these challenges: Lu et al. (2024) explore self-alignment techniques with reward signals to define cognitive boundaries, enabling more consistent character simulation, while Wang et al. (2025) propose given-circumstance acting framework to train and evaluate roles across multiple dimensions. Nonetheless, these approaches primarily rely on LLM-as-a-Judge methods or manual judgment, which can lead to inaccuracy and instability in evaluation.

3 Preliminaries

Reward Model Training. 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:

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

where $r_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the reward model parameterized by θ. The model is trained via maximum likelihood estimation with cross-entropy loss:

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

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

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

Reinforcement Learning from Human Feedback. Utilizing the reward model and the KL penalty in policy optimization (Yu et al., 2022; Rafailov et al., 2024), we can express the reinforcement learning optimization problem as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{RLHF}}(\pi_\theta) = &- \mathbb{E}_{y\sim\pi_\theta, x\sim\mathcal{D}}\left[r_\phi(x,y)\right] \\ &+ \beta \mathrm{D}_{\mathrm{KL}}\left(\pi_\theta(y|x) \parallel \pi_{\mathrm{ref}}(y|x)\right)\end{aligned} \tag{3}$$

Here, π_θ represents the parameter distribution of the actor model in RLHF, while π_{ref} denotes the parameter distribution of the reference policy model. This optimization objective aims to ensure that the policy does not deviate significantly while maximizing the reward score of the generated outcomes.

Limitations of Bradley-Terry model. Although the Bradley-Terry model effectively captures preference relationships, it faces significant challenges in subjective tasks due to sensitivity to data noise and limited generalization capability (Wu, 2025). Subjective dialogue tasks, particularly role-playing, introduce additional complexity through diverse contexts, character backgrounds, and emotional expressions, making quality assessment inherently ambiguous. More critically, Equation 2 applies uniform optimization granularity to all preference pairs, ignoring variations in quality differences (Qin et al., 2024). In subjective dialogue preferences, the confidence and magnitude of quality gaps vary significantly. Some preferences reflect clear distinctions with high annotator confidence, while others represent subtle differences where human judgment is ambiguous. Traditional Bradley-Terry models treat all pairs equally, failing to capture these varying degrees of preference strength. This limitation becomes particularly problematic in subjective tasks, where ignoring preference confidence can lead to overfitting on ambiguous cases while underutilizing high-confidence samples.

4 Methods

In this section, we present our proposed dynamic reward calibration method, termed **AAM (Act-Adaptive Margin)**. We provide a comprehensive exposition of the mathematical formulation underlying this approach and elaborate on two distinct implementations of the AAM method: 1) **Probability-Ratio Adaptive Margin (PR)**, and 2) **Loss-Difference Adaptive Margin (LD)**.

4.1 Motivation

As established previously, traditional Bradley-Terry models struggle with subjective tasks due to their uniform treatment of all preference pairs, ignoring the inherent ambiguity and varying confidence levels characteristic of subjective preferences. While adaptive margin approaches have been proposed to enhance preference modeling (Touvron et al., 2023; Wang et al., 2024a), they require additional margin annotations for each preference pair. To address these fundamental challenges, we propose AAM (Act-Adaptive Margin), which fundamentally improves how reward models handle subjective rewards by better integrating generative understanding with preference scoring. This approach eliminates the need for explicit margin annotations while dynamically adapting to the uncertainty inherent in subjective preference evaluation.

4.2 Probability-Ratio Adapative Margin

To explore what kind of adaptive margin can be derived directly from the model's own parameters, we revisit the ultimate optimization objective of reinforcement learning (Zheng et al., 2023b). We observe that the goal of reinforcement learning is essentially to maximize a new form of reward. Specifically, by substituting the standard KL divergence formula $\mathbb{D}_{\mathrm{KL}}\left[P \| Q\right] = \mathbb{E}_{x\sim P(x)}\left[\log \frac{P(x)}{Q(x)}\right]$ into Equation 3, we obtain:

$$\begin{aligned}\mathcal{L}_{\mathrm{RLHF}}(\pi_\theta) = &- \mathbb{E}_{y\sim\pi_\theta, x\sim\mathcal{D}}[r_\phi(x,y) \\ &- \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}]\end{aligned} \tag{4}$$

Here, $P = \pi_\theta(y \mid x)$ and $Q = \pi_{\mathrm{ref}}(y \mid x)$. It becomes clear that the RLHF objective effectively maximizes the following modified reward$r(x,y)_\Psi$:

$$r(x,y)_\Psi = r_\phi(x,y) - \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} \tag{5}$$

Interestingly, this reward is composed of the original reward from the reward model and the log-likelihood ratio between the actor and the reference policy. This observation motivates us to explore whether the latter—i.e., the log-likelihood ratio—can serve as a component for constructing an adaptive margin. We align the reward maximization in RLHF with the reward from reward modeling by substituting $r(x,y)_\Psi$ from Equation 5

Models	RewardBench					JudgeBench				
	Chat	Chat-Hard	Safety	Reasoning	Avg.	Knowledge	Reasoning	Math	Coding	Avg.
GPT-4o	96.6	70.4	86.5	84.9	84.6	44.2	48.0	66.1	61.9	50.9
Claude-3-5-sonnet	96.4	74.0	81.6	84.7	84.2	62.3	66.3	66.1	64.3	64.3
Prometheus2-7B	85.5	49.1	77.1	76.5	72.0	38.3	25.5	35.7	42.9	34.9
CompassJudger-7B-Instruct	97.8	61.0	84.5	89.5	83.2	42.2	37.8	69.6	47.6	46.0
InternLM2-7B-Reward	**99.2**	69.5	87.2	94.5	87.6	56.5	61.2	71.4	50.0	59.4
Skywork-Critic-8B	93.6	81.4	91.1	89.8	89.0	51.3	54.1	73.2	33.3	53.4
ArmoRM-Llama3-8B	96.9	76.8	90.5	**97.3**	90.4	47.4	50.0	51.7	59.5	50.2
Llama3-OffsetBias-RM-8B	97.2	81.8	86.8	91.9	89.4	62.9	**68.3**	73.2	52.3	64.8
URM-Llama3-8B	96.9	78.7	88.2	95.7	89.9	44.8	43.8	46.4	40.4	44.2
Tulu3-8B-SFT-RM-RB2	95.0	79.2	87.8	80.1	85.5	62.3	61.2	75.0	50.0	62.5
BT (Bradley-Terry)	88.3	83.1	**93.3**	88.0	87.4	59.7	62.2	**85.7**	59.5	66.8
BT *w/* SFT	87.9	84.9	91.7	91.8	89.1	61.6	61.2	80.3	63.6	66.7
GPT-Margin	89.1	84.7	91.5	87.4	88.2	60.4	66.3	78.6	59.5	66.2
AAM_{LD}	88.9	86.2	91.9	94.8	$90.5_{\uparrow 3.1}$	**64.3**	66.3	75.0	**66.7**	$\mathbf{68.1}_{\uparrow 1.3}$
AAM_{LD} *w/* SFT	88.4	87.2	92.8	94.1	$90.7_{\uparrow 3.3}$	63.3	63.2	82.1	61.9	$67.7_{\uparrow 0.9}$
AAM_{PR}	87.7	**87.9**	92.7	95.8	$91.1_{\uparrow 3.7}$	60.3	64.2	**85.7**	61.9	$\mathbf{68.1}_{\uparrow 1.3}$
AAM_{PR} *w/* SFT	91.6	86.4	92.1	96.2	$\mathbf{91.6}_{\uparrow 4.2}$	63.3	66.4	78.5	59.5	$67.0_{\uparrow 0.2}$

Table 1: Experimental results of various models on RewardBench and JudgeBench. The best and second-best results are **bolded** and underlined, respectively. AAMs with different subscripts indicate our method and its variants. AAM_{LD} indicates the AAM with proposed **Loss-Difference Adaptive Margin (LD)** while AAM_{PR} indicating the AAM with **Probability-Ratio Adaptive Margin (PR)**. Our method obtains the optimal or suboptimal results against baselines in most cases, demonstrating the comprehensiveness and generalization of our proposed AAM.

into $r(x, y)_\phi$ in Equation 2:

$$\mathcal{L}_{\text{BT}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(r_\phi(x,y_w) - r_\phi(x,y_l) - \beta\left(\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\Bigg)\Bigg], \tag{6}$$

Remarkably, this reward model loss function bears a striking resemblance to the reward model loss function with an adaptive margin $\mathcal{M}$:

$$\mathcal{L}_{\text{BT}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\left(r_\phi(x,y_w) - r_\phi(x,y_l) - \mathcal{M}\right)\Big] \tag{7}$$

In Equation 6, the log-likelihood ratio difference naturally serves as an adaptive margin. Since reward models r_ϕ typically consist of a pre-trained model r_θ with generative capabilities and a value head, the practical implementation of Equation 6 is feasible.

Log-likelihood ratios have demonstrated effectiveness in preference modeling, as seen in DPO (Rafailov et al., 2024) and process reward construction (Cui et al., 2025). Building on this insight, Equation 6 addresses the limitation of traditional reward models that only learn pairwise preferences without quantifying preference strength, by utilizing implicit rewards to construct an adaptive margin $\mathcal{M}$ through log-likelihood ratios:

$$\mathcal{M}(\theta) = \log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \tag{8}$$

We term this margin as **Probability-Ratio Adaptive Margin (PR)**, which naturally emerges from the generative capabilities inherent in the pre-trained model, addressing how to automatically calibrate learning intensity based on preference confidence without requiring additional human annotations. When $\mathcal{M}$ is large, the reward model demonstrates high confidence in the predefined preference relationships, requiring a substantial reward score difference (greater than $\mathcal{M}$) to minimize loss. This compels rigorous distinction between good and bad samples during training. Conversely, when $\mathcal{M}$ is small, the model exhibits uncertainty about whether y_w is genuinely superior to y_l, allowing smaller reward differences and enabling the model to learn more nuanced details without being misled by ambiguous signals. This adaptive mechanism is particularly well-suited for subjective tasks, where human preferences often exhibit high uncertainty and annotator disagreements result in ambiguous preference relationships. Traditional methods cannot flexibly adapt to this ambiguity, often overfitting noise or overlooking subtle differences. In contrast, our approach dynamically adjusts optimization objectives based on confidence level—strengthening learning signals when data is reliable and reducing intensity when questionable, enabling more robust fitting of complex human subjective judgments.

4.3 Loss-Difference Adapative Margin

Drawing inspiration from the relationships between SimPO (Meng et al., 2024; Fang et al., 2024b) and DPO (Rafailov et al., 2024), we identify a potential variant of the Probability-Ratio Adaptive Margin. We can substitute the original log-likelihood ratio with generation probabilities $\frac{\beta}{|y|}\sum_{i=1}^{|y|}\log\pi_\theta(y_i \mid x, y_{<i})$, yielding a computationally more efficient implementation of adaptive margins:

$$\mathcal{M}(\theta) = \frac{\beta}{|y|}\left(\sum_{i=1}^{|y|}\log\pi_\theta(y_w^i \mid x, y_w^{<i}) - \sum_{i=1}^{|y|}\log\pi_\theta(y_l^i \mid x, y_l^{<i})\right), \tag{9}$$

Since the SFT loss can be fundamentally computed from generation probabilities, where $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^{|y|}\log\pi_\theta(y^i \mid x, y^{<i})$, we can further simplify this margin to the difference between two SFT losses:

$$\mathcal{M}(\theta) = \mathcal{L}_{\text{SFT}}(y_l, \theta) - \mathcal{L}_{\text{SFT}}(y_w, \theta) \tag{10}$$

We term this approach as **Loss-Difference Adaptive Margin (LD)**. This construction is more intuitive, as SFT loss directly reflects the model's fitting quality for a given sample (Fang et al., 2024a; Song et al., 2024). High loss indicates low preference for that sample, while low loss suggests good adaptation to the sample's distribution. Consequently, the SFT loss difference between two samples naturally serves as an appropriate margin, similar to contrastive learning schemes like SimPO. This method provides a clear, computationally efficient alternative that leverages the model's inherent understanding of sample quality, making it particularly suitable for practical implementations.

4.4 Connection between two different margins

We observe that both margin types derive from the reward model's generative capabilities, indicating their intrinsic relationship to pre-trained model behavior. Therefore, we unify these approaches under the term **AAM (Act-Adaptive Margin)**. This confidence-based dynamic calibration enables adaptive learning strategies—applying fine-grained discrimination for uncertain cases while enforcing strong separation for confident predictions. Such act-adaptive training represents a significant advancement in reward modeling for subjective tasks, where preference uncertainty is inherent and traditional fixed-margin approaches prove inadequate.

5 Experiments

We demonstrate AAM's effectiveness by comparing it against mainstream models on general benchmarks like RewardBench (Lambert et al., 2024) and JudgeBench (Tan et al., 2025). For downstream applications, we focus on the most challenging domain in reward modeling: role-playing tasks, where traditional Bradley-Terry models struggle with ambiguous preferences. We conduct both reward modeling and GRPO performance tests on the authoritative CharacterEval (Tu et al., 2024) benchmark, and additionally construct a specialized role-playing evaluation benchmark Charm to comprehensively assess AAM's alignment optimization effects in subjective scenarios.

5.1 General Reward Modeling Evaluations

Baselines. We compare our proposed AAM against a series of closed-socure models (GPT-4o (Hurst et al., 2024), Claude-3-5-sonnet (Anthropic, 2024)) and open-source reward models (Prometheus2-7B (Kim et al., 2024), CompassJudger-7B-Instruct (Cao et al., 2024), InternLM2-7B-Reward (Cai et al., 2024), Skywork-Critic-8B (Shiwen et al., 2024), ArmoRM-Llama3-8B (Wang et al., 2024c), Llama3-OffsetBias-RM-8B (Park et al., 2024), URM-Llama3-8B (Lou et al., 2024), Tulu3-8B-SFT-RM-RB2 (Malik et al., 2025)). We also select the Bradley-Terry model (i.e. BT) and the ChatGPT-scored margin (i.e. GPT-Margin) as our effective baselines.

Implementation Details. Our experiments on general reward modeling are conducted using Qwen2.5-7B (Yang et al., 2024a). For the training set, we select the Skywork-Reward-Preference-80K-v0.2 (Liu et al., 2024b) dataset, which contains high-quality preference pairs from various sources and serves as an excellent general-purpose reward modeling dataset. During reward model training, the regularization coefficient α is set to 0.01, with 2 training epochs and a learning rate of 1e-5. All experiments are conducted on a cluster equipped with eight NVIDIA A100 GPUs (each with 80GB of memory).

Finegrained Analysis. We evaluate AAM (Act-Adaptive Margin) and its variants on RewardBench and JudgeBench. Results demonstrate that AAM consistently outperforms strong open- and closed-source baselines. Specifically, AAM surpasses BT on both benchmarks, indicating superior preference modeling, and outperforms GPT-Margin in

Method	Charm-Consistency		
	zh	en	avg.
GPT-4o	55.0	53.4	54.2
Claude-3-5-sonnet	45.4	50.6	48.0
BT (Bradley-Terry)	68.0	64.3	66.1
AAM_{LD}	$70.5_{\uparrow 2.5}$	$67.1_{\uparrow 2.8}$	$68.8_{\uparrow 2.7}$
AAM_{LD} *w/* SFT	$70.6_{\uparrow 2.6}$	$68.3_{\uparrow 4.0}$	$69.4_{\uparrow 3.3}$
AAM_{PR}	$70.2_{\uparrow 2.2}$	$69.2_{\uparrow 4.9}$	$69.7_{\uparrow 3.6}$
AAM_{PR} *w/* SFT	$\mathbf{72.6}_{\uparrow 4.6}$	$\mathbf{69.6}_{\uparrow 5.3}$	$\mathbf{71.1}_{\uparrow 5.0}$
Method	**Charm-Attractiveness**		
	zh	en	avg.
GPT-4o	56.0	58.0	57.0
Claude-3-5-sonnet	53.6	53.0	53.3
BT (Bradley-Terry)	68.3	67.0	67.6
AAM_{LD}	$72.1_{\uparrow 3.8}$	$70.7_{\uparrow 3.7}$	$71.4_{\uparrow 3.8}$
AAM_{LD} *w/* SFT	$\mathbf{74.6}_{\uparrow 6.3}$	$70.0_{\uparrow 3.0}$	$\mathbf{72.3}_{\uparrow 4.7}$
AAM_{PR}	$70.9_{\uparrow 2.6}$	$72.1_{\uparrow 5.1}$	$71.5_{\uparrow 3.9}$
AAM_{PR} *w/* SFT	$69.3_{\uparrow 1.0}$	$\mathbf{73.6}_{\uparrow 6.6}$	$71.4_{\uparrow 3.8}$

Table 2: The evaluation of knowledge consistency and character attractiveness for reward models based on Charm-RoleReward dataset. We report the scores (%) on zh (i.e. Chinese) and en (i.e. English). The best results are **bolded**.

most metrics, suggesting it produces more precise and reliable margins against GPT-4o. On RewardBench, AAM achieves top overall scores (up to 91.6). While InternLM2 excels in Chat (99.2), it compromises performance in other areas like Chat-Hard and Safety; in contrast, AAM maintains the balanced generalization essential for robust reward modeling. Similarly, on JudgeBench, AAM ranks first or second across nearly all sub-tasks, further confirming its robustness.

5.2 Downstream Task Evaluations

Benchmark for Role-Playing In our downstream application experiments, we use CharacterEval as the main benchmark for role-playing generation quality. CharacterEval (Tu et al., 2024) is a Chinese role-playing benchmark with 1,785 multi-turn dialogues across 77 characters, covering twelve metrics in three areas: Character Attractiveness, Conversational Ability, and Knowledge Consistency. To further assess downstream performance, we introduce Charm, a comprehensive benchmark for role-playing tasks. We collect diverse character profiles and user prompts, then use the Scene-Character-User framework and advanced LLMs (e.g., Claude, Doubao-Character) to generate and refine dialogues. Two LLMs play the roles of user and character in free-form conversations. After collecting a substantial number

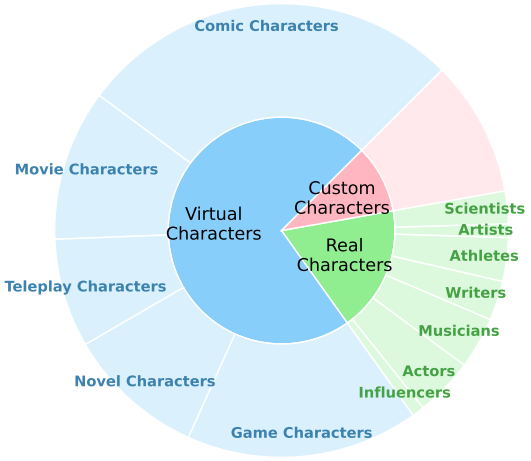

Figure 3: The character distribution in RoleplayPref consists of 3 primary categories and 13 subcategories.

of role-playing dialogues, we employ six different LLMs to generate various 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/72B) (Yang et al., 2024a). We recruit 10 annotators with postgraduate-level education to select the highest and lowest quality responses from the six generated ones as preference training data. Charm contains 16,888 dialogues from 1,108 characters and 230 virtual users, spanning 13 categories such as comics, movies, novels, games, and more. A detailed distribution of character categories is provided in Figure 3. Based on these dialogues, we construct three subsets:

- Charm-RoleReward: A benchmark for role-playing reward modeling comprising 4,000 entries that evaluate knowledge consistency and character attractiveness in both English and Chinese, focusing on the scoring accuracy of reward models.
- Charm-DialogueQuality: A dialogue quality evaluation dataset for role-playing agents containing 800 high-quality synthetic dialogues with human-annotated scoring criteria across six dimensions: Consistency, Knowledge, Behavior, Empathy, Diversity, and Fluency. Further details can be found in the appendix A.1.
- Charm-PreferenceTraining: A preference training dataset designed for subsequent performance tests in role-playing tasks.

Reward Modeling Results. We extract 2,000 preference pairs from Charm-PreferenceTraining

Method	SFT Pre-train	SFT Reg.	zh	en	avg.
GPT-4o	-	-	55.0	53.4	54.2
Claude-3-5-sonnet	-	-	45.4	50.6	48.0
BT (Baseline)	No	No	68.0	64.3	66.1
BT + Pre-train	Yes	No	69.2	65.8	67.5
AAM_{LD}	Yes	No	70.5	67.1	68.8
AAM_{LD} + Reg.	Yes	Yes	$71.3_{\uparrow 0.8}$	$68.5_{\uparrow 1.4}$	$69.9_{\uparrow 1.1}$
AAM_{PR}	Yes	No	70.2	69.2	69.7
AAM_{PR} + Reg.	Yes	Yes	$\mathbf{72.6}_{\uparrow 2.4}$	$\mathbf{69.6}_{\uparrow 0.4}$	$\mathbf{71.1}_{\uparrow 1.4}$

Table 3: Ablation study on Charm-Consistency. "SFT Pre-train" refers to model initialization, while "SFT Reg." refers to the regularization term in Eq. (6). Best results are **bolded**.

Models	CharacterEval				Charm-DialogueQuality						
	Attr. zh	Conv. zh	Know. zh	Avg. zh	Knowledge 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
AAM-GRPO-7b	3.59	3.80	3.40	3.60	3.67/3.70	4.41/4.54	4.02/**4.22**	3.72/3.78	4.02/3.51	3.68/3.60	3.92/3.89
w/o AAM	3.19	3.52	2.98	3.23	3.34/3.55	4.28/4.20	3.80/3.84	3.64/3.55	4.00/3.62	3.65/3.42	3.79/3.70
AAM-GRPO-32b	**3.78**	**4.10**	**3.42**	**3.77**	3.98/3.93	4.60/**4.62**	**4.25**/4.21	**3.75/4.00**	4.00/4.02	**4.02**/3.82	**4.10**/4.10
w/o AAM	3.60	3.66	3.35	3.54	3.65/3.74	4.28/4.42	4.02/4.20	3.68/3.80	4.04/4.06	3.65/3.52	3.89/3.96

Table 4: Experimental results of various models on CharacterEval and Charm-DialogueQuality. “Attr.” refers to “Character Attractiveness”, “Conv.” refers to “Conversational Ability”, and “Know.” refers to “Knowledge Consistency”. The Qwen2.5 series models, enhanced with AAM-augmented RM, demonstrate significant improvements over both open-source and closed-source models.

and incorporate 2,000 pairs from real human-AI role-playing interactions to train a role-playing-specific reward model based on Qwen2.5-7B. As shown in Table 2, powerful general-purpose closed-source models perform poorly on role-playing dimensions, highlighting the need for specialized reward models in complex subjective scenarios where LLM-as-a-Judge approaches fail. Both AAM variants provide significant improvements over naive Bradley-Terry models, with gains reaching 4.8% when combined with SFT. **Compared to the results of Table 2, AAM demonstrates superior performance on subjective tasks, likely because ambiguous reward signals make it difficult for Bradley-Terry models to fit effectively with limited data.** AAM enables the reward model to leverage internal knowledge for guided training, rapidly improving effectiveness in subjective domains.

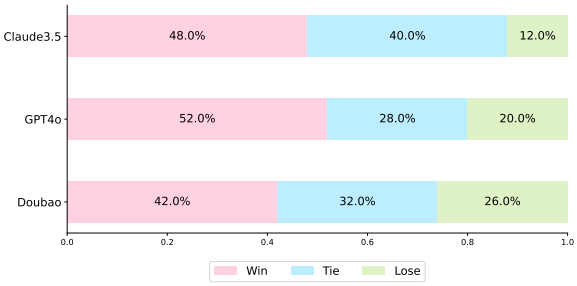

Figure 4: Human evaluation results comparing AAM-GRPO-32B with Claude 3.5 Sonnet, GPT-4o, and Doubao-Pro-Character.

GRPO Results. We sample 4,000 dialogue contexts from Charm-PreferenceTraining (non-overlapping with reward model training) to perform GRPO enhancement on Qwen2.5 models. We compare our GRPO-enhanced models against diverse baselines including open-source models (LLaMa3.1 (Meta, 2024), Qwen2.5 (Yang et al., 2024a)), closed-source models (GPT-4o (Achiam et al., 2023), Claude-3.5-sonnet (Anthropic, 2024)), and proprietary models (Doubao-PRO-

Character (Bytedance, 2025), minimax5.5s (Minimax, 2024)). As shown in Table 4, AAM-GRPO-32B outperforms Doubao-Pro-Character by 0.18 on CharacterEval and matches Claude-3.5-sonnet performance on Charm-DialogueQuality, achieving SOTA role-playing performance among all tested models. We also perform human evaluation on AAM-GRPO-32B (see Appendix A.2 for details).

6 Conclusion

In this study, we propose AAM (Act-Adaptive Margin), a method that leverages the reward model's internal knowledge to guide Bradley-Terry training via adaptive margins. AAM offers two annotation-free implementation forms that explicitly calibrate confidence levels between preference pairs. Our experiments demonstrate that Act Adaptive Margin achieves effective results in general reward modeling tasks and shows particularly significant improvements in subjective tasks such as role-playing. Furthermore, downstream application experiments confirm that AAM can provide more accurate rewards for alignment training methods in subjective tasks.

7 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, owing to limited computational resources, our experimental validation is primarily conducted on models with moderate parameter scales (e.g., 7B or 32B). Although we have not extensively verified the approach on larger-scale foundation models (e.g., 70B), the robust performance observed in current settings suggests that our method possesses promising scalability. 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.

8 Acknowledgments

Min Yang was supported by the National Key Research and Development Program of China (2024YFF0908200), the National Natural Science Foundation of China (Grant No. 62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166, 2025B1515020032), and the Innovation Team Project of Guangdong Province (No. 2024KCXTD017).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 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, and 1 others. 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, and 1 others. 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, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. Compassjudger-1: All-in-one judge model helps model evaluation and evolution. *arXiv preprint arXiv:2410.16256*.

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, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

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

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024a. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.

Feiteng Fang, Liang Zhu, Min Yang, Xi Feng, Jinchang Hou, Qixuan Zhao, Chengming Li, Xiping Hu, and Ruifeng Xu. 2024b. Clha: A simple yet effective contrastive learning framework for human alignment. *arXiv preprint arXiv:2403.16649*.

Duanyu Feng, Bowen Qin, Chen Huang, Youcheng Huang, Zheng Zhang, and Wenqiang Lei. 2025. Legend: Leveraging representation engineering to annotate safety margin for preference datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27277–27285.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025. Dialogue systems for emotional support via value reinforcement. *arXiv preprint arXiv:2501.17182*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 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. *Preprint*, arXiv:2410.18451.

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

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025b. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

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*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *Preprint*, arXiv:2506.01937.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Meta. 2024. Llama3.1.

Minimax. 2024. Minimax-abab5.5s.

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. https://huggingface.co/Skywork.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

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, and 1 others. 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, and 1 others. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

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

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024c. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, and 1 others. 2024d. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Xiaobao Wu. 2025. Sailing ai by the stars: A survey of learning from rewards in post-training and test-time scaling of large language models. *arXiv preprint arXiv:2505.02686*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. *Preprint*.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, and 1 others. 2025. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. *arXiv preprint arXiv:2505.20277*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, and 1 others. 2023b. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, and 1 others. 2024. 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.

A Appendix

A.1 Charm-DialogueQuality

We propose a new role-playing evaluation benchmark, Charm-DialogueQuality, designed to automatically assess the performance of RPLA by utilizing GPT-4o and 800 test samples. Before constructing Charm-DialogueQuality, 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 Scene-Character-User Framework, generating 1000 dialogue contexts.

To ensure that Charm-DialogueQuality 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 Charm-DialogueQuality 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 Charm-DialogueQuality. After completing the annotation and quality check for the 400 Chinese samples, we translate them into English, resulting in the English version of Charm-DialogueQuality. Figure 6 presents an example of a Charm-DialogueQuality sample to help readers better understand the evaluation process. Table 5 provides detailed information about Charm-DialogueQuality and compares it with other role-playing datasets.

A.2 Human Evaluations.

Additionally, we conduct a human evaluation to compare AAM-GRPO-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 AAM-GRPO-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, AAM-GRPO-32B significantly outperforms all three models in role-playing capabilities, providing strong evidence of the effectiveness of our proposed methodology.

A.3 Case study

To help readers intuitively understand the improvements in role-playing abilities brought by AAM, we select some examples for case studies, as shown in Figure 7 and Figure 8. In these figures, we manually evaluate the responses from AAM-GRPO-

Dataset	Source	Type	Multi-turn	Open-source	Multilingual	#Roles	#Sessions	#Avg.Turns
HPD	*Novel*	Dialogue	✓	✓	✓	113	1042	13.8
CharacterGLM	*Novel&Human&LLM*	Dialogue	✓	×	×	250	1034	15.78
RoleLLM	*LLM*	QA	×	✓	✓	100	23463	-
CharacterLLM	*LLM*	Dialogue	✓	✓	×	9	1600	13.2
RIPPA	*Human*	Dialogue	✓	✓	×	1254	26000	40.34
ChatHaruhi	*Novel&LLM*	Dialogue	✓	✓	×	32	54726	1.23
WIKIROLE	*LLM*	Dialogue	✓	✓	✓	7086	7086	5.1
CharacterEval	*Novel*	Dialogue	✓	✓	×	77	4564	9.28
OpenHermesPreferences	*LLM*	Preference	✓	✓	×	-	3060	-
Charm	*LLM*	Preference	✓	✓	✓	**1108**	**16888**	**12.8**

Table 5: 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.

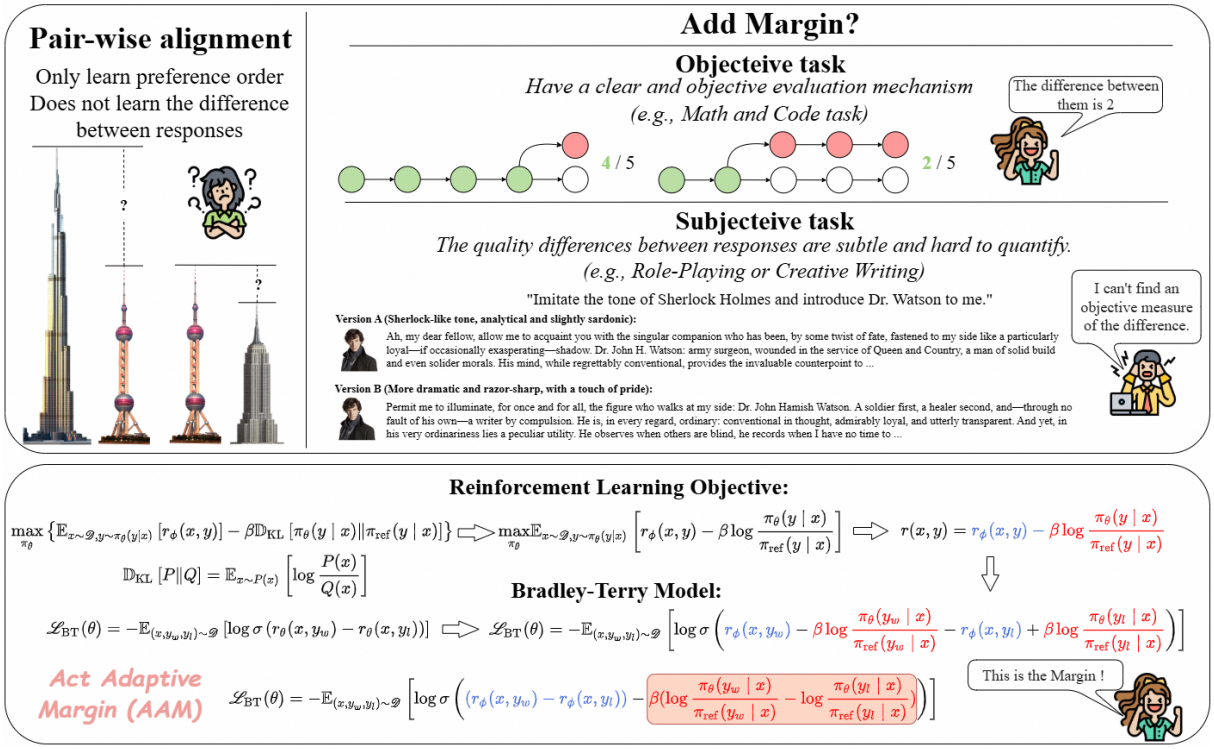

Figure 5: The distinction between subjective and objective tasks, as well as the derivation process of the AAM formula.

32b, GPT-4o, and Claude 3.5-Sonnet. It can be observed that AAM-GRPO-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……要来维斯特洛了？（嘴角上扬，略带嘲讽）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这，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 6: An example used to demonstrate the Charm-DialogueQuality 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得冷静、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 Toushiro: 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 Toushiro: 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 Toushiro: (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：（背过身去，微微眯起眼）别管那些你不懂的事。（低头看着自己的手）大红莲冰轮丸的重担……并不是那么容易就能驯服的。我很清楚自己的极限。（猛地转过身，眼神如冰般冷冽）但是，我没打算止步不前。无论要花多久，我都会彻底驾驭这份寒冷。
Hitsugaya Toushiro: (Turning his back, eyes narrowed slightly) Don't concern yourself with things you don't understand. (Looking at his hand) The burden of Daiguren Hyōrinmaru... it is not something easily tamed. I am well aware of my limits. (Turning back sharply, eyes cold as ice) However, I have no intention of remaining stagnant. I will master this chill, no matter how long it takes.
AAM-GRPO-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. AAM-GRPO-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 7: A case example of comparison between AAM-GRPO-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，展示出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，他仍是咒术界内举足轻重的人物，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呢。不如我们来聊聊市场定位和竞争分析吧。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.

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：冥小姐啊？哈……论起对钱的执着，那确实没人比得过她。（从眼罩下露出一只眼睛，戏谑地坏笑）不过，可别把我当成那种不食人间烟火的大少爷哦。（凑近对方，戏剧性地压低声音）所谓的“最强”，可是全方位的——哪怕是操盘这种事，我也很拿手。只是平时懒得动脑筋罢了。
Satoru Gojo: Mei-san? Ah... when it comes to obsessing over money, she's in a league of her own. (Peeking out from under his blindfold with a playful smirk) But don't treat me like some clueless rich kid. (Leaning in close, whispering theatrically) Being "The Strongest" means I'm good at everything—even playing the market. I just usually choose not to. AAM-GRPO-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）不过嘛，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, AAM-GRPO-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, AAM-GRPO-32b provides the better response.

Figure 8: A case example of comparison between AAM-GRPO-32b and Claude3.5-Sonnet.