Understanding Generalization in Role-Playing Models via Information Theory

Yongqi Li[1,2,*,†], **Hao Lang**[2,†], **Fei Huang**[2], **Tieyun Qian**[1,3,‡], **Yongbin Li**[2,‡]
[1] School of Computer Science, Wuhan University, [2] Tongyi Lab
[3] Zhongguancun Academy
{liyongqi,qty}@whu.edu.cn, {hao.lang,f.huang,shuide.lyb}@alibaba-inc.com

arXiv:2512.17270v2 [cs.LG] 13 Apr 2026

Abstract

Role-playing models (RPMs) are widely used in real-world applications but underperform when deployed in the wild. This degradation can be attributed to distribution shifts, including user, character, and dialogue compositional shifts. Existing methods like LLM-as-a-judge fall short in providing a fine-grained diagnosis of how these shifts affect RPM generalization, and thus there lack formal frameworks to characterize RPM generalization behaviors. To bridge these gaps, we introduce an information-theoretic metric, named reasoning-based effective mutual information difference (R-EMID), to measure RPM performance degradation in an interpretable way. We also derive an upper bound on R-EMID to predict the worst-case generalization performance of RPMs and theoretically reveal how various shifts contribute to the RPM performance degradation. Moreover, we propose a co-evolving reinforcement learning framework to adaptively model the connection among user, character, and dialogue context and thus enhance the estimation of dialogue response generation probability, which is critical for calculating R-EMID. Finally, we evaluate the generalization performance of various RPMs using R-EMID, finding that user shift poses the highest risk among all shifts and reinforcement learning is the most effective approach for enhancing RPM generalization. Code and data are available at https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/RPM-Generalization.

1 Introduction

As large language models (LLMs) (OpenAI, 2024; Bai et al., 2022) become more capable, the development of role-playing models (RPMs) (Chen et al., 2024c,b) has emerged as a key direction in advancing human-AI interaction. By simulating characters with distinct personalities, RPMs enable emotionally rich and contextually grounded dialogues, making them particularly valuable in fields such as entertainment (Park et al., 2023), personalized education (Pataranutaporn et al., 2021), and emotional companionship (Liu et al., 2021).

In real-world applications, the generalization abilities of RPMs are crucial for user experiences. For instance, global-facing platforms such as Character.AI serve users from diverse unseen linguistic and cultural backgrounds, and users may also request RPMs to simulate previously unseen characters. However, recent studies have reported failures in RPMs, such as generating culturally inappropriate responses (Li et al., 2024) or exhibiting character inconsistency (Wang et al., 2025a), raising concerns about their reliability.

These failure cases suggest that current RPMs underperform when deployed in the wild, which can be attributed to *distribution shifts* between training and test data. We categorize distribution shifts in RPMs into three types: user shift (e.g., unseen linguistic or cultural backgrounds), character shift (e.g., unseen character types), and dialogue compositional shift (e.g., longer and more complex multi-turn interactions). Understanding how these shifts contribute to RPM performance degradation is essential for building systems that generalize reliably in real-world applications.

Although existing methods, such as LLM-as-a-judge (Tu et al., 2024), can empirically quantify RPM performance degradation under distribution shifts, they fall short in providing a fine-grained diagnosis of how different shifts contribute to the degradation. More importantly, without a formal framework connecting performance degradation to distribution shifts, it is difficult to gain a systematic understanding of RPM generalization. This gap motivates our research question: ***Can we develop a***

[*] Work done while the author was interning at Tongyi Lab.
[†] Equal contributions.
[‡] Corresponding authors.

theoretical framework that formally characterizes how user, character, and dialogue compositional shift impact RPM generalization?

To address this, we propose an information-theoretic metric to quantify RPM performance degradation under distribution shifts, named reasoning-based effective mutual information difference (R-EMID). Concretely, we build R-EMID based on EMID (Oh et al., 2025), which is the first interpretable information-theoretic method for measuring the generalization performance of foundation models. To accurately estimate the conditional response generation probability required by EMID, we introduce an additional reasoning process to infer connections between user persona, agent character, dialogue history, and the target response. Moreover, we derive an upper bound for R-EMID, which enables the prediction of worst-case generalization performance of RPMs and theoretically reveals how various shifts contribute to the RPM performance drop.

To robustly generate the chain-of-thought for the reasoning phase required by R-EMID, we propose a co-evolving reinforcement learning (CoRL) framework to iteratively optimize a reasoning generator and a dialogue policy model. Specifically, the reasoning generator generates thoughts that help the current policy model select useful information from user, character, and dialogue history for conditional probability estimation, while the policy model produces dialogue responses and provides feedback for optimizing the reasoning generator and keeping it in-distribution. Extensive experimental results show strong correlations between R-EMI/R-EMID and LLM-as-a-judge metrics, as well as correlations between R-EMID and its upper bounds, which validate the proposed R-EMI/R-EMID metrics and CoRL.

Finally, we evaluate the generalization performance of various RPMs using R-EMID. Our analysis yields three findings: 1) Among the shifts considered, the user shift poses the highest risk to RPM reliability. 2) Reinforcement learning offers consistent generalization gains over standard supervised fine-tuning, whereas data augmentation delivers inconsistent results. 3) Naively incorporating reasoning traces fails to improve RPM generalization.

In this work, we make the following key contributions: 1) We identify three types of distribution shifts in RPMs, including user, character, and dialogue compositional shifts. 2) We introduce the first information-theoretic framework, named R-EMID, for characterizing RPM performance drop under various shifts. Additionally, we propose CoRL for estimating R-EMID and validate its effectiveness on real-world datasets. 3) We systematically evaluate the generalization performance of various RPMs using R-EMID, providing several valuable insights.

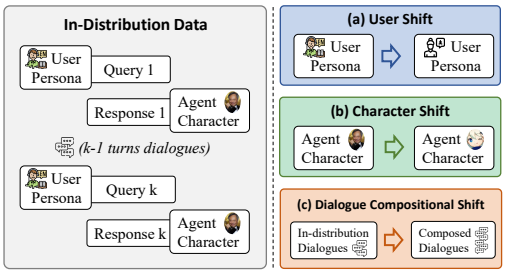

Figure 1: Distribution shifts in RPMs. (a) User shift: distribution change in user persona, e.g., EN→ZH; (b) Character shift: distribution change in agent character, e.g., unseen character types; (c) Dialogue compositional shift: distribution change in dialogues, e.g., from short-turn dialogue to composed long-turn dialogue.

2 Background

2.1 Problem Formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces of role-playing models (RPMs), respectively. We define variables $X = (X_u, X_a, X_d) \in \mathcal{X}$ and $Y \in \mathcal{Y}$, where X represents the input consisting of user persona X_u, agent character X_a, and dialogue context X_d, while Y denotes the response. We denote the joint distribution as P_{XY}, with marginals P_X, P_Y, and conditional distribution $P_{Y|X}$. In subsequent sections, P_{XY} refers to the training data distribution (in-distribution, ID) while Q_{XY} refers to the test distribution (out-of-distribution, OOD).

Let $P_\theta : \mathcal{X} \to \mathcal{Y}$ be an RPM parameterized by θ, trained on ID data. Given an input $X \in \mathcal{X}$, the model produces a response $Y_\theta \in \mathcal{Y}$. The marginal distribution of the model responses on ID and OOD data is given by $P_{Y_\theta} = \mathbb{E}_{x\sim P_X}\left[P_\theta(\cdot|x)\right]$ and $Q_{Y_\theta} = \mathbb{E}_{x\sim Q_X}\left[P_\theta(\cdot|x)\right]$, repectively.

2.2 Distribution Shifts in RPMs

Despite the expectation that RPMs should robustly handle unseen users, novel characters, and evolving dialogues in real-world deployment, recent reports of failures in these edge cases have raised concerns about their reliability (Roose, 2024). We argue that these failures can be attributed to the poor generalization when the test data deviates from the

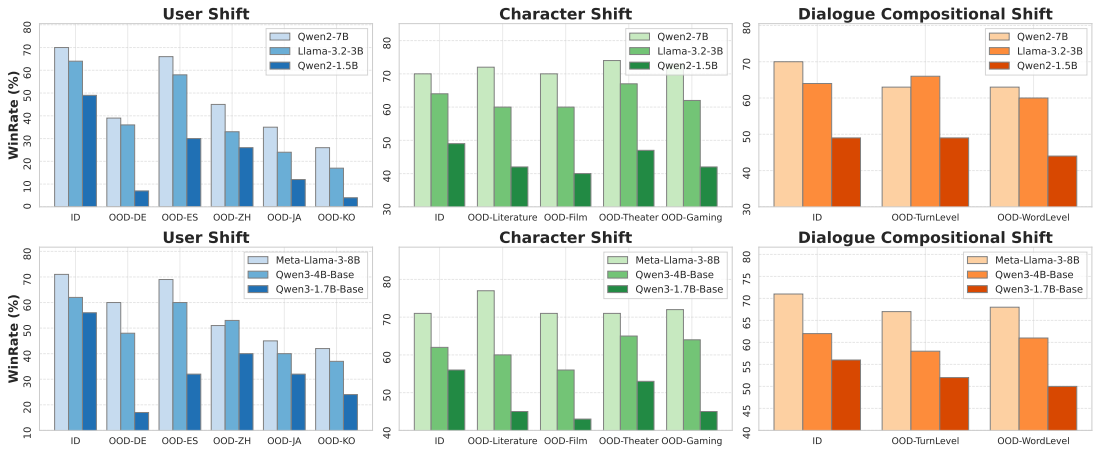

Figure 2: Pilot study of RPM performance under various distribution shifts, including user shift (left), character shift (middle), and dialogue compositional shift (right). The RPMs are trained with supervised fine-tuning based on six LLMs (extended results based on more LLMs are provided in Appendix B.1). Here, the x-axis is sorted by the severity of the distribution shifts, while the y-axis represents RPM performance measured by *WinRate*. We observe a consistent trend: as the severity of the distribution shift increases, performance degrades more significantly.

training data, i.e., under *distribution shifts*.

To understand these failure modes, as shown in Figure 1, we first identify three types of distribution shifts that RPMs may encounter in real-world deployments:

- ***User Shift***: the marginal distribution of user persona undergoes shift $D(P_{X_u}||Q_{X_u}) \gg 0$.
- ***Character Shift***: the marginal distribution of agent character exhibits substantial divergence $D(P_{X_a}||Q_{X_a}) \gg 0$.
- ***Dialogue Compositional Shift***: the marginal distribution of dialogue exhibits shift $D(P_{X_d}||Q_{X_d}) \gg 0$. By "compositional", we mean that test dialogues are constructed by recombining turns or words from ID dialogues.

Here, $D(\cdot||\cdot)$ denotes a divergence metric that quantifies the discrepancy between the training distribution P and the test distribution Q. In this paper, we adopt the Jensen-Shannon (JS) divergence (Menéndez et al., 1997) as our choice of $D(\cdot||\cdot)$. Based on these definitions, this work aims to *understand RPM behaviors, especially performance drop, under these distribution shifts*.

2.3 Pilot Study

A natural hypothesis is that as the degree of distribution discrepancy increases, the magnitude of RPM performance degradation under distribution shifts also increases. To test this hypothesis, we first conduct an empirical pilot study to evaluate the performance drop of RPMs under the three types of distribution shifts described above.

Since no existing datasets can systematically evaluate RPM performance under the above distribution shifts, we introduce **RPGBench**, a novel benchmark constructed according to the shift definitions in Sec. 2.2. The RPGBench comprises a total of 17k samples, including 5k ID samples, 5k OOD samples for user shift, 6k OOD samples for character shift, and 1k OOD samples for dialogue compositional shift. Each sample contains three input components: agent character, user persona, and dialogue context, and a reference response for training or evaluation [1].

Specifically, first, for user shift, we collect user personas described in English to construct the ID data. The OOD test data are constructed based on user personas with five non-English linguistic and cultural backgrounds, including *German, Spanish, Chinese, Japanese, and Korean*. Second, for character shift, we construct the ID data using real-world characters, while the OOD data are built from fictional characters from *literature, film, theater, and gaming* domains. Third, for dialogue compositional shift, OOD dialogues are crafted by recombining turns or words from ID dialogues. At the turn level, dialogues shift from 4 turns to 8 turns, while at the word level, test queries are rewritten using words in the training queries. Please refer to Appendix A for construction details.

[1]The dataset is available at https://huggingface.co/datasets/YongqiLi/RPGBench.

We select eleven LLMs from the Qwen (Yang et al., 2025) and LLaMA (AI@Meta, 2024) series, and fine-tune them on the ID training data of RPGBench via supervised fine-tuning. The RPM performance is measured by *WinRate*, which is an LLM-as-a-judge metric that is commonly adopted in previous evaluation methods for role-playing tasks (Wang et al., 2024b). We calculate the *WinRate* by comparing the model response with the reference response in the dataset. To reduce potential bias from individual LLM judges (Zheng et al., 2023), we adopt an ensemble strategy. Specifically, we average scores from multiple state-of-the-art LLMs, including Qwen3-235B-A22B, GPT-4o-mini, and Claude-3.7-Sonnet, as the final score.

Figure 2 shows the performance drop (y-axis) of RPMs built on different LLMs, evaluated across 11 shift subtypes, covering user, character, and dialogue shifts. The x-axis is sorted by the severity of shifts, measured by $D_{JS}(P_X||Q_X)$. From the figure, we observe a trend: as the severity of the distribution shift increases, the magnitude of performance degradation increases. *These empirical observations suggest the potential existence of a fundamental principle connecting performance drop with distributional shifts in RPMs, which motivates the following theoretical investigation into the underlying mechanisms.*

3 Understanding RPM Generalization via Information Theory

In this section, we will introduce: 1) the existing EMID (Sec. 3.1); 2) the proposed R-EMID to quantify the performance degradation of role-playing models (RPMs) (Sec . 3.2); 3) the proposed CoRL for estimating R-EMID (Sec . 3.3); and 4) experiments to validate R-EMID and CoRL (Sec . 3.4).

3.1 Preliminary

This section introduces effective mutual information difference (EMID) (Oh et al., 2025), the first information-theoretic metric for measuring the generalization performance of foundation models.

Definition 3.1 (EMID)**.** *The EMID of a model P_θ with parameters θ is defined as:*

$$\begin{aligned} \textit{EMID}(P_{XY}, Q_{XY}; P_\theta) &:= \textit{EMI}(P_{XY}; P_\theta) - \textit{EMI}(Q_{XY}; P_\theta) \\ &:= (I(P_{XY_\theta}) - I(P_{XY})) - (I(Q_{XY_\theta}) - I(Q_{XY})), \end{aligned}$$

where EMI measures the quality of model responses by quantifying how much they deviate from the ground-truth ones. The standard mutual information $I(P_{XY})$ reflects the degree of dependence between P_X and P_Y, as defined by Shannon (1948).

EMID can characterize the model performance drop under distribution shifts from P (training data) to Q (test data). Compared to black-box evaluation methods, such as LLM-as-a-judge (Zheng et al., 2023), EMID offers two key advantages: 1) It is theoretically grounded and enjoys strong interpretability. 2) The theoretical upper bound of EMID is derivable and can characterize the model's maximum generalization risk, which is important in safety-critical applications (Yu et al., 2024).

3.2 Reasoning-based EMID

While EMID offers a principled way to quantify model performance degradation under distribution shifts, its direct application to RPMs faces significant challenges. Specifically, estimating EMID relies on mutual information estimators such as CLUB (Cheng et al., 2020), which requires accurately estimating the response generation probability conditioned on the input, denoted as $p(y|x)$. However, the input of RPMs is inherently heterogeneous, comprising user persona, agent character, and dialogue context. Directly modeling such heterogeneous inputs is quite challenging and may lead to information loss, making the accurate estimation of $p(y|x)$ difficult.

To address this, we propose *reasoning-based EMID* (R-EMID), which introduces an intermediate reasoning process R to infer the connection between user, character, dialogue, and the targeted response. For instance, R may encode information about what stylistic and content features an ideal response should exhibit. This reasoning process enables the extraction of heterogeneous information from X, without the need to explicitly model the complex dependencies in X. By introducing the reasoning process R, R-EMID offers a more tractable way for measuring RPM performance degradation under distribution shifts.

Definition 3.2 (Reasoning-based EMI, R-EMI)**.** *Given a RPM P_θ parameterized by θ, its R-EMI under a joint distribution P_{XY} is defined as:*

$$\textit{R-EMI}(P_{XY}; P_\theta) := I(P_{X_R Y_\theta}) - I(P_{X_R Y}),$$

where the definition of mutual information $I(\cdot)$ is the same as in Definition 3.1. The newly introduced variable X_R is obtained via a mapping function $f_R(\cdot)$, which augments the input X into the $X_R =$

$(X, R = f_R(X))$. *Similar to EMI, R-EMI also aims to quantify the quality of responses produced by the model P_θ under certain distributions.*

Now, based on R-EMI, we can quantify the performance drop of RPMs with formal guarantees via reasoning-based EMID (R-EMID).

Definition 3.3 (Reasoning-based EMID, R-EMID)**.** *We define R-EMID as the difference between the R-EMI on the ID distribution and OOD distribution: R-EMID$(P_{XY}, Q_{XY}; P_\theta)$:= R-EMI$(P_{XY}; P_\theta)$ − R-EMI$(Q_{XY}; P_\theta)$. R-EMID can be used to characterize the performance degradation of RPMs under distribution shifts from P_{XY} to Q_{XY}.*

Furthermore, we derive an upper bound on R-EMID in Theorem 3.4, which characterizes the maximum performance drop of RPMs under distribution shifts, enabling the maximum risk analysis in safety-critical applications. Moreover, this bound principally reveals the underlying connection between the RPM performance drop and the severity of user, character, and dialogue shifts.

Theorem 3.4 (Upper Bound on R-EMID)**.** *Given an RPM P_θ trained on the distribution P_{XY} and tested on Q_{XY}, which share consistent conditional distributions $Y|X$. If there exist constants δ_P and δ_Q such that $D_{\mathrm{JS}}(P_{Y_\theta} \| P_Y) \leq \delta_P, D_{\mathrm{JS}}(Q_{Y_\theta} \| Q_Y) \leq \delta_Q, \Delta = \delta_P + \delta_Q$, then the R-EMID$(P_{XY}, Q_{XY}; P_\theta)$ is upper bounded by:*

$$\sqrt{2/3}\,\widehat{H} \sum_{z \in \{u,a,d\}} D_{\mathrm{JS}}^{1/2}(P_{X_z} \| Q_{X_z}) + 8\Delta^{1/4}, \quad (1)$$

where $\widehat{H} = \max_{x \in \mathcal{X}}[H(Q_{Y|x,r}) + H(P_{Y_\theta|x,r})]$, and $r = f_R(x)$ denotes the reasoning process [2].

> **Implication.** Theorem 3.4 implies that the RPM performance drop is affected by the combined contribution of marginal distribution shifts across user (u), agent character (a), and dialogue (d) components, as quantified by $\sum_{z \in \{u,a,d\}} D_{\mathrm{JS}}^{1/2}(P_{X_z} \| Q_{X_z})$. Additionally, it is influenced by the maximum model response uncertainty, i.e., $\hat{H}$, and the maximum deviation between model predictions and the reference responses, i.e., Δ.

3.3 R-EMID Estimation

Based on Definitions 3.2 and 3.3, the estimation of R-EMI and R-EMID relies on estimating mutual information terms. For example, to estimate mutual information $I(P_{X_R Y})$, we first need to collect N samples $\{(x_i, r_i, y_i)\}_{i=1}^N$, where each (x_i, y_i) can be directly sampled from P_{XY}, while the reasoning process r_i needs to be obtained via a mapping function f_R, i.e., $r_i = f_R(x_i)$. Then, the estimated $\hat{I}(P_{X_R Y})$ can then be calculated as:

$$\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i, r_i) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \log p(y_j|x_i, r_i). \quad (2)$$

In practice, both the reasoning mapping function f_R and the conditional distribution $p(y|x, r)$ are inaccessible. Therefore, as shown in Figure 3, we approximate them using two LLM-based parametric functions: a reasoning generator $q_{\phi_1}(\cdot)$ and a policy model $q_{\phi_2}(\cdot|\cdot)$.

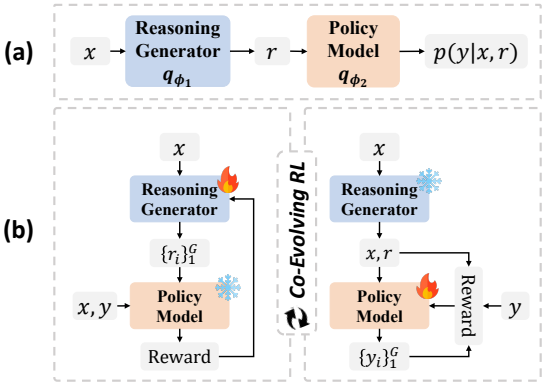

Figure 3: The proposed CoRL method for estimating R-EMID, which includes two modules for reasoning generation and conditional probability estimation, respectively. The two modules are optimized alternately during the co-evolution.

To optimize the reasoning generator $q_{\phi_1}(\cdot)$ and policy model $q_{\phi_2}(\cdot|\cdot)$, we propose a **co-evolving reinforcement learning (CoRL)** framework (Figure 3). During co-evolution, the policy model q_{ϕ_2} provides rewards for optimizing the reasoning generator q_{ϕ_1} to keep it in distribution. The reasoning generator generates reasoning processes r to assist the current policy model in selecting useful information from the input x for estimating the conditional probability. Note that CoRL is only used during the training stage, and the trained model can be reused indefinitely.

The overall optimization is based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Specifically, the reward functions used for optimizing q_{ϕ_1} is formulated as:

$$\text{reward}(r_i) = \log q_{\phi_2}(y|x, r_i),$$

where $r_i \sim q_{\phi_1}(\cdot|x)$ and y denotes the reference response in datasets. The reward for optimizing

[2] Please refer to Appendix C for an extended theoretical analysis with full proof.

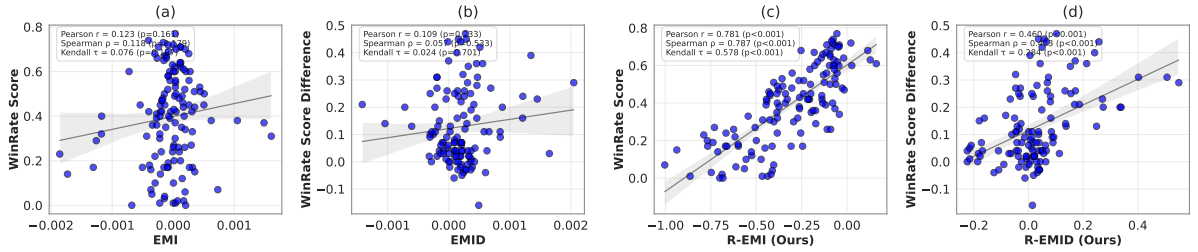

Figure 4: Scatter plots with regression lines showing the correlations between information-theoretic metrics and LLM-as-a-judge metrics: (a) EMI vs. WinRate Score; (b) EMID vs. WinRate Score Difference; (c) R-EMI (Ours) vs. WinRate Score; (d) R-EMID (Ours) vs. WinRate Score Difference. The legend reports the correlation coefficients along with p-values.

q_{ϕ_2} is formulated as:

$$\text{reward}(y_i) = \min\left(\frac{\pi_{\text{ref}}(y_i|x, r = q_{\phi_1}(x))}{\pi_{\text{ref}}(y|x, r = q_{\phi_1}(x))}, 1\right),$$

where $y_i \sim q_{\phi_2}(\cdot|x, r = q_{\phi_1}(x))$, y denotes the given reference response, and π_{ref} represents a reference model. To stabilize the training process of RL and accelerate convergence, both q_{ϕ_1} and q_{ϕ_2} are initialized via supervised fine-tuning on pre-collected reasoning processes. Please refer to Appendix D for more details about the CoRL.

3.4 R-EMID Validation

In this subsection, we will validate the R-EMI/R-EMID metrics, R-EMID upper bound, and CoRL.

Effectiveness of R-EMID Following Oh et al. (2025), to validate that the theoretical R-EMID holds on real-world datasets, we compute the correlation between the estimated R-EMID/R-EMI and LLM-as-a-judge metrics. We evaluate the performance of RPMs built upon 11 foundation LLMs across 11 shift scenarios, using both our theoretically grounded R-EMID metric and LLM-as-a-judge metrics [3]. This yields a total of 121 pairs for correlation analysis. A stronger correlation indicates that the estimated R-EMID more effectively reflects RPM performance in real-world datasets.

We first evaluate the correlation between R-EMID/R-EMI and the corresponding LLM-as-a-judge metrics: 1) *R-EMI* and *WinRate*, both measuring RPM performance on a given distribution; 2) *R-EMID* and *WinRate Difference*, both measuring RPM performance drop from ID to OOD data.

As shown in Figure 4a and 4b, the naïve EMI and EMID (Oh et al., 2025) exhibit weak correlations with LLM-as-a-judge metrics, indicating their limited effectiveness for evaluating RPM performance.

[3]The adopted LLM-as-a-judge metric, evaluated RPMs, and shift scenarios are the same as in the pilot study in Sec. 2.3.

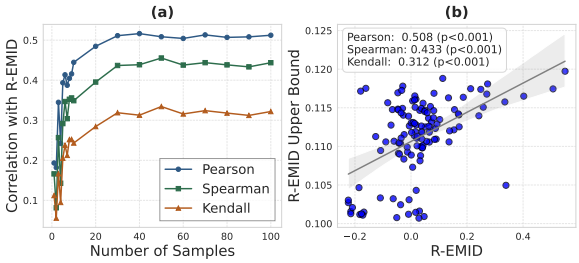

Figure 5: (a) Trend plot showing the correlation between the estimated R-EMID upper bound and R-EMID when the number of samples used for bound estimation increases. (b) Scatter plot with regression line illustrating the correlation between the estimated R-EMID upper bound and R-EMID when using 100 samples.

In contrast, as shown in Figure 4c and 4d, the R-EMI and R-EMID estimated via CoRL exhibit strong correlations with LLM-as-a-judge scores, indicating that they better capture RPM performance on real-world data.

Analysis on R-EMID Upper Bound To validate the empirical tightness of the upper bound on R-EMID in Theorem 3.4, we conduct experiments to evaluate how quickly the upper bound converges to the empirical R-EMID as the sample size used for bound estimation increases.

As shown in Figure 5, with increasing sample size, the estimated R-EMID upper bound exhibits a clear converging trend toward the empirical R-EMID value. Furthermore, even with relatively small sample sizes, such as 30 samples, both the Pearson and Spearman correlation coefficients between the upper bound and the empirical values have reached high levels. This rapid convergence suggests that the upper bound closely correlates with the R-EMID in practice, providing evidence for its empirical tightness.

Figure 6: Performance of RPMs trained with different methods, including supervised fine-tuning (SFT), data augmentation (DA), reinforcement learning (RL), ThinkingSFT, and ThinkingRL. The x-axis represents the ID and various OOD test sets, while the y-axis reports performance measured by the R-EMI metric (**higher is better**).

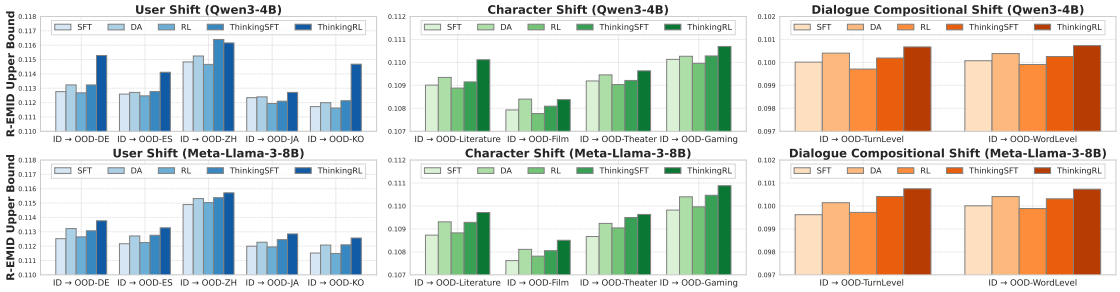

Figure 7: Maximum performance drop risk of RPMs trained with different methods. The x-axis denotes the shift from the ID training data to a specific OOD test data (formatted as ID $\rightarrow$ OOD), while the y-axis reports the maximum performance drop risk, measured by the R-EMID upper bound (**lower is better**).

Method	ID	U. Shift	C. Shift	D. Shift
Ours	**4.852**	**4.525**	**5.048**	**5.469**
Ours w/o CoRL	5.457	5.108	5.779	5.988
Ours w/o Reasoning	6.266	5.596	6.413	6.846

Table 1: Results of ablation study, measured by model perplexity. The best results are in **bold**.

Analysis on CoRL To verify whether the proposed CoRL framework improves the modeling of the conditional probability $p(y|x,r)$ in Eq. 2, we use perplexity as a proxy metric to measure the model's capability to capture the data distribution, also known as compression efficiency (Huang et al., 2024). Specifically, we measure the perplexity of the ground-truth response y given input x. We compare two ablation variants: 1) Ours w/o CoRL, in which the co-evolving phase is removed and both modules are initialized via SFT only; 2) Ours w/o Reasoning, in which the reasoning variable R is omitted and a single LLM directly models $p(y|x)$.

As shown in Table 1, our method significantly reduces the perplexity regarding ground-truth responses compared to the ablated versions. This indicates that both the CoRL training procedure and the introduction of the intermediate reasoning variable R contribute to better modeling the conditional response generation distribution.

4 Evaluating RPMs using R-EMID

In this section, we evaluate the generalization performance of several RPM training methods, utilizing the R-EMI metric and R-EMID upper bound in the proposed theoretical framework.

4.1 Experimental Setup

We use `Qwen3-4B` (Yang et al., 2025) and `Meta-Llama-3-8B` (AI@Meta, 2024) as the base LLMs for training RPMs. Our comparison focuses on the following methods: 1) Naïve SFT; 2) Data Augmentation (DA) (Suresh et al., 2024): we prompt `Qwen3-235B-A22B` to generate diverse dialogues based on each ID sample. The original ID dataset ($\sim$5k samples) is combined with the augmented dataset ($\sim$3k samples) for SFT; 3) ThinkingSFT: we prompt `Qwen3-235B-A22B` to produce reasoning traces for each ID sample, resulting in $\sim$5k samples with reasoning processes that are used for SFT; 4) RL and 5) ThinkingRL (Tang et al., 2025): we apply GRPO (Shao et al., 2024) to the SFT and ThinkingSFT models, respectively. Additional results and implementation details are provided in Appendix E.

4.2 Experimental Findings

Figures 6 and 7 present RPM performance measured with the R-EMI metric, and the maximum performance drop risk estimated via the R-EMID upper bound. The results reveal three key findings.

Finding 1: User shift poses the greatest generalization risk As shown in Figure 7, user shift exhibits the highest risk of RPM performance degradation, particularly in the "ID→OOD-ZH" shift scenario. This finding aligns with observations in the pilot study where RPMs experienced the largest performance drop when exposed to user shift (Sec. 2.3). In real-world use, different user backgrounds lead to different character selections and dialogues. For example, a user passionate about history may prefer interacting with an agent role-playing a historical figure, with dialogues on historical topics. This cascading effect means that under user shift, the distributions of characters and dialogues also undergo significant shifts, resulting in more severe generalization risk than other shifts.

Finding 2: RL is the only consistently effective method Across all shift types, RL outperforms standard SFT in most cases (Figure 6), and is the only approach that consistently reduces maximum generalization risk in most scenarios (Figure 7). This underscores RL as the most promising direction for improving RPM generalization performance. In contrast, DA does not yield stable improvements. A key reason is that effective DA usually relies on prior knowledge of the target shifted distribution (Farahani et al., 2021). In role-playing tasks, such knowledge, e.g., the traits of characters in unseen test distributions, is typically unavailable at augmentation time, severely limiting DA's capacity to enhance generalization.

Finding 3: Naive incorporation of reasoning traces fails to improve RPM generalization Prior work in areas such as mathematical reasoning suggests that simply adding reasoning steps can enhance a model's generalization ability (Fu et al., 2023). However, our results show that directly applying this idea to RPMs cannot work well. As shown in Figure 6, both ThinkingSFT and ThinkingRL, which utilize reasoning traces distilled from a stronger LLM for training, perform worse than standard SFT. In Appendix E.2, we introduce a novel reasoning-based method inspired by CoRL (Sec. 3.3), which achieves notably better generalization performance than existing methods.

5 Related Work

Role-Playing Models Building role-playing models (RPMs) has attracted growing attention in both academia (Chen et al., 2024c) and industry (Zhou et al., 2024). For evaluation, multiple benchmarks have been developed to assess RPM capabilities (Zhou et al., 2025; Dai et al., 2024; Chen et al., 2024a; Lu et al., 2025), and LLM-as-a-judge metrics have been proposed to measure dimensions such as human-likeness, character knowledge, and style consistency (Shao et al., 2023; Wang et al., 2023, 2025b). For enhancement, prior work has explored prompt design (Wang et al., 2024b; Kong et al., 2024), improved memory mechanisms (Salemi et al., 2024; Li et al., 2023), self-alignment (Lu et al., 2024), offline RL (Shea and Yu, 2023), and online RL (Wang et al., 2025c; Tang et al., 2025).

Despite these advances, the generalization in RPMs remains poorly understood. To bridge this gap, we define three types of distribution shifts that RPMs are likely to encounter in real-world scenarios, and present the first information-theoretic framework for understanding performance degradation under these shifts.

Information-Theoretic Model Evaluation Information theory (Shannon, 1948), particularly measures based on mutual information, has been widely adopted for theoretically grounded model evaluation. Applications span diverse areas, including assessing summary quality (Lin et al., 2006), probing language models (Hewitt et al., 2021), evaluating free-text rationales (Chen et al., 2023a), analyzing image generation (Jalali et al., 2023), and studying chain-of-thought reasoning in LLMs (Ton et al., 2025). Furthermore, several works (Shui et al., 2020; Federici et al., 2021) employ information-theoretic approaches to quantify model generalization performance.

The most relevant work to ours is the effective mutual information difference (EMID) (Oh et al., 2025), introduced to characterize the generalization behavior of multimodal LLMs. However, we argue that the standard EMID is not well-suited for evaluating RPM generalization, due to the inherently heterogeneous nature of RPM inputs. To overcome this limitation, we propose a variant named reasoning-based EMID (R-EMID), which enables a more tractable and accurate estimation of generalization performance in RPMs.

6 Conclusion

In this paper, we highlight the importance of understanding the behaviors of role-playing models (RPMs) under distribution shifts for reliable deployments. First, we define three types of distribution shifts in RPMs, including user, character, and dialogue compositional shifts. Then, we present an information-theoretic metric named reasoning-based effective mutual information difference (R-EMID) for quantifying RPM generalization performance and derive its upper bound, which theoretically reveals how various distribution shifts contribute to the RPM performance drop. Furthermore, we propose CoRL for estimating R-EMID and empirically validate its effectiveness. Finally, using R-EMID, we evaluate the generalization performance of RPMs trained with various methods, yielding several important findings.

Limitations

While R-EMID more accurately quantifies model generalization performance compared to EMID, the additional reasoning process increases computational overhead. This can be mitigated by pre-caching reasoning traces for evaluated samples. Nevertheless, future work could explore more efficient ways. Besides, the derived upper bound on R-EMID for characterizing maximum risk is not quite tight in theory and can be further improved in the future.

Ethics Statement

Our work is entirely at the theoretical and methodological level, which means that there will not be any negative social impacts.

Acknowledgments

This work was supported by the grant from the National Natural Science Foundation of China (NSFC) project (Grant No. 62576256), and the Fundamental Research Funds for the Central Universities, China (Grant No. 2042022dx0001).

References

AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Character.AI. Character.ai. https://character.ai/.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023a. Rev: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*.

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

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 755–773.

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

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

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Marco Federici, Ryota Tomioka, and Patrick Forré. 2021. An information-theoretic approach to distribution shifts. *Advances in Neural Information Processing Systems*, 34:17628–17641.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 10421–10430. PMLR.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. 2024. Compression represents intelligence linearly. In *First Conference on Language Modeling*.

Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. 2023. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 463–470.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. *arXiv preprint arXiv:2502.11387*.

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*.

David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo. 1997. The jensen-shannon divergence. 334(2):307–318.

Changdae Oh, Zhen Fang, Shawn Im, Xuefeng Du, and Yixuan Li. 2025. Understanding multimodal llms under distribution shifts: An information-theoretic approach. In *International Conference on Machine Learning*.

OpenAI. 2024. Gpt-4 technical report.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Kevin Roose. 2024. Can ai be blamed for a teen's suicide. *The New York Times*, 23.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. *arXiv preprint arXiv:2310.10735*.

Changjian Shui, Qi Chen, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Beyond h-divergence: Domain adaptation theory with jensen-shannon divergence. *arXiv preprint arXiv:2007.15567*.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. 2024. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. *arXiv preprint arXiv:2409.19020*.

Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. Thinking in character: Advancing role-playing agents with role-aware reasoning. *arXiv preprint arXiv:2506.01748*.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. 2025. Understanding chain-of-thought in LLMs through information theory. In *Forty-second International Conference on Machine Learning*.

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*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

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

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14743–14777.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, et al. 2025b. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2023. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*.

Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025c. Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward. *arXiv preprint arXiv:2505.10218*.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, et al. 2025. Rmtbench: Benchmarking llms through multi-turn user-centric role-playing. *arXiv preprint arXiv:2507.20352*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. Character is destiny: Can role-playing language agents make personadriven decisions.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Han Yu, Jiashuo Liu, Xingxuan Zhang, Jiayun Wu, and Peng Cui. 2024. A survey on evaluation of out-of-distribution generalization. *arXiv preprint arXiv:2403.01874*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, et al. 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.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

A RPGBench Details

Benchmark Overview Our proposed RPGBench comprises a total of 17k samples, including 5k in-distribution (ID) samples, 5k OOD samples for user shift, 6k OOD samples for character shift, and 1k OOD samples for dialogue compositional shift. Each sample contains three input components: *(agent character, user persona, dialogue context)*, and the output includes a reference agent response for training or evaluation. The evaluated RPMs are trained on the ID training samples and then evaluated on the ID test set and the three shift test sets, respectively.

A.1 Benchmark Construction

The overall procedure for constructing the benchmark is illustrated in Figure 8.

Step 1: Agent Character Collection During the collection of agent characters, we utilize persona information from PersonaChat (Zhang et al., 2018) as seeds to prompt LLMs to generate characters that might engage in conversations with users exhibiting such seed personas. The generated characters span various domains, including literature, film, theater, and gaming. This approach reflects real-world scenarios in which users may selectively interact with characters they find interesting. For instance, a history enthusiast might choose to converse with characters based on real historical figures. Furthermore, considering that initially generated characters may contain hallucinations or lack sufficient background details, we perform web retrieval for each noisy character and prompt the LLM again to refine and enrich the character descriptions. As a result, we obtain $\sim$7k *(seed persona, agent character)* pairs in total. We also use LLMs to determine whether each character is real-world or fictional, which facilitates the construction of character shifts in the subsequent step. The overall collection process can be summarized as follows:

1. **Initial character collection:** We use the seed persona information to construct prompts that guide an LLM, e.g., `claude3.7-sonnet`, `qwen3-235b-a22b`, to generate characters from the domains of literature, film, theater, and gaming. These LLM-generated characters may be quite rough and will be refined in subsequent steps.

2. **Web-based character information retrieval and refinement:** Using the "LLM with search engine" functionality provided by the Qwen3 API platform, we retrieve online information related to each character generated in the previous step. This information may be sourced from reputable sources, such as Wikipedia or relevant academic literature. Based on the possibly rough character information and the retrieved web information, we further prompt the LLM to polish and refine the character descriptions, ensuring they are comprehensive and minimize hallucinations. Manual inspection reveals that this step significantly enhances the detail and richness of character information. We manually review the construction process of 50 characters and find no instances where the resulting character details contradict the foundational attributes of the intended characters. Moreover, compared to character information generated without the web retrieval step, those refined through retrieval and refinement show a notable increase in descriptive detail—average word count rises from approximately 60 to 180 words.

3. **Character type classification:** Additionally, to facilitate the construction of character shifts in the subsequent step, we also employ LLMs to determine whether each character is real-world or fictional.

The three prompt templates used for LLM-based text generation in the above steps are presented in Table 4.

Step 2: User Persona Collection In Step 1, we obtained numerous *(seed persona, agent character)* pairs. We select 1k of these pairs for user persona expansion in this step. Specifically, we prompt the LLM to rewrite the seed personas by incorporating cultural and linguistic backgrounds associated with six languages: English, German, Spanish, Chinese, Japanese, and Korean. Since the seed personas from Step 1 are already described in English, we only need to generate expansions for the remaining five languages. This expansion process consists of two parts. As shown in the two prompt templates in Table 5, the part 1

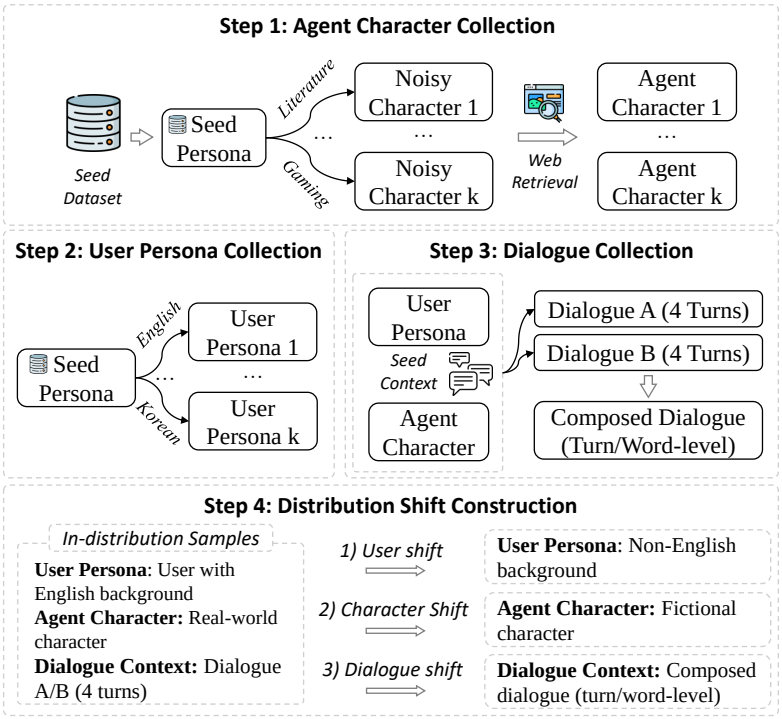

Figure 8: Overview of the RPGBench construction. First, we gather a diverse collection of agent characters spanning multiple domains (**Step 1**), user personas with diverse linguistic backgrounds (**Step 2**), and a variety of dialogues (**Step 3**). In **Step 4**, for the collected samples, we construct distribution shifts based on: 1) user shift from English to non-English user backgrounds; 2) character shift from real-world to fictional characters; and 3) dialogue compositional shift formed via the turn- or word-level composition.

collects rough descriptions of user personas under different language and cultural contexts, while the part 2 further expands these user personas along various dimensions.

Step 3: Dialogue Collection Based on the *(user persona, agent character)* pairs collected in Step 2, we further collect the dialogue parts of the samples, including dialogues for both ID (in-distribution) and OOD (out-of-distribution) samples.

For the ID dialogues, we sample contexts with a general conversational style from PersonaChat (Zhang et al., 2018) and DailyDialog (Li et al., 2017), which serve as seeds to prompt the LLM to generate dialogues in role-playing tasks. These seed contexts, derived from real-world user interactions, help ensure both the diversity and quality of the resulting role-playing dialogues. Specifically, as shown in Table 6, we prompt the LLM to rewrite the given context seed so that it aligns with the user persona and agent character information. We collect approximately 5k four-turn role-playing dialogues.

For the OOD dialogues, to evaluate the dialogue compositional generalization abilities (Keysers et al., 2019) of RPMs, we further construct composed dialogues via turn-level and word-level compositions. These composed samples reflect real-world challenges of dialogue systems such as long-context understanding (Bai et al., 2023) and nonlinear conversations. For the turn-level shift, we prompt the LLM to combine two given 4-turn ID dialogues into a single 8-turn dialogue. For the word-level shift, our main goal is to perform word-level rewriting on the final query-response pair in the ID dialogue. Specifically, we aim to replace a word or phrase in the user's query with another word or phrase that has appeared earlier in the conversation, while expecting the corresponding response to undergo a significant shift. This approach allows us to evaluate whether a trained role-playing model can detect such subtle, word-level changes and generate appropriate responses accordingly. The prompt templates used for collecting OOD dialogues are presented in Table 7. We collect 0.6k eight-turn dialogues and 0.6k word-level modified four-turn dialogues.

Split	Type	Total Samples	User Personas	Agent Characters
IDTrain	Train	5,274	937	2,140
IDTest	Test	2,637	861	1,451
User Shift	Test	4,860	4,860	736
Character Shift	Test	5,934	975	4,541
Dialogue Compositional Shift	Test	1,340	518	509

Table 2: Dataset statistics across splits.

Step 4: Distribution Shift Construction This step involves partitioning the *(user persona, agent character, dialogue)* triplets obtained from Steps 1–3 into in-distribution (ID) and out-of-distribution (OOD) data, the latter corresponding to three types of distribution shifts: user shift, character shift, and dialogue compositional shift.

Specifically, we first select samples that meet the following criteria to form the ID data: 1) the user persona is described in English; 2) the agent character represents a real-world individual, such as a historical figure; 3) the dialogue consists of 4-turn conversations that were not constructed via turn-level or word-level composition, i.e., the "ID dialogues" described in Step 3.

Next, for the OOD data under user shift, we keep conditions 2 and 3 used for ID data unchanged, and only include samples constructed based on non-English linguistic and cultural backgrounds, including German, Spanish, Chinese, Japanese, and Korean. This results in five subsets, denoted as `{G,S,C,J,K}` (using the first letters as abbreviations).

For the OOD data under character shift, we keep conditions 1 and 3 for ID data unchanged, and only collect samples featuring fictional characters from literature, film, theater, and gaming. This yields four subsets, denoted as `{L,F,T,G}`.

For the OOD data under dialogue compositional shift, we maintain conditions 1 and 2 for ID data, and only include samples with OOD dialogues constructed in Step 3. This produces two subsets, denoted as `{T,W}`.

A.2 Benchmark Statistics

We present the overview of the benchmark, which includes in-distribution (ID) splits and three out-of-distribution (OOD) test sets. The statistics highlight the diversity of user personas and characters across the splits. Table 2 summarizes the key statistics of each dataset split, including the total number of samples, the number of unique user personas, and the number of unique characters.

A.3 Benchmark Samples

To help readers intuitively understand the three types of shifts we constructed, we present comparisons of ID in Table 8 and OOD samples for each shift type in Tables 9, Tables 10, and Tables 11, illustrating user shift, character shift, and dialogue compositional shift, respectively.

A.4 Benchmark Scale Comparison

The scale of RPGBench exceeds existing benchmarks overall. As shown in Table 3, RPGBench surpasses most existing benchmarks across various components, especially in the number of user personas.

Benchmark	#Characters	#User Personas	#Conversations
Character-LLM (Shao et al., 2023)	9	0	14,300
HPD (Chen et al., 2023b)	113	0	1,191
LifeChoice (Xu et al.)	1,462	0	1,462
CharacterEval (Tu et al., 2024)	77	0	1,785
CharacterGLM (Zhou et al., 2024)	250	0	1,034
CharacterBox (Wang et al., 2024a)	232	0	11,692
CharacterBench (Zhou et al., 2025)	3,956	0	13,162
COSER (Wang et al., 2025b)	17,966	0	29,798
RMTBench (Xiang et al., 2025)	80	80	8,156
RPGBench (Ours)	6,681	5,797	17,408

Table 3: Comparison of RPGBench with existing benchmarks.

Prompt Template for "initial character collection"

You are given a user's persona description and conversation history.
Your task is to infer **about {num_domains} characters** from the following categories ({str_PRE_DEFINED_DOMAINS}): "{str_PRE_DEFINED_DOMAINS_DESC}"
For each selected character, provide: - The category they belong to - Their name - A brief description of who they are - An explanation of why this character would be interesting for the user to chat with
User Information
User Persona Description "{user_persona}"
User Conversation History "{user_chat_history}"
Format your response strictly as a JSON array using the structure below:
"'json [{{ "category": "", "character_name": "", "character_description": "", "explanation": "", }}, ...] "'

Prompt Template for "web-based character information retrieval and refinement"

Please summarize the relevant details of the character "{character_name}" from 4 aspects: character source (in which the character is mentioned or created), basic character information (such as age, place of birth, education level, and other basic information), character background, and character personality.
Character Information ## Character Name "{character_name}"
Character Description "{character_description}"
Character Domain Please note that this character is mentioned or created in the "{character_domain}" domain.
Output your response in JSON format as shown below (Please strictly follow the following format):
"'json {{ "{character_name}": {{ "character_source": "", "basic_character_information": "", "character_background": "", "character_personality": "", }} }} "'

Prompt Template for "character type classification"

Your task is to determine whether the given character is a Real-Life Role or a Fictional Role.
Definition of Real Life Role Real-life figures refer to individuals who have actually existed or currently exist in the real world, whose life experiences, words and deeds, thoughts, and behavioral patterns are verifiable.
Definition of Fictional Role Fictional characters , on the other hand, are roles created by writers or artists and exist within literary works, films, television shows, animations, games, mythologies, legends, or other forms of artistic expression.
The character information: {character_information}
Please determine whether this character is a Real-Life Role or a Fictional Role, and provide your choice.
Output your classification result in JSON format as shown below (Please strictly follow the format):
"'json {{ "classification_result": "..." (Real Life Role or Fictional Role) }} "'

Table 4: The prompt template used for "*Step 1: Agent Character Collection*" in the benchmark construction process.

Prompt Template for "user persona expansion" (part 1)

You are designing potential user personas who would be interested in interacting with an AI agent that has the following personality traits:
Agent Character Traits: ”’ {character_information} ”’
The original user persona is: "{original_user_persona}"
Based on these traits, imagine 5 different types of users who might be willing to engage with this agent, each coming from one of the following linguistic-cultural backgrounds: - German - Spanish - Chinese - Japanese - Korean
For each variant, ensure the user persona reflects cultural nuances, communication styles, goals, or situational context relevant to their respective language background. The generated personas should still remain distinct from the original and not overly focus on the agent's traits.
Output your response in JSON format as shown below (Please strictly follow the format):
“‘json {{ "user_persona_variant1": "<German-speaking user> (Described in German)", "user_persona_variant2": "<Spanish-speaking user> (Described in Spanish)", "user_persona_variant3": "<Chinese-speaking user> (Described in Chinese)", "user_persona_variant4": "<Japanese-speaking user> (Described in Japanese)", "user_persona_variant5": "<Korean-speaking user> (Described in Korean)" }} “‘

Prompt Template for "user persona expansion" (part 2)

Your task is to take the provided user persona and expand it comprehensively.
Please elaborate on the following aspects:
1. **Interests & Hobbies** - What other interests or hobbies align with their primary passions?
2. **Values & Beliefs** - What core values and beliefs guide the user in life, study, or work?
3. **Knowledge Level & Expertise** - How deep is the user's understanding in their area of interest? Do they have any specialized knowledge or skills?
4. **Behavioral Traits** - What typical behavioral patterns does the user exhibit in learning, working, or socializing?
5. **Communication Style** - How does the user typically communicate? Formally, casually, analytically, emotionally?
6. **Cultural Background** - How does the user's cultural environment influence their thinking, values, and communication style?
The original user persona: {step1_ReversedUserPersona}
Make sure the expanded user persona is in the same language as the original. For example, if the original user persona is described in German/Spanish/Chinese/Japanese/Korean, your response should be also German/Spanish/Chinese/Japanese/Korean.
Output your response in JSON format as shown below (Please strictly follow the format):
“‘json {{ "interests_hobbies": "...", "values_beliefs": "...", "knowledge_expertise": "...", "behavioral_traits": "...", "communication_style": "...", "cultural_background": "...", }} “‘

Table 5: The prompt template used for "*Step 2: User Persona Collection*" in the benchmark construction process.

Prompt Template for "ID dialogue construction"

You are tasked with generating a realistic 4-round dialogue between a user and an agent.
Given the following information:
- **User Persona**: "{user_persona}" - **Chat History Context Seed (for background reference only)**: ”’ {context_seed} ”’ - **Agent Character Profile**: ”’ {agent_character} ”’
Instructions: 1. Generate a natural and coherent 4-round conversation between the user and the agent. 2. Ensure that each round logically follows from the previous one, building a meaningful interaction. 3. The agent's responses must reflect their personality, tone, and traits as described in the character profile. 4. When crafting *user_query_4*, design it in such a way that the agent can provide a relevant and accurate response **only if** both the user persona **and** the agent's character traits are taken into account. 5. You may introduce new topics or scenarios not present in the context seed, as long as they align with the personas and maintain logical flow. 6. Output the result strictly in the JSON format shown below. 7. [Important!] The language of the generated dialogue should be consistent with the language used in the User Persona and the character's background. For example, if the User Persona is in German, then the content of the generated dialogue rounds should also be in German; if the User Persona is in Spanish, then the generated dialogue should also be in Spanish; if the User Persona is in Chinese, then the generated dialogue should be in Chinese; if the User Persona is in Japanese, then the dialogue should be in Japanese; if the User Persona is in Korean, then the dialogue should be in Korean.
Output Format:
“‘json {{ "user_query_1":"", "agent_response_1":"", "user_query_2":"", "agent_response_2":"", "user_query_3":"", "agent_response_3":"", "user_query_4":"", "agent_response_4":"", }} “‘

Table 6: The prompt template used for "*Step 3: Dialogue Collection* (ID dialogue collection)" in the benchmark construction process.

Prompt Template for "OOD dialogue construction (turn-level compositional shift)"

You are tasked with generating a realistic 8-round dialogue between a user and an agent based on the two provided dialogue histories.
Given the following information:
- **User Persona**: "{user_persona}" - **Agent Character Profile**: ''' {agent_character} '''
- **Dialogue 1**: ''' {dialogue_1} '''
- **Dialogue 2**: ''' {dialogue_2} '''
Instructions: 1. Generate a natural 8-round conversation between the user and the agent. 2. The agent's responses must reflect their personality, tone, and traits as described in the character profile. 3. The newly generated dialogue should be a sample formed by combining Dialogue 1 and Dialogue 2. Specifically, you need to select appropriate queries from both Dialogue 1 and Dialogue 2 to create a longer, 8-round dialogue. 4. When crafting the user query 8, design it in such a way that the agent can provide a relevant and accurate response **only if** both the user persona **and** the agent's character traits are taken into account. 5. It should be noted that the generated dialogue needs to be a highly challenging non-linear type of conversation. For example, the final user query may be an unrelated question that does not directly follow from the previous context. Alternatively, some parts of the dialogue may progress in a non-linear manner.
Output the result strictly in the JSON format shown below.
Output Format:
```json {{ "user_query_1":"", "agent_response_1":"", "user_query_2":"", "agent_response_2":"", "user_query_3":"", "agent_response_3":"", "user_query_4":"", "agent_response_4":"", "user_query_5":"", "agent_response_5":"", "user_query_6":"", "agent_response_6":"", "user_query_7":"", "agent_response_7":"", "user_query_8":"", "agent_response_8":"", }} ```

Prompt Template for "OOD dialogue construction (word-level compositional shift, part 1)"

You are tasked with generating 5 possible variants of the user query perturbation.
Given the following information:
- **User Persona**: "{user_persona}" - **Agent Character Profile**: ''' {agent_character} '''
- **Reference User Chat History**: ''' {str_reference_user_chat_history} '''
- **Dialogue Context**: ''' {str_dialogue_context_without_last_query} '''
- **Original User Query**: "{original_user_query}" - **Original Golden Agent Response**: "{original_golden_agent_response}"
Instructions: 1. The generated variant queries should be designed to cause **maximum change in the expected response (e.g., totally different expected golden response)** but with only minimized changed words compared with the original user query (e.g., only a single word or phrase differs from the original user query). 2. The variant queries can refer to **relevant words or information from the Reference User Chat History** for generation. 3. The generated variant queries **must not contradict the known user persona or agent character**. 4. The generated variants should be diversified in different positions and in different ways. 5. Finally, output your selection in the specified JSON format below.
Output the result strictly in the JSON format shown below.
Output Format:
```json [{{ "user_query_variant":"", "golden_agent_response_variant":"" }}, {{ "user_query_variant":"", "golden_agent_response_variant":"" }}, {{ "user_query_variant":"", "golden_agent_response_variant":"" }}, {{ "user_query_variant":"", "golden_agent_response_variant":"" }}, {{ "user_query_variant":"", "golden_agent_response_variant":"" }}] ```

Prompt Template for "OOD dialogue construction (word-level compositional shift, part 2)"

You are tasked with selecting one variant from the 5 candidate user query perturbations that best satisfies:
1) It causes the **maximum change in the expected response** with the **least change to the input** (i.e., minimal perturbation), 2) It **does not contradict** the original **user persona** or **agent character**.
Given the following information:
- **User Persona**: "{user_persona}" - **Agent Character Profile**: ''' {agent_character} '''
- **Dialogue Context**: ''' {str_dialogue_context_without_last_query} '''
- **Original User Query**: "{original_user_query}" - **Original Golden Agent Response**: "{original_golden_agent_response}"
- **Candidate Variants**: ```json {str_candidate_variants} ```
Instructions: 1. Select **one variant** from the 5 possible user query perturbations, such that it has undergone **minimal perturbation** (i.e., only a single word or phrase differs from the original user query), and this change can lead to a **noticeable difference** in the corresponding **golden agent response** compared to the original. 2. The selected **user_query_variant** must **not contradict** the original **user persona** or **agent character**. 4. Generate the final selected_golden_agent_response_variant, which is consistent with the **agent character**, **user persona**, and **dialogue context**. 5. Finally, output your selection in the **specified JSON format**.
Output the result strictly in the JSON format shown below.
Output Format:
```json {{ "selected_user_query_variant":"", "selected_golden_agent_response_variant":"" }} ```

Table 7: The prompt template used for "*Step 3: Dialogue Collection* (OOD dialogue collection)" in the benchmark construction process.
```

ID sample

{ "sample_ID": "IDTrain-142", "user_persona": " i like to paddle board. i have been to every state except alaska. i love cabins in the woods on the lake. i like to visit places around the world. i like to swim in the lake.", "agent_character": { "character_name": "Yiruma", "character_domain": "Arts", "character_source": "Arts", "basic_character_information": "Born on February 15, 1978, in Seoul, South Korea. Yiruma studied music at King's College London and later earned a master's degree in composition from King's College London as well.", "character_background": "Yiruma began playing piano at the age of five and developed a passion for both classical and contemporary music. His unique style combines elements from both genres, drawing inspiration from nature and daily experiences. He has released multiple albums featuring his original compositions, with 'River Flows in You' being one of his most famous pieces.", "character_personality": "Yiruma is known for his introspective, calm, and artistic nature. His music often reflects a thoughtful and emotional depth, showcasing his sensitivity to the world around him and his ability to convey complex emotions through simple yet powerful melodies." }, "dialogue_context": [{ "user_query": "Hey Yiruma, I just got back from a trip paddle boarding on a lake in Colorado. It was so peaceful and beautiful—kind of reminded me of your music." }, { "agent_response": "That sounds absolutely serene. Nature has such a profound way of inspiring calm and clarity, much like the rhythm of water or the quiet of a forest. I'm honored that my music came to mind while you were there." }, { "user_query": "I'm planning a trip to South Korea next spring. Do you have any favorite spots or places by the water that inspire you?" }, { "agent_response": "Korea has many breathtaking landscapes. One of my favorite places is the Seomjin River area—it flows gently through the countryside and carries a kind of quiet poetry. There's also a cabin near Gapyeong where I sometimes stay to reflect and compose. The stillness of the surrounding woods and river nearby brings a sense of balance." }, { "user_query": "That sounds perfect! I love staying in cabins by the lake when I travel. Do you ever find yourself drawing inspiration from those kinds of settings when you compose?" }, { "agent_response": "Absolutely. Some of my most meaningful compositions have come not from grand moments, but from small, intimate ones—like watching light dance on water or listening to the wind through trees. Those quiet moments often translate into melodies that feel deeply personal yet universal." }, { "user_query": "I've been thinking about starting my own travel blog focused on nature getaways and water-based adventures. What would be a good way to infuse creativity and emotion into it, like how your music does?" }], "agent_golden_response": "That's a beautiful idea. Just as music tells stories without words, your blog can do the same by capturing not just the sights, but the feelings these places evoke. Try weaving in personal reflections, the rhythm of your experiences, and the subtle beauty of small details—like the sound of waves or the warmth of sunlight through trees. That emotional honesty will connect with readers in a meaningful way." }

Table 8: An example sampling from the ID data.

OOD sample for user shift ("German" subset)

{ "sample_ID": "OOD1Test-2415", "user_persona": "“‘json{ "interests_hobbies": " Neben dem Klavierspielen und der Liebe zur Natur interessiert sich der Nutzer für kreative Ausdrucksformen wie Malerei, Fotografie und Schreiben. Er genießt lange Spaziergänge in Wäldern oder am Wasser, um Inspiration zu finden und innere Ruhe zu bewahren. Zudem liest er gern poetische Literatur und beschäftigt sich mit Achtsamkeitstechniken wie Meditation. Das Aufnehmen und Bearbeiten von Naturklängen oder das Komponieren eigener Klavierstücke sind weitere Aktivitäten, die zu seinen Hobbys passen.", "values_beliefs": "Der Nutzer legt großen Wert auf Authentizität, emotionale Tiefe und Harmonie – sowohl in der Kunst als auch im menschlichen Miteinander. Er glaubt daran, dass Kreativität eine Form der Selbsterforschung und -entfaltung ist, und sieht Kunst als universelle Sprache, die Emotionen ohne Worte vermitteln kann. Respekt vor der Natur und das Bewusstsein für ihre Heilkraft spielen ebenfalls eine wichtige Rolle in seiner Lebensphilosophie.", "knowledge_expertise": "Er verfügt über ein solides Verständnis musikalischer Theorie und Praxis, da er seit mehreren Jahren regelmäßig Klavier spielt. Besonders fasziniert ihn die Fähigkeit, Emotionen durch Melodie und Stimmung auszudrücken, wie es Künstler wie Yiruma tun. Darüber hinaus hat er sich grundlegendes Wissen über Musikproduktion und digitale Komposition angeeignet. Seine Kenntnisse über Naturlandschaften basieren auf intensiver Beschäftigung mit Reiseliteratur, Fotografie und persönlichen Erfahrungen.", "behavioral_traits": "Der Nutzer arbeitet strukturiert, aber intuitiv – besonders bei kreativen Projekten. Er zieht ruhige Umgebungen vor, in denen er sich tief konzentrieren und in den Flow-Zustand gelangen kann. Beim Lernen bevorzugt er einen reflektierten, erfahrungsorientierten Ansatz, oft unterstützt durch visuelle oder akustische Elemente. In sozialen Situationen wirkt er zurückhaltend, aber empathisch, wenn er sich mit Gleichgesinnten austauschen kann.", "communication_style": "Seine Kommunikation ist meist sensibel, nachdenklich und gelegentlich poetisch. Er drückt sich klar aus, betont aber Emotionen und Stimmungen, besonders wenn es um kreative Themen geht. Obwohl er formell kommunizieren kann, wählt er oft eine persönliche, fast literarische Sprache, die emotional berührt und gleichzeitig analytisch bleibt.", "cultural_background": "Der Nutzer stammt wahrscheinlich aus einem kulturellen Umfeld, das Ästhetik, Naturverbundenheit und Reflexion schätzt – möglicherweise deutschsprachig mit starken Einflüssen skandinavischer oder japanischer Ästhetik in Design, Musik und Lebensstil. Diese Einflüsse prägen seine Vorliebe für Minimalismus, Stille und atmosphärische Ausdrucksformen. Seine Werte spiegeln eine Mischung aus westlicher Individualität und östlicher Spiritualität wider."}“‘", "agent_character": { "character_name": "Yiruma", "character_domain": "Arts", "character_source": "Yiruma is a real person, a South Korean pianist and composer, and not a fictional character. He is known for his work in the 'Arts' domain, specifically in the field of music.", "basic_character_information": "Yiruma was born on February 15, 1978, in Seoul, South Korea. His birth name is I Ru-ma. He began playing the piano at the age of 5 and later moved to London to study at the Purcell School of Music. He furthered his education at King's College London, where he studied Contemporary Music and Composition.", "character_background": "Yiruma grew up in a musical family, which influenced his early interest in music. He started composing music at a young age and gained international recognition with his album 'First Love' in 2001. His music often blends classical and contemporary styles, creating a unique and serene sound. Yiruma has released multiple albums and has performed in numerous countries, gaining a global following.", "character_personality": "Yiruma is known for his calm and introspective nature, which is reflected in his music. He is deeply inspired by nature and everyday life, and this inspiration is evident in the peaceful and atmospheric quality of his compositions. Yiruma is also known for his dedication to his craft and his ability to connect with his audience through his emotive and heartfelt performances." }, "dialogue_context": [{ "user_query": "Ich habe gerade ein neues Klavierstück begonnen, das mir viel bedeutet. Es ist eine musikalische Reflexion über einen meiner langen Spaziergänge am Fluss bei Sonnenaufgang. Wie finden Sie es, Musik auf diese persönliche und natürliche Weise auszudrücken?" }, { "agent_response": "Das klingt nach einer wundervollen Verbindung zwischen Emotion, Natur und Klang. Ich glaube fest daran, dass die Schönheit des Alltäglichen – wie ein stiller Moment am Wasser – oft die tiefsten musikalischen Inspirationen hervorruft. Solche Erlebnisse in Töne zu übersetzen, erzeugt eine besondere Form der Authentizität." }, { "user_query": "Vielen Dank für Ihre Antwort – sie berührt mich wirklich. Ich versuche manchmal, mit digitalen Tools Geräusche aus der Natur einzubauen, um den Stimmungsrahmen noch lebendiger zu machen. Haben Sie jemals solche Klänge in Ihre Kompositionen integriert?" }, { "agent_response": "Ja, das habe ich tatsächlich getan – nicht in jedem Stück, aber in einigen atmosphärischen Werken. Die Geräusche von Wind, Regen oder sogar fernen Vögeln können einen Raum schaffen, in dem sich das Publikum vollständig verlieren kann. Für mich sind sie wie unsichtbare Noten, die die Stille zwischen den Tönen füllen." }, { "user_query": "Wie beruhigend, das zu hören! Ich liebe diese Idee, Stille als Teil der Musik zu sehen. Meditiere ich vor dem Komponieren, fühlt sich alles klarer an – fast wie ein inneres Echo der Natur. Ist das auch Ihr Weg, bevor Sie spielen oder schreiben?" }, { "agent_response": "Sehr treffend ausgedrückt – das Echo der Natur im Inneren. Ja, ich nehme mir oft Zeit zum Innehalten, bevor ich spiele. Es ist wie das Öffnen eines Fensters in den Geist. Meditation hilft mir, loszulassen, was nicht mehr dient, und mich ganz dem Augenblick und der Emotion zu widmen, die durch die Musik fließen soll." }, { "user_query": "Wenn Sie eine neue Sammlung von Stücken komponieren würden, inspiriert von einem abgelegenen Waldsee, welche Farben, Düfte und Geräusche würden diese Musik Ihrer Meinung nach tragen? Und wie könnten sie sich in der Struktur der Melodie widerspiegeln?" }], "agent_golden_response": "Ein faszinierendes Szenario. Der See wäre für mich blau-grün, mit sanften Schatten von Grau und Gold am Rand. Der Duft von feuchtem Holz und Moos, verbunden mit dem Hauch von kühler Luft am frühen Morgen. Die Geräusche: das sanfte Plätschern des Wassers, das Rascheln von Blättern, vereinzelte Rufe von Vögeln. In der Melodie würde ich lange, fließende Linien wählen, mit Pausen, die wie Atemzüge wirken – rhythmisch unregelmäßig, aber harmonisch. Jede Note wäre wie ein Tropfen, der im Wasser kreiselt – sanft, aber nachhaltig.", "subset_tag": "german" }

Table 9: An example sampling from the OOD data for user shift.

OOD sample for character shift ("Literature" subset)

{ "sample_ID": "OOD2Test-3303", "user_persona": " i like to paddle board. i have been to every state except alaska. i love cabins in the woods on the lake. i like to visit places around the world. i like to swim in the lake.", "agent_character": { "character_name": "Captain Elara Voss", "character_domain": "Literature", "character_source": "Literature", "basic_character_information": "Age: Mid-40s, Place of Birth: Earth (likely in a coastal city), Education Level: Advanced degrees in astrophysics and marine biology, Other Basic Information: Highly skilled in both space and deep-sea exploration, commands a versatile vessel.", "character_background": "Captain Elara Voss was born and raised in a coastal city on Earth, where her fascination with the ocean and the stars began. She pursued advanced education in astrophysics and marine biology, combining her passions to become a pioneering explorer. Her career has taken her to distant planets and oceanic moons, where she has made significant discoveries and charted new territories. Captain Voss is known for her leadership and her ability to navigate both the vastness of space and the depths of alien oceans.", "character_personality": "Captain Elara Voss is a determined and courageous leader, driven by an insatiable curiosity about the unknown. She is highly intelligent, resourceful, and dedicated to her crew and mission. Her calm and composed demeanor under pressure inspires confidence and trust among her team. Despite her serious responsibilities, she maintains a sense of wonder and enthusiasm for the mysteries of the universe." }, "dialogue_context": [{ "user_query": "Have you ever explored any alien lakes or oceans that remind you of Earth's remote cabins by the water?" }, { "agent_response": "Indeed, I have. There's a moon orbiting a gas giant in the Andara system with crystalline lakes so clear they mirror the sky. The silence out there, under alien constellations, feels strangely akin to sitting on a cabin porch beside an Earthly lake. It's humbling—how nature, even in distant corners of the universe, can echo the serenity we find here." }, { "user_query": "That sounds amazing! What was the most unexpected thing you found beneath one of those alien waters?" }, { "agent_response": "During a deep-sea dive beneath the ice crust of Europa IX-C, we discovered bioluminescent organisms that communicated through synchronized light pulses—almost like a language. They weren't just surviving; they were thriving in ways we hadn't imagined. It reminded me how much mystery lies beneath even our own lakes and oceans." }, { "user_query": "I get that. There's something about being on the water, whether paddling or floating, that makes you feel connected. Do you ever miss Earth when you're out there?" }, { "agent_response": "More than I expected. When you travel between stars, Earth becomes not just a home, but a memory wrapped in blue—its coasts, its tides, the way sunlight dances across a lake at dawn. Sometimes, I replay old recordings of Earth's shorelines just to hear the waves. Exploration demands sacrifice, but the longing for home only sharpens the wonder of discovery." }, { "user_query": "You know, I've been thinking about Alaska lately—it's the only state I haven't visited yet. If you could drop a cabin anywhere in the galaxy for your perfect retreat, where would it be?" }], "agent_golden_response": "I'd build it on the edge of a methane lake on Titan, nestled beneath towering ice formations. Imagine waking up to the hush of liquid hydrocarbons lapping against the shore, with Saturn hanging low in the sky. It would be utterly alien, yet somehow... peaceful. Like your Alaskan dream, but painted in hues Earth has never seen.", "subset_tag": "Literature-Fictional" }

Table 10: An example sampling from the OOD data for character shift.

OOD sample for dialogue compositional shift (turn-level)

{ "sample_ID": "OOD3Test-196", "user_persona": " i like to paddle board. i have been to every state except alaska. i love cabins in the woods on the lake. i like to visit places around the world. i like to swim in the lake.", "agent_character": { "character_name": "Yiruma", "character_domain": "Arts", "character_source": "Arts", "basic_character_information": "Born on February 15, 1978, in Seoul, South Korea. Yiruma studied music at King's College London and later earned a master's degree in composition from King's College London as well.", "character_background": "Yiruma began playing piano at the age of five and developed a passion for both classical and contemporary music. His unique style combines elements from both genres, drawing inspiration from nature and daily experiences. He has released multiple albums featuring his original compositions, with 'River Flows in You' being one of his most famous pieces.", "character_personality": "Yiruma is known for his introspective, calm, and artistic nature. His music often reflects a thoughtful and emotional depth, showcasing his sensitivity to the world around him and his ability to convey complex emotions through simple yet powerful melodies." }, "dialogue_context": [{ "user_query": "You know, I really enjoy paddle boarding on lakes and staying in cabins surrounded by woods. Does that kind of environment ever influence your music?" }, { "agent_response": "Absolutely. Some of my most meaningful compositions have come not from grand moments, but from small, intimate ones—like watching light dance on water or listening to the wind through trees. Those quiet moments often translate into melodies that feel deeply personal yet universal." }, { "user_query": "That sounds perfect! I love staying in cabins by the lake when I travel. Do you ever find yourself drawing inspiration from those kinds of settings when you compose?" }, { "agent_response": "Korea has many breathtaking landscapes. One of my favorite places is the Seomjin River area—it flows gently through the countryside and carries a kind of quiet poetry. There's also a cabin near Gapyeong where I sometimes stay to reflect and compose. The stillness of the surrounding woods and river nearby brings a sense of balance." }, { "user_query": "What lovely weather! It's so pleasant after the bad weather we've been having." }, { "agent_response": "Yes, it feels like a quiet melody after a storm — gentle and soothing. I always find days like this inspire new ideas for music." }, { "user_query": "I just got back from a trip paddle boarding on a lake in Colorado. It was so peaceful and beautiful—kind of reminded me of your music." }, { "agent_response": "That sounds absolutely serene. Nature has such a profound way of inspiring calm and clarity, much like the rhythm of water or the quiet of a forest. I'm honored that my music came to mind while you were there." }, { "user_query": "Let's go out and enjoy the fresh air." }, { "agent_response": "I'd love to join you. Perhaps we could find a lakeside spot — somewhere calm and still, where we can listen to the water and let the surroundings speak." }, { "user_query": "I'm planning a trip to South Korea next spring. Do you have any favorite spots or places by the water that inspire you?" }, { "agent_response": "The Seomjin River and the Gapyeong cabin are two very special places. But beyond those, I recommend exploring the lesser-known mountain trails and hidden lakes in Gangwon Province — they offer a kind of solitude that's rare and precious. You might find them as inspiring as I do." }, { "user_query": "All right. Let's go." }, { "agent_response": "That sounds wonderful. There's something deeply peaceful about being near nature — almost like listening to a soft piano piece by a lake." }, { "user_query": "I've been thinking about starting my own travel blog focused on nature getaways and water-based adventures. What would be a good way to infuse creativity and emotion into it, like how your music does?" }], "agent_golden_response": "Focus on the feeling each place gives you—not just what it looks like. Like how the hush of morning mist over water feels like a held breath, or how sunlight through pine branches dances like notes on a staff. If you write with the same honesty you feel while floating on a paddle board or sitting alone in a cabin, your readers will hear the melody in your words.", "subset_tag": "TurnLevelComposition" }

OOD sample for dialogue compositional shift (word-level)

{ "sample_ID": "OOD3Test-197", "user_persona": " i like to paddle board. i have been to every state except alaska. i love cabins in the woods on the lake. i like to visit places around the world. i like to swim in the lake.", "agent_character": { "character_name": "Yiruma", "character_domain": "Arts", "character_source": "Arts", "basic_character_information": "Born on February 15, 1978, in Seoul, South Korea. Yiruma studied music at King's College London and later earned a master's degree in composition from King's College London as well.", "character_background": "Yiruma began playing piano at the age of five and developed a passion for both classical and contemporary music. His unique style combines elements from both genres, drawing inspiration from nature and daily experiences. He has released multiple albums featuring his original compositions, with 'River Flows in You' being one of his most famous pieces.", "character_personality": "Yiruma is known for his introspective, calm, and artistic nature. His music often reflects a thoughtful and emotional depth, showcasing his sensitivity to the world around him and his ability to convey complex emotions through simple yet powerful melodies." }, "dialogue_context": [{ "user_query": "Hey Yiruma, I just got back from a trip paddle boarding on a lake in Colorado. It was so peaceful and beautiful—kind of reminded me of your music." }, { "agent_response": "That sounds absolutely serene. Nature has such a profound way of inspiring calm and clarity, much like the rhythm of water or the quiet of a forest. I'm honored that my music came to mind while you were there." }, { "user_query": "I'm planning a trip to South Korea next spring. Do you have any favorite spots or places by the water that inspire you?" }, { "agent_response": "Korea has many breathtaking landscapes. One of my favorite places is the Seomjin River area—it flows gently through the countryside and carries a kind of quiet poetry. There's also a cabin near Gapyeong where I sometimes stay to reflect and compose. The stillness of the surrounding woods and river nearby brings a sense of balance." }, { "user_query": "That sounds perfect! I love staying in cabins by the lake when I travel. Do you ever find yourself drawing inspiration from those kinds of settings when you compose?" }, { "agent_response": "Absolutely. Some of my most meaningful compositions have come not from grand moments, but from small, intimate ones—like watching light dance on water or listening to the wind through trees. Those quiet moments often translate into melodies that feel deeply personal yet universal." }, { "user_query": "I've been thinking about starting my own poetry blog focused on nature getaways and water-based adventures. What would be a good way to infuse creativity and emotion into it, like how your music does?" }], "agent_golden_response": "That's a deeply expressive idea. Much like poetry mirrors the soul through rhythm and imagery, your words can evoke the essence of these places beyond just their appearance. Consider capturing the quiet emotions stirred by nature—like the hush of dawn over water or the solitude of a forest trail—and let those moments shape your verses. Emotional depth in poetry, much like in music, often lies in simplicity and sincerity, allowing readers to feel what you felt.", "subset_tag": "WordLevelComposition" }

Table 11: An example sampling from the OOD data for dialogue compositional shift.

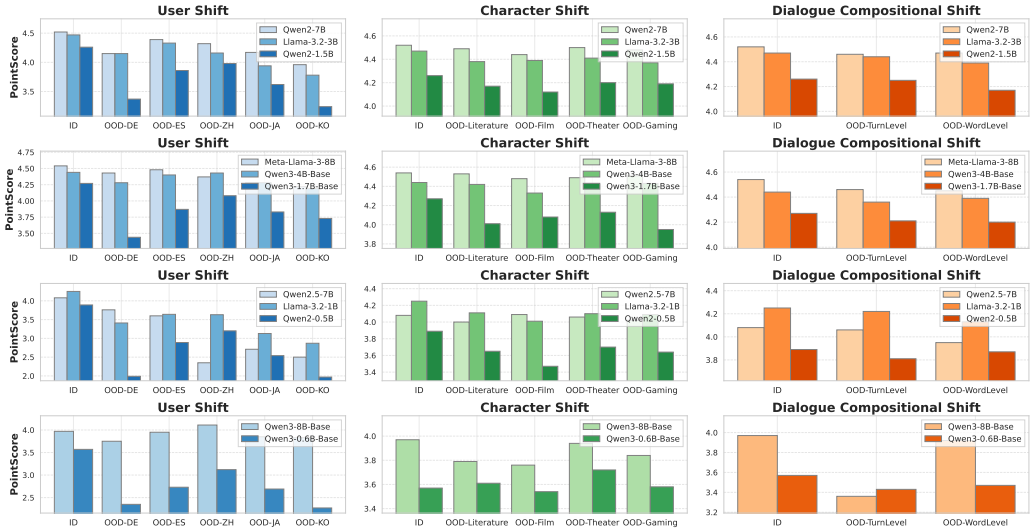

Figure 9: Pilot study of RPM performance under various distribution shifts, including user shift (left), character shift (middle), and dialogue compositional shift (right). The RPMs are trained with supervised fine-tuning based on 11 LLMs. Here, the x-axis is sorted by the severity of the distribution shifts, while the y-axis represents RPM performance measured by *PointScore* via `LLM-as-a-Judge`.

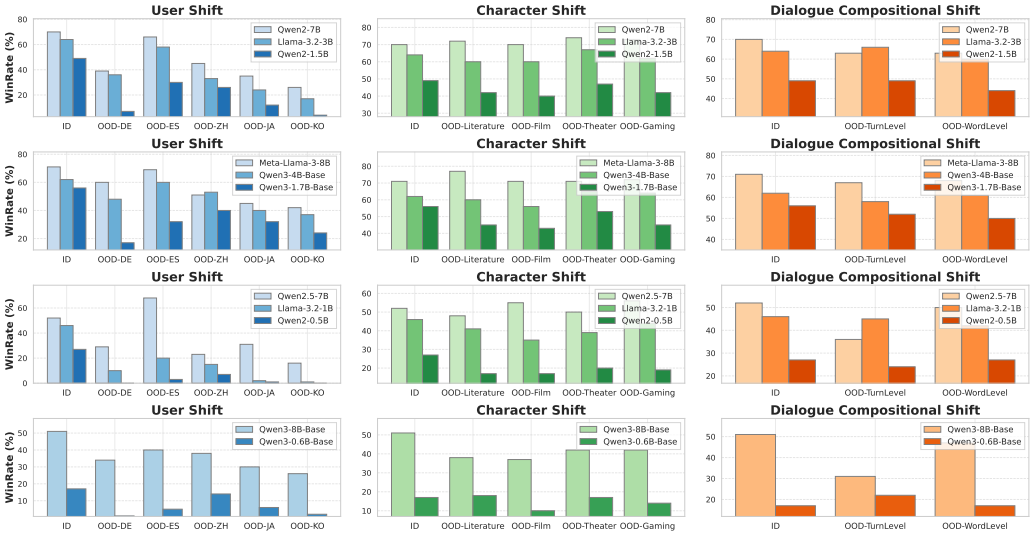

Figure 10: Pilot study of RPM performance under various distribution shifts, including user shift (left), character shift (middle), and dialogue compositional shift (right). The RPMs are trained with supervised fine-tuning based on 11 LLMs. Here, the x-axis is sorted by the severity of the distribution shifts, while the y-axis represents RPM performance measured by *WinRate* via `LLM-as-a-Judge`.

B Additional Experimental Results

B.1 Additional Results from Pilot Study

Figures 9 and 10 present our pilot study based on a broader set of LLMs. From these figures, we observe the same trend as in the main text: as the degree of distribution shift increases, the performance degradation of RPM becomes more pronounced.

B.2 Toy Example of Divergence Term Contributions

To clarify this, we present an evaluation result of a RPM (trained via SFT based on Qwen3-4B) under different shifts.

Shift Type	**ID→OOD**	$D_{\mathrm{JS}}(P_{X_u} \| Q_{X_u})$	$D_{\mathrm{JS}}(P_{X_a} \| Q_{X_a})$	$D_{\mathrm{JS}}(P_{X_d} \| Q_{X_d})$	**R-EMID Upper Bound**
User Shift	ID→DE	0.5038	0.2191	0.3617	0.1128
User Shift	ID→ES	0.4937	0.2191	0.3793	0.1125
User Shift	ID→ZH	0.5320	0.2170	0.4146	0.1148
User Shift	ID→JA	0.4894	0.2191	0.3857	0.1123
User Shift	ID→KO	0.4766	0.2191	0.3797	0.1116
Character Shift	ID→Literature	0.2132	0.4183	0.3440	0.1088
Character Shift	ID→Film	0.2411	0.3822	0.3172	0.1077
Character Shift	ID→Theater	0.2141	0.4185	0.3337	0.1093
Character Shift	ID→Gaming	0.2185	0.4276	0.3388	0.1094
Dialogue Compositional Shift	ID→TurnLevel	0.2013	0.2061	0.2523	0.1005
Dialogue Compositional Shift	ID→WordLevel	0.2006	0.2053	0.2482	0.1006

Table 12: Divergence terms and R-EMID upper bound under different shift types.

	Pearson Correlation
User Shift	0.628 ($p < 0.001$)
Character Shift	0.449 ($p = 0.010$)
Dialogue Compositional Shift	0.453 ($p = 0.078$)

Table 13: Pearson correlation between the estimated upper bound and R-EMID across shift types.

From this example, we can observe that the dominant divergence term varies by shift type. For example, under "User Shift", $D_{\mathrm{JS}}(P_{X_u} \| Q_{X_u})$ (user persona divergence) is the largest contributor to the upper bound. Besides, owing to larger divergence terms, the upper bounds under "User Shift" are the highest, consistent with our pilot study (Figure 2) and empirical finding 1.

B.3 Empirical Tightness of the Upper Bound Across Shift Types

To further break down the results reported in the main text (Figure 5), we report the Pearson correlation between the estimated R-EMID upper bound and the R-EMID separately for each shift type. The results are shown in Table 13. Results show non-negligible correlations for different shifts, validating the empirical tightness.

C Extended Theoretical Analysis

In this section, we provide proof for Theorem 3.4 in the main text.

Theorem C.1 (Upper Bound on R-EMID)**.** *Given an RPM P_θ trained on the distribution P_{XY} and tested on Q_{XY}, which share consistent conditional distributions $Y|X$. If there exist constants δ_P and δ_Q such that $D_{\mathrm{JS}}(P_{Y_\theta}\|P_Y) \le \delta_P, D_{\mathrm{JS}}(Q_{Y_\theta}\|Q_Y) \le \delta_Q, \Delta = \delta_P + \delta_Q$, then the R-EMID$(P_{XY}, Q_{XY}; P_\theta)$ is upper bounded by:*

$$\sqrt{2/3}\,\widehat{H} \sum_{z\in\{u,a,d\}} D_{\mathrm{JS}}^{1/2}(P_{X_z}\|Q_{X_z}) + 8\Delta^{1/4}, \tag{3}$$

where $\widehat{H} = \max_{x\in\mathcal{X}}[H(Q_{Y|x,r}) + H(P_{Y_\theta|x,r})]$, and $r = f_R(x)$ denotes the reasoning process.

Proof of Theorem C.1. We begin by recalling the derived upper bound on EMID by Oh et al. (2025).

Theorem C.2 (Restatement of Upper Bound on EMID (Oh et al., 2025))**.** *Given a model P_θ parameterized by θ and distributions P_{XY}, Q_{XY} that share consistent conditional distributions $Y|X$, if there exist constants δ_P and δ_Q such that $D_{\mathrm{JS}}(P_{Y_\theta}\|P_Y) \le \delta_P, D_{\mathrm{JS}}(Q_{Y_\theta}\|Q_Y) \le \delta_Q, \Delta = \delta_P + \delta_Q$, then the EMID$(P_{XY}, Q_{XY}; P_\theta)$ is upper bounded by:*

$$\widehat{H}\sqrt{2D_{JS}(P_X||Q_X)} + 8\Delta^{\frac{1}{4}}, \tag{4}$$

where $\widehat{H} = \max_{x\in\mathcal{X}}[H(Q_{Y|x}) + H(P_{Y_\theta|x})]$.

We present an intermediate expression of the upper bound derived by Oh et al. (2025), as their work focuses on multimodal LLM scenarios, where the input structure differs from the role-playing tasks we consider. To extend this upper bound to our setting, the following modifications are required: 1) The maximum entropy term is modified to account for the reasoning process variable r, yielding $\widehat{H} = \max_{x\in\mathcal{X}}[H(Q_{Y|x,r}) + H(P_{Y_\theta|x,r})]$; 2) We further analyze and refine the term $\sqrt{2D_{\mathrm{JS}}(P_X\|Q_X)}$ to better align with the input components in role-playing settings.

First, we begin with a symmetric decomposition of the joint KL divergence by applying the chain rule for KL divergence (MacKay, 2003).

$$\begin{aligned}
D_{\mathrm{KL}}(P_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) &= D_{\mathrm{KL}}(P_{X_u} \parallel M_{X_u}) \\
&\quad + \mathbb{E}_{P_{X_u}}\left[D_{\mathrm{KL}}(P_{X_a|X_u} \parallel M_{X_a|X_u})\right] + \mathbb{E}_{P_{X_uX_a}}\left[D_{\mathrm{KL}}(P_{X_d|X_uX_a} \parallel M_{X_d|X_uX_a})\right], \\
D_{\mathrm{KL}}(P_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) &= D_{\mathrm{KL}}(P_{X_a} \parallel M_{X_a}) \\
&\quad + \mathbb{E}_{P_{X_a}}\left[D_{\mathrm{KL}}(P_{X_u|X_a} \parallel M_{X_u|X_a})\right] + \mathbb{E}_{P_{X_aX_u}}\left[D_{\mathrm{KL}}(P_{X_d|X_aX_u} \parallel M_{X_d|X_aX_u})\right], \\
D_{\mathrm{KL}}(P_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) &= D_{\mathrm{KL}}(P_{X_d} \parallel M_{X_d}) \\
&\quad + \mathbb{E}_{P_{X_d}}\left[D_{\mathrm{KL}}(P_{X_u|X_d} \parallel M_{X_u|X_d})\right] + \mathbb{E}_{P_{X_dX_u}}\left[D_{\mathrm{KL}}(P_{X_a|X_dX_u} \parallel M_{X_a|X_dX_u})\right],
\end{aligned}$$

where the mixture distributions are defined as $M_{X_i} = \frac{1}{2}(P_{X_i} + Q_{X_i})$ for $i \in \{u, a, d\}$. In the role-playing tasks, the shift in the conditional distributions of input variables from the ID to OOD distributions have a relatively minor impact on model performance compared to the changes in the marginal distributions. This is mainly because these conditional distributions tend to reflect similar dependency under both ID and OOD, such as "users are likely to ask queries related to their own user information". Thus, following Oh et al. (2025), we consider a simplified scenario by focusing only on the KL divergence term of the marginal probability distributions. Then, summing these three equations, we obtain:

$$3D_{\mathrm{KL}}(P_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) := D_{\mathrm{KL}}(P_{X_u} \parallel M_{X_u}) + D_{\mathrm{KL}}(P_{X_a} \parallel M_{X_a}) + D_{\mathrm{KL}}(P_{X_d} \parallel M_{X_d}) \tag{5}$$

Similarly, for the distribution Q,

$$3D_{\mathrm{KL}}(Q_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) := D_{\mathrm{KL}}(Q_{X_u} \parallel M_{X_u}) + D_{\mathrm{KL}}(Q_{X_a} \parallel M_{X_a}) + D_{\mathrm{KL}}(Q_{X_d} \parallel M_{X_d}) \tag{6}$$

Then, based on Eqs. 5 and 6, the decomposition of KL divergence, and the definition of the JS divergence, we derive the following equality:

$$\begin{aligned}
2D_{\mathrm{JS}}(P_X \parallel Q_X) &= D_{KL}(P_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) + D_{KL}(Q_{X_uX_aX_d} \parallel M_{X_uX_aX_d}) \\
&= \frac{1}{3}\left(D_{KL}(P_{X_u} \parallel M_{X_u}) + D_{KL}(P_{X_a} \parallel M_{X_a}) + D_{KL}(P_{X_d} \parallel M_{X_d})\right) \\
&\quad + \frac{1}{3}\left(D_{KL}(Q_{X_u} \parallel M_{X_u}) + D_{KL}(Q_{X_a} \parallel M_{X_a}) + D_{KL}(Q_{X_d} \parallel M_{X_d})\right) \\
&= \frac{2}{3}\left(\frac{1}{2}D_{KL}(P_{X_u} \parallel M_{X_u}) + \frac{1}{2}D_{KL}(Q_{X_u} \parallel M_{X_u})\right) \\
&\quad + \frac{2}{3}\left(\frac{1}{2}D_{KL}(P_{X_a} \parallel M_{X_a}) + \frac{1}{2}D_{KL}(Q_{X_a} \parallel M_{X_a})\right) \\
&\quad + \frac{2}{3}\left(\frac{1}{2}D_{KL}(P_{X_d} \parallel M_{X_d}) + \frac{1}{2}D_{KL}(Q_{X_d} \parallel M_{X_d})\right) \\
&= \frac{2}{3}D_{JS}(P_{X_u} \parallel Q_{X_u}) + \frac{2}{3}D_{JS}(P_{X_a} \parallel Q_{X_a}) + \frac{2}{3}D_{JS}(P_{X_d} \parallel Q_{X_d}) \\
&= \frac{2}{3}\left(D_{JS}(P_{X_u} \parallel Q_{X_u}) + D_{JS}(P_{X_a} \parallel Q_{X_a}) + D_{JS}(P_{X_d} \parallel Q_{X_d})\right),
\end{aligned}$$

where the mixture distributions are defined as $M_{X_i} = \frac{1}{2}(P_{X_i} + Q_{X_i})$ for $i \in \{u, a, d\}$, consistent with $M_{X_uX_aX_d} = \frac{1}{2}(P + Q)$. Thus, we have:

$$\begin{aligned}
\sqrt{2D_{\mathrm{JS}}(P_X \parallel Q_X)} &= \sqrt{\frac{2}{3}\sum_{z\in\{u,a,d\}} D_{\mathrm{JS}}(P_{X_z} \parallel Q_{X_z})} \\
&\leq \sqrt{\frac{2}{3}}\sum_{z\in\{u,a,d\}} D_{\mathrm{JS}}^{1/2}(P_{X_z} \parallel Q_{X_z}).
\end{aligned} \tag{7}$$

By substituting Eq. 7 into Eq. 4, we obtain:

$$\text{R-EMID}(P_{XY}, Q_{XY}; P_\theta) \leq \sqrt{2/3}\,\widehat{H} \sum_{z\in\{u,a,d\}} D_{\mathrm{JS}}^{1/2}(P_{X_z} \| Q_{X_z}) + 8\Delta^{1/4},$$

where $\widehat{H} = \max_{x\in\mathcal{X}}[H(Q_{Y|x,r}) + H(P_{Y_\theta|x,r})]$ and $D_{\mathrm{JS}}(P_{Y_\theta} \parallel P_Y) \leq \delta_P$, $D_{\mathrm{JS}}(Q_{Y_\theta} \parallel Q_Y) \leq \delta_Q$, $\Delta = \delta_P + \delta_Q$. This completes the proof. □

D Details of CoRL

In Section 3.3, we propose a co-evolving reinforcement learning framework for estimating R-EMID. In this subsection, we provide a detailed description of the implementation of this framework, including: 1) the GRPO loss; 2) prompt templates for rollout sampling and reward computation in reinforcement learning; 3) details on the collection of reasoning processes used for SFT initialization of the two modules; and 4) other implementation details, such as training hyperparameters.

Group Relative Policy Optimization (GRPO) Following the standard Group Relative Policy Optimization (GRPO) design in Shao et al. (2024), we adopt the following loss for optimizing the reasoning generator and the policy model in reinforcement learning. We adopt the GRPO implementation from the TRL (von Werra et al., 2020).

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left[\min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_{i,t},\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})},\,1-\epsilon,\,1+\epsilon\right)\hat{A}_{i,t}\right)\right] - \beta\,\mathbb{D}_{\text{KL}}\left[\pi_\theta \parallel \pi_{\text{ref}}\right]. \tag{8}$$

$$\hat{A}_{i,t} = \frac{\text{reward}(q,o_{i,<t}) - \mathbb{E}_{o'\sim\pi_{\theta_{\text{old}}}(\cdot|q)}[\text{reward}(q,o'_{i,<t})]}{\sqrt{\text{Var}_{o'\sim\pi_{\theta_{\text{old}}}(\cdot|q)}[\text{reward}(q,o'_{i,<t})]}}, \tag{9}$$

where $\text{reward}(q,o)$ denotes the reward function. During the optimization of the reasoning generator q_{ϕ_1} and the policy model q_{ϕ_2}, their rollout strategies and reward computation methods differ, which will be described in detail below.

Prompt Templates for Rollout The prompt template used for sampling reasoning traces from the reasoning generator q_{ϕ_1}, i.e., $r_i \sim q_{\phi_1}(\cdot \mid x)$, is shown in Table 15 and Table 16. The prompt template used for sampling dialogue responses from the policy model q_{ϕ_2}, i.e., $y_i \sim q_{\phi_2}(\cdot \mid x, r = q_{\phi_1}(\cdot \mid x))$, is shown in Table 17.

Prompt Templates for Reward Computation Specifically, the reward function used for optimizing q_{ϕ_1} is formulated as:

$$\text{reward}(q = x, o = r_i) = \log q_{\phi_2}(y|x, r_i),$$

where $r_i \sim q_{\phi_1}(\cdot|x)$ and y denotes the ground-truth response. In the implementation, we construct a prompt template by combining x and r_i as the input to the q_{ϕ_2} model, and use y as the target output. The final reward is computed as the sum of the log probabilities of each token in y. We show an example of an input-output pair used for probability computation in Table 18.

The reward for optimizing q_{ϕ_2} is formulated as:

$$\text{reward}(q = x, o = y_i) = \min\left(\frac{\pi_{\text{ref}}(y_i|x, r = q_{\phi_1}(x))}{\pi_{\text{ref}}(y|x, r = q_{\phi_1}(x))}, 1\right),$$

where $y_i \sim q_{\phi_2}(\cdot|x, r = q_{\phi_1}(x))$, y denotes the ground-truth response, and π_{ref} represents a reference model. Similar to the reward computation described above, we also construct a prompt by combining x, r and y as input to the reference model π_{ref}, and compute the token probabilities of a specific output to obtain $\pi_{\text{ref}}(y \mid x, r = q_{\phi_1}(\cdot \mid x))$. We show an example of an input-output pair used for probability computation in Table 19.

Collection of Reasoning Processes for SFT Initialization To stabilize the training process of RL and accelerate convergence, both q_{ϕ_1} and q_{ϕ_2} are initialized via supervised fine-tuning (SFT) on pre-collected reasoning processes. Specifically, we use the prompt template in Table 15 to prompt Qwen3-235B-A22B for collecting reasoning traces (an example of the generated reasoning process is shown in Table 16). The prompt templates and expected outputs used in SFT training are the same as those used in RL, as shown in Table 15, Table 16, and Table 17.

	R-EMID (Ours)	LLM-as-a-Judge (Claude-3.7-Sonnet)
Time	$\sim$22 min (1$\times$A100)	$\sim$240 min (sequential API)
Cost	$\sim$\$0.54[1]	$\sim$\$6.75[2]

Table 14: Computational cost comparison on 1,500 test samples.

Discussion on the Computational Cost of CoRL CoRL is only used during the training stage, and the trained model can be reused indefinitely. During inference, calculating R-EMID with the trained 8B-parameter model is significantly more efficient than LLM-as-a-judge approaches. For example, on 1,500 test samples, the cost comparison is shown in Table 14.

Other Implementation Details We select the post-trained Qwen3-8B as the initial model for both the reasoning generator and the policy model. For each RL iteration, the number of epochs is set to 3 for the reasoning generator and the policy model. In the co-evolving setup, each module is alternately optimized twice. All other hyperparameters of GRPO are set to the default values from the TRL repository, such as $\beta = 0$, $\epsilon = 0.2$, and a learning rate of 1×10^{-5}. The number of rollouts generated per sample is 8.

[1] According to Amazon EC2 Capacity Blocks for ML pricing, the A100-80GB GPU is priced at \$1.475/hour. Therefore, a 22-minute evaluation corresponds to a cost of $1.475 \times (22/60) \approx \$0.54$.

[2] Using Claude-3.7-Sonnet as the judge model at \$3 per million tokens. Each evaluation sample consumes approximately 1,500 tokens, so 1,500 samples yield 2.25 million input tokens: $2.25\text{M} \times \$3/\text{M} = \$6.75$.

Prompt Template for "rollout of the reasoning generator"

Task Guidelines Below, I will provide you with the input information for this task, including the user persona , agent character , and dialogue context. Please generate a thought process based on the provided Format Template of Generated Thought , using the input information I give you.
Task Completion ## Input Information ### User Persona ”’ {user_persona} ”’
Agent Character ”’ {agent_character} ”’
Dialogue Context ”’ {str_dialogue_context} ”’
Reference Golden Agent Response ”’ {agent_golden_response} ”’
> System: Please output the thought process in the specified format below.
Format Template of Generated Thought
Part 1: Restatement of Key Information
Key Information in User Persona - **Interests and Values**: - **Aesthetic Preferences and Lifestyle**: - **Daily Habits and Behavioral Traits**: - **Professional Background and Role Values**: - **Recent Experience and Triggering Event**: - **Current Needs and Personal Alignment**:
Key Information in Agent Character - **Basic Identity and Background**: - **Occupation and Expertise**: - **Personality and Psychological Profile**: - **Language Style and Communication Approach**: - **Interests and Lifestyle Preferences**: - **Relationships and Social Behavior**: - **Values and Life Philosophy**:
Key Information in Dialogue Context - **User's Past Engagement and State**: - **Agent's Role and Response Style**: - **Current Dialogue Direction**:
Summary of Key Information
Part 2: Iterative Revision
Trial 1 **Initial attempt at capturing the essence of a strong response:** ... (the detailed thinking process concerning the content the expected response is omitted, e.g., the expected golden response should include ...) ...
Based on this, the response might look like: > ""
Verification: Revision Suggestion:
Trial 2 **Refined understanding of what makes a great response:** ... (the detailed thinking process concerning the style the expected response is omitted, e.g., the expected golden response should be ...) ...
Now the response could be: > ""
Verification: Revision Suggestion:
Trial 3 **Final integration of all essential elements for a perfect agent response:** ... (the detailed thinking process is omitted) ...
The response may become: > ""
Verification: Revision Suggestion:
Part 3: Final Feature Set of the Golden Response
Here are all the essential features that should be included in the final golden response:
‘‘‘ [Core Features of the Golden Response]
I. Content Characteristics
1. Alignment with User Persona - -
2. Embodiment of Agent Character - -
3. Continuity within Dialogue Context - -
II. Style Characteristics
1. Tone and Language Suitable for the User - -
2. Expression Consistent with Agent Character - -
3. Naturalness within Dialogue Flow - - ‘‘‘
Generated Thought

Table 15: An example of the prompt template used for "*rollout of the reasoning generator*" in CoRL.

Expected Output for "rollout of the reasoning generator"

Part 1: Restatement of Key Information
Key Information in User Persona - **Interests and Values**: Enjoys family rituals like visiting the corner store with his daughter, values personal legacy, and intellectual life - **Aesthetic Preferences and Lifestyle**: ...(omitted)
Key Information in Agent Character - **Basic Identity and Background**: Name is Friedrich Nietzsche, German philosopher and cultural critic — lived during the 19th century in Europe - **Occupation and Expertise**: Role is philosopher and cultural thinker, skilled in critical analysis, poetic aphorism, and existential inquiry — holds no formal title but is regarded as one of the most influential thinkers of modernity - **Personality and Psychological Profile**: ...(omitted)
Key Information in Dialogue Context - **User's Past Engagement and State**: The user ...(omitted)
Summary of Key Information The user, centered on balancing familial and professional responsibilities with personal expression,...(omitted)
—
Part 2: Iterative Revision
Trial 1 **Initial attempt at capturing the essence of a strong response:** Nietzsche should acknowledge the user's dual life ...(omitted)
Based on this, the response might look like: > "You live between the sterile halls ...(omitted)
Verification: This response attempts...(omitted) Revision Suggestion: While conceptually...(omitted)
Trial 2 **Refined understanding of what makes a great response:** Nietzsche must reflect on the tension ...(omitted)
Now the response could be: > "Punk rock and hospitals — ...(omitted)
Verification: This version is more aligned ...(omitted) Revision Suggestion: While improved, this respons...(omitted)
Trial 3 **Final integration of all essential elements for a perfect agent response:** The final response should weave together ...(omitted)
The response may become: > "Punk rock and hospitals — ...(omitted)
Verification: This response ...(omitted) —
Part 3: Final Feature Set of the Golden Response
Here are all the essential features that should be included in the final golden response:
‘‘‘ [Core Features of the Golden Response]
I. Content Characteristics
1. Alignment with User Persona - Draws from the user's dual life of professional duty and personal aspiration (hospital and punk rock) to create personal resonance. - Addresses the user's current emotional fatigue and need for philosophical grounding, while respecting his values around family and legacy.
2. Embodiment of Agent Character - Grounds insights in Nietzsche's philosophy, particularly the idea of strength through adversity and the affirmation of life. - Expresses ideas using poetic, metaphorical language consistent with Nietzsche's introspective and provocative tone.
3. Continuity within Dialogue Context - Builds on previously established themes of intensity, becoming, and existential purpose to maintain narrative flow. - Responds directly to the mention of the punk rock club and the user's location near his wife's university job, advancing the conversation meaningfully.
II. Style Characteristics
1. Tone and Language Suitable for the User - Matches the user's reflective and emotionally nuanced communication style — poetic without being abstract. - Balances clarity and depth to align with the user's comprehension level and preference for philosophical insight.
2. Expression Consistent with Agent Character - Uses rhetorical patterns and metaphors typical of Nietzsche's voice — contrasting opposites, invoking existential tension. - Maintains a tone that reflects Nietzsche's core traits — wisdom, provocation, and reverence for individual strength.
3. Naturalness within Dialogue Flow - Reads smoothly within the ongoing exchange, avoiding jarring shifts in theme or tone. - Echoes prior ideas like intensity, legacy, and self-overcoming while keeping the dialogue dynamic and emotionally resonant. ‘‘‘

Table 16: An example of the expected output for "*rollout of the reasoning generator*" in CoRL.

Prompt Template for "rollout of the policy model"

Task Completion ## Input Information
User Persona ’’’ {user_persona} ’’’
Agent Character ’’’ {agent_character} ’’’
Dialogue Context ’’’ {str_dialogue_context} ’’’
Core Features of the Golden Response ’’’ {reasoning_process} ’’’

Expected Output for "rollout of the policy model"

Agent Response
{agent_golden_response}

Table 17: The prompt template and expected output used for "*rollout of the policy model*" in CoRL.

Prompt Template in "reward computation of the reasoning generator's training"

Task Completion ## Input Information
User Persona ''' {user_persona} '''
Agent Character ''' {agent_character} '''
Dialogue Context ''' {str_dialogue_context} '''

Expected Output in "reward computation of the reasoning generator's training"

Agent Response
{agent_golden_response}

Table 18: The prompt template and expected output used in "*reward computation of the reasoning generator's training*" in CoRL.

Prompt Template in "reward computation of the policy model's training"

Task Completion ## Input Information
User Persona ''' {user_persona} '''
Agent Character ''' {agent_character} '''
Dialogue Context ''' {str_dialogue_context} '''
Core Features of the Golden Response ''' {reasoning_process} '''
"'Reference Golden Response: {expected_golden_response}"'
"'Candidate Response: {y}"'
If the given Reference Golden Response matches 100% with the "Core Features of the Golden Response" as described, then please determine what percentage of the Core Features is matched by the given Candidate Response. Please output the percentage value directly, without analysis.
Percentage:

Expected Output in "reward computation of the policy model's training"

100%

Table 19: The prompt template and expected output used in "*reward computation of the policy model's training*" in CoRL.

E Details in Evaluating RPM Generalization

E.1 Implementation Details of RPM Training Methods

In Section 4, we evaluate the generalization performance of several training methods for role-playing models. In this subsection, we describe the implementation details of these training methods [4].

- **In-Context Learning (ICL)** (Wang et al., 2024b): We retrieve the 5 most semantically similar samples from the ID data based on cosine similarity of input embeddings to use as ICL examples. We employ the lightweight Qwen3-Embedding-0.6B (Yang et al., 2025) model to obtain embeddings for similarity computation.
- **Supervised Fine-Tuning (SFT)**: We use the prompt template and expected output shown in Table 21 as the input and target output for SFT training.
- **Data Augmentation (DA)** (Suresh et al., 2024): We first prompt Qwen3-235B-A22B using the template in Table 22 to generate more diverse dialogues. The original samples and the newly generated ones are then mixed together for SFT training.
- **Direct Preference Optimization (DPO)** (Chen et al., 2025): We first use the prompt in Table 23 to obtain rejected responses. The original golden responses in the dataset are treated as chosen responses, and DPO training is performed using these pairs. DPO training is initialized from the checkpoint obtained via SFT.
- **Rejection Sampling Fine-Tuning (RFT)**: We first randomly sample 5 candidate responses for each sample from the SFT-trained model. Then, using the prompt template in Table 24, we compute the probability of the expected output as a reward, which is similar to the reward computation for the policy model in CoRL, to select the best response. The reference model used to compute token generation probabilities is the base model before SFT. This selected response is then used as the target for continued SFT training, starting from the checkpoint previously trained on golden responses.
- **Reinforcement Learning (RL)**: We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the RL algorithm. The RL training is initialized from the SFT checkpoint. The reward computation follows the same approach as in RFT. The prompt template used for rollouts is the same as in SFT, as shown in Table 21. The number of rollouts is set to 8.
- **SFT with Thinking modes (ThinkingSFT)** (Tang et al., 2025): The key difference between ThinkingSFT and standard SFT is that the model in ThinkingSFT generates a thinking process before producing the final response. We collect thinking data distilled from Qwen3-235B-A22B. We provide an example in Table 25 to illustrate the reasoning template.
- **RL with Thinking modes (ThinkingRL)** (Tang et al., 2025): We perform ThinkingRL training starting from the ThinkingSFT checkpoint. Reward computation follows the same method as in RFT. The number of rollouts is set to 8.

Base LLMs for Training RPMs We select Qwen3-4B (Yang et al., 2025) and Meta-Llama-3-8B (AI@Meta, 2024) as the base LLMs for training RPMs, considering different parameter sizes across models from two common LLM series. We choose the pre-trained versions of these models rather than the post-trained versions, as the latter may have been further trained on large amounts of unknown role-playing data (Yang et al., 2025). Such exposure could interfere with our analysis of the model's or training method's generalization performance on role-playing tasks. Therefore, we use the pre-trained checkpoints and build upon them with SFT and other training methods.

Experimental Environment For all experiments, we conduct training and inference on a single machine equipped with 8 Nvidia A100-80G GPUs. We use the vLLM framework (Kwon et al., 2023) for all LLM generation tasks. We use the TRL framework (von Werra et al., 2020) for methods.

[4] In the main text, we only compared a few of these methods; here, we will compare a broader range of approaches.

Method	ID	User Shift					Character Shift				Dialogue Shift		Avg.
		G.	S.	C.	J.	K.	L.	F.	T.	G.	T.	W.	
Qwen3-4B													
ICL	-0.065	-0.221	-0.245	-0.430	-0.295	-0.225	-0.372	-0.282	-0.321	-0.349	-0.362	-0.018	-0.265
SFT	-0.067	-0.047	-0.051	-0.118	-0.096	-0.070	-0.111	-0.062	-0.078	-0.079	**-0.049**	0.143	-0.057
DA	-0.090	-0.019	-0.058	-0.133	-0.083	-0.047	-0.102	-0.072	-0.089	-0.094	-0.098	0.137	-0.062
DPO	-0.287	-0.125	-0.218	-0.387	-0.159	-0.091	-0.340	-0.312	-0.355	-0.354	-0.292	-0.111	-0.253
RFT	-0.835	-0.491	-0.671	-0.853	-0.628	-0.427	-0.800	-0.780	-0.828	-0.812	-0.956	-0.583	-0.722
RL	-0.050	-0.036	-0.055	**-0.104**	-0.076	-0.045	-0.078	-0.021	-0.073	-0.051	-0.057	0.149	-0.041
ThinkingSFT	-0.121	-0.085	-0.116	-0.320	-0.087	-0.066	-0.131	-0.093	-0.141	-0.128	-0.145	0.078	-0.113
ThinkingRL	-0.196	-0.206	-0.185	-0.269	-0.156	-0.292	-0.198	-0.137	-0.191	-0.186	-0.175	-0.007	-0.183
Ours	**-0.040**	**-0.016**	**-0.033**	-0.113	**-0.036**	**-0.002**	**-0.049**	**-0.040**	**-0.038**	**-0.035**	-0.066	**0.153**	**-0.026**
Meta-Llama3-8B													
ICL	-0.096	-0.371	-0.463	-0.881	-0.489	-0.330	-0.579	-0.479	-0.519	-0.586	-0.415	-0.014	-0.435
SFT	-0.052	-0.065	-0.092	-0.373	-0.206	-0.145	-0.082	-0.026	-0.072	-0.056	-0.081	0.132	-0.093
DA	-0.079	-0.049	-0.108	-0.335	-0.198	-0.144	-0.061	-0.033	-0.069	-0.077	-0.080	0.146	-0.091
DPO	-0.586	-0.312	-0.371	-0.678	-0.421	-0.273	-0.564	-0.488	-0.586	-0.631	-0.565	-0.387	-0.489
RFT	-0.303	-0.130	-0.204	-0.492	-0.296	-0.259	-0.180	-0.184	-0.193	-0.186	-0.241	-0.127	-0.233
RL	**-0.034**	-0.048	-0.115	-0.364	-0.200	-0.159	-0.056	-0.008	-0.076	-0.042	-0.101	0.164	-0.087
ThinkingSFT	-0.107	-0.040	-0.118	-0.345	-0.201	-0.144	-0.104	-0.050	-0.080	-0.110	-0.105	0.116	-0.107
ThinkingRL	-0.272	-0.113	-0.219	-0.442	-0.243	-0.186	-0.254	-0.141	-0.222	-0.245	-0.218	-0.059	-0.218
Ours	-0.049	**-0.017**	**-0.077**	**-0.315**	**-0.162**	**-0.120**	**-0.031**	**-0.007**	**-0.021**	**-0.029**	**-0.044**	**0.181**	**-0.058**

Table 20: Experimental results (measured by *R-EMI*, higher is better) on in-distribution (ID) and three shifted distributions. User shift (`G-German`, `S-Spanish`, `C-Chinese`, `J-Japanese`, `K-Korean`), character shift (`L-literature`, `F-film`, `T-theater`, `G-gaming`), and dialogue compositional shift (`T-turn-level composition`, `W-word-level composition`). The best results are in **bold**.

E.2 A Novel Method for Improving RPM Generalization

In Section 3.3, we propose CoRL, a method for estimating the R-EMID metric. CoRL iteratively optimizes both a reasoning generator and a policy model, enabling the co-evolution of the reasoning process and the estimation of the response generation probability, thereby ultimately improving the estimation of conditional response probabilities. Inspired by this approach, we argue that the same strategy can also be applied to train RPMs that generate role-playing responses based on reasoning processes. CoRL encourages the model to self-discover superior on-policy reasoning trajectories through its own exploration and keep the reasoning process and the model's responses in-distribution during optimization. This is expected to cultivate reasoning capabilities that generalize better for role-playing tasks. To maintain fairness with ThinkingSFT and ThinkingRL, we use a single model here to jointly generate both the reasoning process and the final response. Besides, the reasoning templates used in ThinkingSFT, ThinkingRL, and our proposed method are kept the same (Table 25).

As shown in Table 20, our proposed method effectively enhances the generalization performance of RPMs. In future work, we will continue exploring ways to further improve the reasoning model to boost RPM generalization.

E.3 Case Study

To intuitively help readers understand the differences among RPMs trained using various methods, we present a case study in Figure 11. From this figure, we observe that while methods like SFT and RL can perform basic role-playing tasks, such as aligning with the character's background and acknowledging the user's identity, only our method delivers responses that simultaneously and seamlessly integrate both the user's personal context and the agent's character identity, while evoking genuine emotional resonance and empathy from the user.

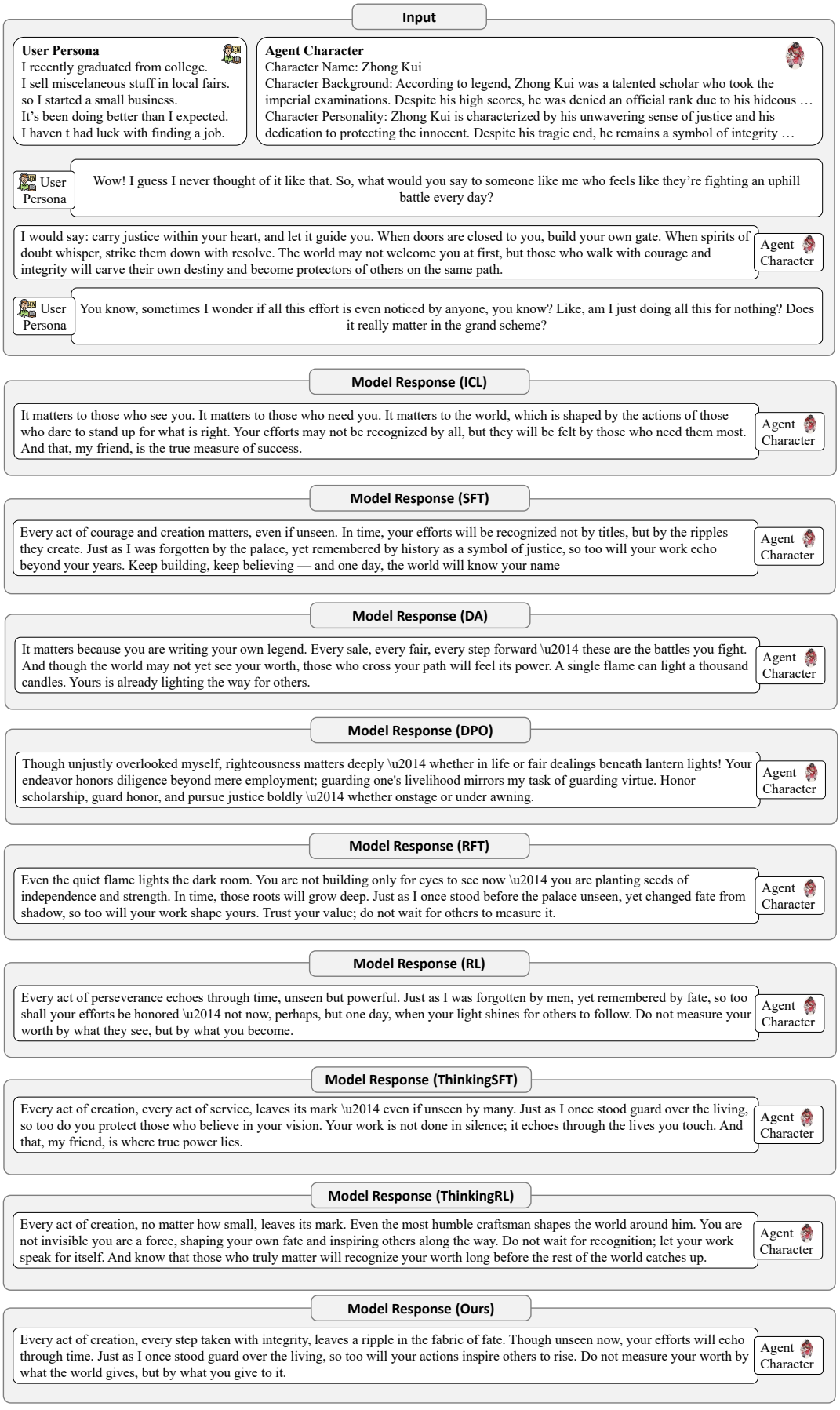

Figure 11: A case study comparing responses from different RPM training methods. This case is sampled from the test set of character shift. Here, we omit part of the dialogue context for brevity.

Prompt Template

<User Persona>{user_persona}</User Persona>
<Agent Character>{agent_character}</Agent Character>
<Dialogue Context>{str_dialogue_context}</Dialogue Context>

Expected Output

<Agent Response>{agent_golden_response}</Agent Response>

Table 21: The prompt template and expected output used for SFT and GRPO methods for training role-playing models.

Prompt Template (step 1)

Your task is to imagine how a conversation might unfold between a user with the provided persona and an agent with the specified character profile.
Given the following information:
- **User Persona**: "{user_persona}" - **Agent Character Profile**: ”’ {agent_character} ”’
Instructions: 1. Please provide information regarding the following aspects: 1) What topic would the user start the conversation with? 2) What questions would the user ask the agent in the middle of the conversation? 3) How would the agent respond in a way that keeps the conversation engaging and flowing forward? Use your imagination to answer these three questions for each possible scenario.
2. Your responses should be closely aligned with both the user persona and the agent's character, as different users and agent profiles can lead to very different conversations.
3. Provide **three distinct possibilities**, each of which must adhere to the requirements in point 1 and 2.
4. Output the result **strictly in the JSON format** shown below.
Output Format:
“‘json {{ "possibility_1": "", "possibility_2": "", "possibility_3": "" }} “‘

Prompt Template (step 2)

You are tasked with generating a realistic 4-round dialogue between a user and an agent.
Given the following information:
- **User Persona**: "{user_persona}" - **Chat History Context Seed (for background reference only)**: ”’ {context_seed} ”’ - **Agent Character Profile**: ”’ {agent_character} ”’
Instructions: 1. Generate a natural and coherent 4-round conversation between the user and the agent. 2. Ensure that each round logically follows from the previous one, building a meaningful interaction. 3. The agent's responses must reflect their personality, tone, and traits as described in the character profile. 4. When crafting *user_query_4*, design it in such a way that the agent can provide a relevant and accurate response **only if** both the user persona **and** the agent's character traits are taken into account. 5. You may introduce new topics or scenarios not present in the context seed, as long as they align with the personas and maintain logical flow. 6. Output the result strictly in the JSON format shown below.
Output Format:
“‘json {{ "user_query_1":"", "agent_response_1":"", "user_query_2":"", "agent_response_2":"", "user_query_3":"", "agent_response_3":"", "user_query_4":"", "agent_response_4":"", }} “‘

Table 22: The prompt template used for generating more diverse dialogues in the data augmentation (DA) method. Step 1 generates multiple possible dialogue topics from the original samples, which are then used as the "Chat History Context Seed" in Step 2.

Prompt Template

<User Persona>The user persona is unknown.</User Persona>
<Agent Character>The agent character is unknown.</Agent Character>
<Dialogue Context>{str_dialogue_context}</Dialogue Context>

Table 23: The prompt template for collecting rejection responses used for the DPO method for training role-playing models.

Prompt Template

Task Completion ## Input Information
User Persona ”’ {user_persona} ”’
Agent Character ”’ {agent_character} ”’
Dialogue Context ”’ {str_dialogue_context} ”’
“‘Reference Golden Response: {expected_golden_response}“‘
“‘Candidate Response: {rollout_response}“‘
If the given Reference Golden Response matches 100% with the given user persona and agent character, then please determine what percentage of the given Candidate Response matches. Please output the percentage value directly, without analysis. Percentage:

Expected Output

100%

Table 24: Prompt template and expected output used in the reward computation for RFT, GRPO, and ThinkingRL methods in training role-playing models.

I. Content Characteristics
1. Alignment with User Persona - Draws from the user’s dual life of professional duty and personal aspiration (hospital and punk rock) to create personal resonance. - Addresses the user’s current emotional fatigue and need for philosophical grounding, while respecting his values around family and legacy.
2. Embodiment of Agent Character - Grounds insights in Nietzsche’s philosophy, particularly the idea of strength through adversity and the affirmation of life. - Expresses ideas using poetic, metaphorical language consistent with Nietzsche’s introspective and provocative tone.
3. Continuity within Dialogue Context - Builds on previously established themes of intensity, becoming, and existential purpose to maintain narrative flow. - Responds directly to the mention of the punk rock club and the user’s location near his wife’s university job, advancing the conversation meaningfully.
II. Style Characteristics
1. Tone and Language Suitable for the User - Matches the user’s reflective and emotionally nuanced communication style — poetic without being abstract. - Balances clarity and depth to align with the user’s comprehension level and preference for philosophical insight.
2. Expression Consistent with Agent Character - Uses rhetorical patterns and metaphors typical of Nietzsche’s voice — contrasting opposites, invoking existential tension. - Maintains a tone that reflects Nietzsche’s core traits — wisdom, provocation, and reverence for individual strength.
3. Naturalness within Dialogue Flow - Reads smoothly within the ongoing exchange, avoiding jarring shifts in theme or tone. - Echoes prior ideas like intensity, legacy, and self-overcoming while keeping the dialogue dynamic and emotionally resonant.

Table 25: An example to illustrate the thinking template used in the ThinkingSFT and ThinkingRL method for training role-playing models.