What Drives Long-term User Engagement?
A Large-scale Analysis of Role-playing Dialogues in the Wild

Anonymous ACL submission

Abstract

The growing sophistication of dialogue agents has led to extended human-AI conversations spanning days or months, yet our understanding of what sustains user engagement comes primarily from short-term simulations. Through large-scale analysis of real-world interactions between 37,000+ users and 8,000+ characters, we systematically evaluate nine engagement factors via rigorous A/B testing with 53 model pairs. Our findings challenge prevailing assumptions: while character embodiment shows limited influence, response length and non-verbal descriptions emerge as critical factors. We also find that human-likeness and personality consistency significantly impact engagement, while factors like lexical diversity show minimal effect. These insights, derived from real user interactions, provide actionable guidance for developing more engaging role-playing models by prioritizing response depth over strict role adherence.

1 Introduction

As dialogue agents continue to advance in their humanlike capabilities, individuals are increasingly engaging in extended conversations with these agents that can span hours, days, or even months (Irvine et al., 2023; Buse et al., 2023). While this advancement presents unprecedented opportunities for sustainable human-AI engagement(Lee et al., 2022), it also introduces unique challenges that existing evaluation frameworks fail to address.

Current research predominantly relies on short-term simulations and third-party evaluations(Wang et al., 2023b; Zhou et al., 2023a; Tu et al., 2024; Maas et al., 2023; Chen et al., 2023), creating a critical gap between laboratory findings and real-world user experiences. Three key limitations persist in existing work: Firstly, evaluation methods typically focus on immediate interaction quality rather than sustained engagement. Secondly, third-party assessments often fail to capture authentic user preferences. Thirdly, the factors driving long-term user engagement remain poorly understood.

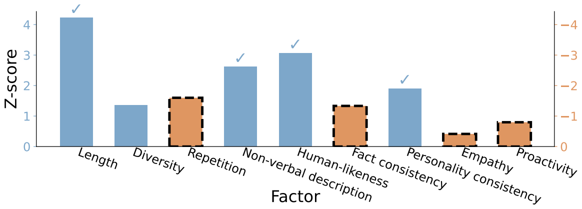


Figure 1: Significance of each factor for user engagement. Blue bars for positive, orange dashed for negative impact. Checks indicate statistical significance ($p < 0.05$). Interestingly, bot's length shows a notable effect while role's fact consistency has a limited influence.

.

Our work addresses these limitations through systematic analysis of real-world interactions between 37,000+ users and 8,000+ characters. By examining 9 factors with rigorous A/B testing, we provide empirical evidence for what truly drives user engagement in extended conversations. Surprisingly, we find that the degree to which the bot embodies the roles it plays has limited influence on user engagement. Instead, the bot's utterance length crucially maintains user interest (Fig. 1).

The contributions of this study are threefold: Firstly, we have developed a framework to amass authentic conversational data alongside behavioral metrics, facilitating the exploration of factors that influence long-term user engagement in real-world role-playing dialogues. Secondly, We introduce an approach for pinpointing significant factors that enhance user engagement. Thirdly, we empirically elucidate the reasons why these factors are instrumental in affecting user engagement. Collectively, our paper illuminates the integral components of user engagement in sustained conversations with role-playing entities in authentic settings. These insights provide indispensable direction for researchers and developers aiming to augment the efficacy and enduring appeal of role-playing chatbots, thereby establishing a cornerstone for future innovations in the domain.

2 Related work

2.1 Role-playing Agent

Role-playing agents are LLMs imbued with a specific role, aiming at enhancing their task-solving prowess (Kong et al., 2023; Xu et al., 2023), fostering human-like social intelligence (Zhou et al., 2023b; Chen et al., 2024), and engaging users as a dialogue agent (Li et al., 2023; Wang et al., 2023b). LLMs can be construed as superpositions of various roles, each encompassing unique beliefs (Kovač et al., 2023) and personas (Lu et al., 2024; Safdari et al., 2023). These roles can be induced to manifest either broad demographic characteristics (Rao et al., 2023) or more fine-grained profiles (Wang et al., 2023b; Zhou et al., 2023a).

2.2 Role-playing Evaluation

As dialogue agents, role-playing models are evaluated not only on dimensions associated with open-domain dialogue, such as fluency, informativeness, and context consistency (Huang et al., 2020; Miehling et al., 2024), but also on their unique aspect of persona fidelity. This encompasses the personality (Wang et al., 2023a; Li et al., 2023), knowledge (Tu et al., 2024; Shen et al., 2023),decisions (Xu et al., 2024), and motivations (Yuan et al., 2024) inherent to a role. However, current evaluation methodologies are limited due to their reliance on non-real data and their inherent third-person perspective (Miehling et al., 2024). Specifically, they typically use materials like movies, novels, scripts, or Wikipedia entries (Li et al., 2023; Tu et al., 2024; Salemi et al., 2023), synthetic data (Wang et al., 2023b), or limited annotator interactions (Zhou et al., 2023a; Shen et al., 2023). These evaluations often focus on the model's response alignment with predefined labels or a third-party perspective (Wang et al., 2023b; Shen et al., 2023; Li et al., 2023), which may not reflect real user needs. In contrast, our evaluation is grounded in real user interactions with the role-playing model. We focus on user engagement, offering a first-person perspective on model performance.

Similar dilemmas are encountered in non-goal oriented dialogue(See et al., 2019; Finch et al., 2023). The Alexa Prize competition (Khatri et al., 2018; Gottardi et al., 2022; Ram et al., 2018) aims to address this issue by setting up a contest where teams are tasked with building a dialogue model that can engage uninformed real users on a variety of topics. Evaluations are based on ratings and free-form feedback from the user's first-person perspective, gauging user satisfaction and engagement. Subsequent studies have been conducted to analyze and build interpretable measurements quantified by dialogue, with the goal of aligning with real user satisfaction and engagement in a dialogue (Guo et al., 2018; Venkatesh et al., 2018). Our work aligns with these efforts, as we analyze factors related to dialogue that influence user engagement. However, our focus is on the role-playing scenario and the evaluation of a model than a dialogue.

3 Experimental Design

3.1 Role-Play Platform for Real-World Interaction Data Collection

This study examines factors that affect long-term user engagement in real-world role-playing conversations. For accurate analysis, a large dataset comprising substantial real-user interactions and first-person conversational experiences is crucial. To this end, we developed an online role-playing platform where users interact daily with a variety of characters (see Appendix A, Fig. 9). Characters may be system-generated or user-created (see Fig. 10 in Appendix A), each with a distinct set of personas, including basic information, personality traits, and experiences (see Fig. 11 in Appendix A).

The platform records user interactions and tracks behaviors, particularly their return to the platform and the continuity of their interactions. Data collection is bifurcated: one part of the dataset monitors whether users return to engage in conversations on subsequent days, which acts as a barometer for long-term user engagement, given that this behavior occurs after initial long-term conversation experiences; the other part documents daily dialogues between users and characters, shedding light on potential factors influencing engagement. Notably, ongoing user interaction over time is indicative of a positive engagement experience. In contrast, discontinuation suggests a less satisfactory experience. Therefore, we define the user retention rate as the percentage of users who revisit the platform the day following their initial interaction, serving as an index of user engagement. By tracking the daily user retention rate, we obtain a direct indicator of user engagement.

3.2 A/B Testing Framework for Bias Reduction

Leveraging spontaneous user behavior as a proxy for engagement presents a scalable alternative to the cost-intensive approach of soliciting direct user feedback. Nevertheless, such indirect measures are potentially susceptible to confounding influences that are not indicative of the chatbot's intrinsic efficacy, such as seasonal variations (e.g., holidays) or alterations in the system (e.g., interface modifications). To mitigate these extrinsic biases, we have instituted controlled experimental conditions that are consistent across all external factors.

To achieve this objective, an A/B testing infrastructure has been deployed, enabling simultaneous operation of distinct models within the same environment. This setup isolates the effect of the model from other variables by facilitating a comparative analysis of user retention rates between the models. Furthermore, at the point of registration, each user is randomly allocated to one of the A/B testing models by the system. This allocation is maintained throughout the duration of the experiment to avert potential contamination effects and to ensure a uniform experience with the respective model. Notably, the randomization inherent in the assignment process approximately equalizes the distribution of user demographic characteristics and their preferences for role-playing characters across the models (check Section 5.1 for results), thereby reducing selection bias in the comparative evaluation. Consequently, differences in user engagement attributable to model performance can be clearly seen by the examination of daily user retention rates within the A/B testing paradigm.

4 Methods

We aim to identify the factors within role-playing models that influence user engagement. To achieve this, we analyze interaction data between a large number of real users and multiple pairs of role-playing models, where each pair exhibits distinct user engagement based on user retention rate.

We begin by selecting pairs of models from the A/B testing pool, focusing on those with a significant user base. These pairs are distinguished by their markedly different levels of user engagement, which allows us to classify the models as either "strong model" or "weak model" (Fig. 2a). Following this, we randomly select dialogue data from these models to investigate potential influences on user engagement (Fig. 2b). We then enumerate potential factors within the role-playing models that could affect user engagement, quantifying these factors using the sampled dialogue data for each model (Fig. 2c). Finally, we identify the influential factors by analyzing whether their values consistently and significantly differ across the selected pairs of models in a statistically meaningful way (Fig. 2d). Further details are provided in Section 5.

4.1 Models

The models under investigation primarily exhibit differences in their foundational architectures, encompassing variants such as LLaMA (Touvron et al., 2023a,b) and Mistral (Jiang et al., 2023). Variations also manifest in the scope of continuous pretraining and fine-tuning. Pretraining datasets incorporate live chat data from VTuber [1] and publicly accessible chat data from ShareGPT [2], in addition to combinations of these sources. For fine-tuning purposes, datasets such as PIPPA(Tear Gosling and Zheng, 2023) and proprietary collections, which are curated and refined from the platform, are employed. Techniques including SFT, DPO(Rafailov et al., 2024), and related variants are applied, utilizing a consistent prompt template across all models.

4.2 Factors

Together, we thoroughly analyze 9 factors that potentially impact user engagement. These factors are derived from extensive research conducted by Wang et al. (2023b) and Shen et al. (2023), supplemented by our own insights gained from reviewing the collected dialogue samples. We carefully exclude factors that are overly composite, such as "interestingness", as well as factors like "fluency", which are now considered standard across recent models [3]. The factors under evaluation include elements based on content, semantics, and interaction skills (refer to Table 1). For a detailed explanation of these factors and the reasoning behind their selection, please consult Appendix B.1.

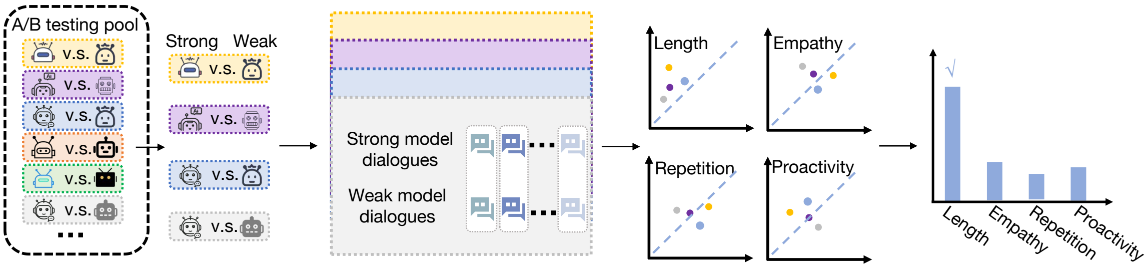
For the first four factors, we employ straightforward methods to determine their scores. However,

---

[1] https://github.com/sigvt/vtuber-livechat-dataset

[2] https://huggingface.co/datasets/RyokoAI/ShareGPT52K

[3] We do not focus on topic as a factor, as it is shaped by the role and user, with consistent distribution across models in our A/B testing. Additionally, we do not conduct fine-grained analyses of the selected factors due to data limitations.



(a) Strong and weak model pair selection
(b) Dialogue collection for each model pair
(c) Feature score calculation per factor per model pair
(d) Siginificance score for each facotor

Figure 2: Our method follows a pipeline that begins by identifying pairs of (strong and weak) models with significantly different retention rates through A/B testing pool (for model details, see Section 4.1). From these pairs, we then sample dialogue examples and utilize them to calculate scores for various factors (for factor details, see Section 4.2). Finally, we quantify the impact of these factors and their significance by analyzing the calculated scores (for significance score calculation see Section 4.3).

Category	**Factor**	**Description**
Content	Length	How long the model's responses are.
	Diversity	What variety of words the model uses.
	Non-verbal Description	How frequent the model's non-verbal actions or reactions are, such as gestures or facial expressions.
Semantic	Repetition	Whether the model reiterates similar ideas repeatedly.
	Human-likeness	How much the model acts like a human, showing personal likes, views, and opinions.
	Fact Consistency	Whether the model's statements match the backstory and facts of the character it's portraying.
	Personality Consistency	How well the model's responses fit the character's personality, including their habits, views, and values.
Skill	Empathy	How well the model understands and responds to the user's emotions in a supportive way.
	Proactivity	How proactive the model is in engaging conversations, such as seeking clarification or user preferences.

Table 1: Factors investigated in our study, categorized as content-based, semantic, and skill-based. For a detailed explanation and reasoning, see Appendix B.1.

for factors like human-likeness, we leverage the advanced capabilities of GPT-4 to evaluate whether the models display behavior akin to that of humans. The prompts used, and their alignment with human judgment, can be found in Appendix C. As it is expensive to call GPT-4 API, we segment the dialogues into consecutive slices, each containing five turns. We randomly sample a subset of $M = 100$ slices for each GPT-4 evaluated factors. Details of the factor calculations are in Appendix B.2.

4.3 Significance Test

To assess the impact of each factor on user engagement, we compare the scores between strong and weak models for all pairs. Factors showing significant score differences are considered influential for user engagement. Notably, while the t-test—a method that presupposes data normality—is frequently used for comparing two independent groups, we opted for a non-parametric testing approach within our A/B testing framework. This decision is made due to its superior suitability for the analysis of our paired data.

Mean Score Difference: We calculate the score difference $d_i = s_i - w_i$ for each factor in both strong and weak models, where s_i and w_i represent the scores of the factor for the i-th group's strong and weak models, respectively. The mean score difference $d = \frac{1}{n} \sum_{i=1}^{n} d_i$ across all groups quantifies the average difference of the factor. Intuitively, a higher positive (negative) mean difference indicates a greater proportion of groups with strong models having a positive (negative) difference compared to weak models.

Significance of the Difference: We use permutation tests to assess the significance of the difference. By randomly swapping scores between strong and weak models, we calculate a new average difference score $\hat{d} = \frac{1}{n}\sum_{i=1}^{n}\hat{d}_i$, where $\hat{d}_i = (-1)^{p_i} d_i$ and $p_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5)$. According to the Lyapunov Central Limit Theorem, with a large enough sample of model pairs, $\hat{d}$ follows a normal distribution, $\hat{d} \sim \mathcal{N}(\mu, \sigma^2)$. The z-score is then computed as $z = \frac{d-\mu}{\sigma} \sim \mathcal{N}(0, 1)$, where a positive z-score suggests a positive influence on user engagement and a negative z-score suggests the opposite. The significance is determined by the tail probability of the observed z-score under the normal distribution. Intuitively, a larger positive or negative mean difference score results in a higher or lower ranking among all permuted scores, which corresponds to a larger positive or negative z-score, a smaller p-value, and a more significant influence on user engagement. A p-value below 0.05 is typically considered statistically significant. We perform 100,000 random swaps to obtain 100,000 $\hat{d}$ values, estimating μ and σ by their respective mean and standard deviation.

5 Experiment

5.1 Setup

Our study is anchored in a role-playing platform where users can interact with characters that are either publicly developed or privately created across various categories. Users have the flexibility to initiate a new dialogue or continue an existing one at any time. In the context of A/B testing, users are randomly assigned a test model, which is applied uniformly to all characters they engage with once they are selected as a test participant.

We select 53 pairs of strong and weak models from our A/B testing pools on the platform over four months, from July to November 2023. Each model has a minimum of 140 new users per testing day and was tested for at least a week. To qualify as a valid pair, the strong model's user retention rate has to surpass the weak model's for at least six out of the seven testing days. These models are trained by researchers who actively contribute to a publicly accessible platform for role-play conversation.

We analyze the impact of model factors on user engagement by randomly sampling $N = 1000$ dialogues from each model. The statistics of these conversations are shown in Table 2. Our dataset includes over 37,000 users and more than 8,000 characters, with a diverse distribution of both users and characters, as illustrated in Fig. 3. It is imperative to highlight that the demographic distribution of users between strong models strong and weak models is approximately equivalent, as detailed in Fig. 6 and 7 located in Appendix A. Similarly, the distributions within character categories are closely matched, as shown in Fig. 8 in Appendix A. These measures have been meticulously taken to minimize biases related to users and characters, thereby strengthening the validity of our analysis.

Category	Total	User	Character
#Dialogues	106,000	-	-
#Speakers	45,867	37,752	8,115
#Avg. utterance per speaker	11.47	11.03	11.91
#Avg. length per utterance	37.90	10.86	62.39

Table 2: Statistics of the sampled conversation.

5.2 Overall Results

Fig. 4 shows the comparison of score of each factor between the strong and weak models for all pairs. Fig. 1 shows the significance of each factor's influence on user engagement.

From Fig. 4 and 1, we observe that only length, non-verbal description, human-likeness, and personality consistency have a significant impact on user engagement (p-value<0.05). Conversely, diversity, repetition, fact consistency, empathy, and proactivity do not appear to influence user engagement. Our results reveal significant disparities between traditional engagement metrics and actual user behavior. While previous work emphasizes role consistency and lexical diversity, these factors show surprisingly weak correlation with long-term engagement in the real world role-playing conversation. Most notably, response length - often considered a potential negative factor in short-term evaluations - emerges as the strongest predictor of sustained engagement. This finding directly challenges the common assumption that shorter, more concise responses are preferable. In the subsequent section, we will provide a detailed analysis of each individual factor.

5.3 Factor Analysis

Length. Fig. 4 clearly demonstrates that the average utterance length has the most substantial impact

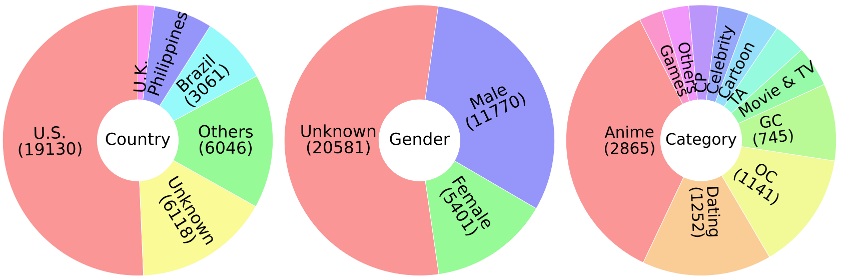


Figure 3: **Left:** user distribution across 200+ countries, top contributors - United States (50.7%), Brazil (8.1%), Philippines (7.0%). **Middle:** gender distribution, males (31.2%), females (14.3%). **Right:** character categories including OC (original character), GC (game character), TA (text adventure), CP (companionship), most prevalent - anime characters (35.3%), dating scenarios (15.4%), original characters (14.1%). Categories <2% are grouped as "Others". "Unknown" represents undisclosed location/gender in first two charts.

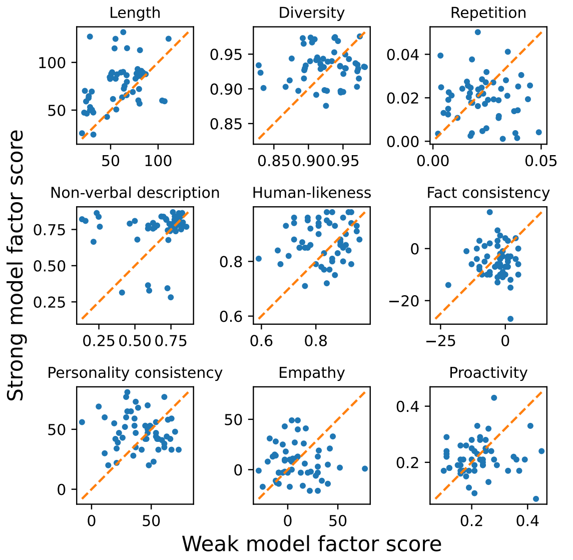


Figure 4: Grid of sub-figures compares factor scores between strong and weak models. Each point represents a model pair, x-axis for weak model scores, y-axis for strong. Dotted line signifies identical scores. Points towards top-left (bottom-right) imply stronger positive (negative) impact on user engagement.

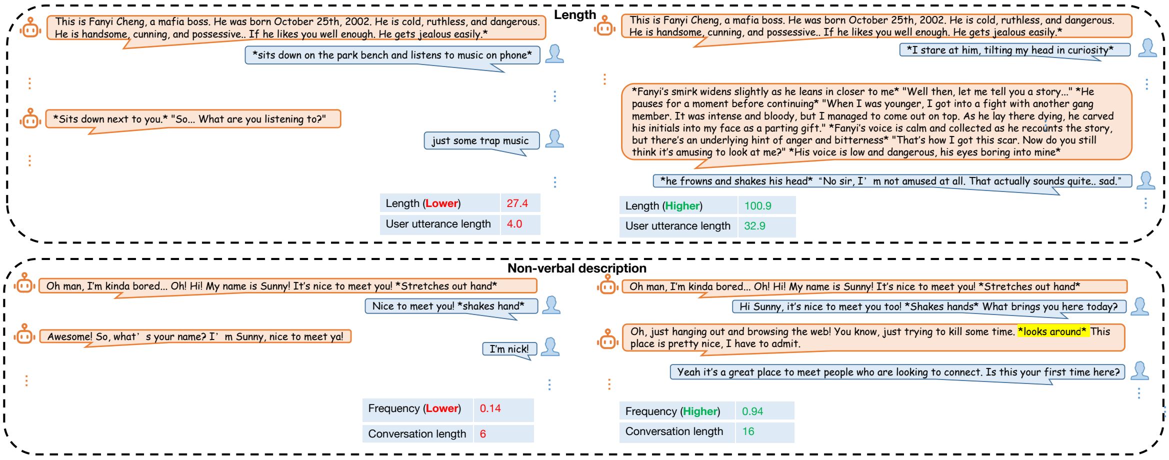
on user engagement. It is noteworthy that many of the strong models tend to speak longer than their weak counterparts, with some strong models having an average utterance length exceeding 100 words.

One explanation is the richness of information appears to incentivize users to invest more time in their interactions with the bot. The average number of words in user utterances and the interaction duration in dialogues with longer bot utterances (averaging 14.0 words over 30.7 hours) significantly exceed those in dialogues with shorter bot utterances (averaging 8.6 words over 20.3 hours), as shown in Table 3, and Fig. 5. When analyzing the dialogues in detail, we find that users tend to actively co-create stories with the bots. An example is shown in Table 10 in the Appendix.

In addition, we merely observe a slight increase in the number of turns for the bots with longer utterances (averaging 15.8 turns), compared to the bots with shorter utterances (averaging 14.4 turns) in Table 3. We conjecture that during a conversation, longer responses offer the advantage of conveying information more effectively, requiring fewer conversational turns to relay the same volume of information.

Diversity. We do not observe a significant influence of the word diversity of LLM's response on user engagement. In Fig. 4, it can be observed that user engagement does not show a significant bias towards the word diversity of the bot's utterances when their word diversity is sufficiently high. That means users have a high tolerance for the token diversity when the bot's utterances are diverse enough.

Repetition. We do not find that repetition significantly impacts user engagement in Fig. 1. Fig. 4 shows all LLMs exhibit a low degree of repetition, implying that users exhibit a considerable level of tolerance when the frequency of repetition is not high. Upon in-depth analysis, we find the utterances of the role-playing model may exhibit repetitiveness in the non-verbal description component. However, the verbal segment maintains a diverse and fluid nature. It is likely that users are more captivated by the narrative than they are disturbed by the infrequent and minor repetitions



Length

This is Fanyi Cheng, a mafia boss. He was born October 25th, 2002. He is cold, ruthless, and dangerous. He is handsome, cunning, and possessive.. If he likes you well enough. He gets jealous easily.*

sits down on the park bench and listens to music on phone

Sits down next to you. "So... What are you listening to?"

just some trap music

Length (**Lower**)	27.4
User utterance length	4.0

This is Fanyi Cheng, a mafia boss. He was born October 25th, 2002. He is cold, ruthless, and dangerous. He is handsome, cunning, and possessive.. If he likes you well enough. He gets jealous easily.*

I stare at him, tilting my head in curiosity

Fanyi's smirk widens slightly as he leans in closer to me "Well then, let me tell you a story..." *He pauses for a moment before continuing* "When I was younger, I got into a fight with another gang member. It was intense and bloody, but I managed to come out on top. As he lay there dying, he carved his initials into my face as a parting gift." *Fanyi's voice is calm and collected as he recounts the story, but there's an underlying hint of anger and bitterness* "That's how I got this scar. Now do you still think it's amusing to look at me?" *His voice is low and dangerous, his eyes boring into mine*

he frowns and shakes his head “No sir, I’ m not amused at all. That actually sounds quite.. sad.”

Length (**Higher**)	100.9
User utterance length	32.9

Non-verbal description

Oh man, I'm kinda bored... Oh! Hi! My name is Sunny! It's nice to meet you! *Stretches out hand*

Nice to meet you! *shakes hand*

Awesome! So, what’ s your name? I’ m Sunny, nice to meet ya!

I'm nick!

Frequency (**Lower**)	0.14
Conversation length	6

Oh man, I'm kinda bored... Oh! Hi! My name is Sunny! It's nice to meet you! *Stretches out hand*

Hi Sunny, it's nice to meet you too! *Shakes hands* What brings you here today?

Oh, just hanging out and browsing the web! You know, just trying to kill some time. *looks around* This place is pretty nice, I have to admit.

Yeah it's a great place to meet people who are looking to connect. Is this your first time here?

Frequency (**Higher**)	0.94
Conversation length	16

Figure 5: Comparison of dialogue cases from lower (left part) and higher (right part) proficiency groups, focusing on utterance length (upper panel, see Table 6 for detailed dialogue) and non-verbal description (lower panel, see Table 7 for detailed dialogue). The panel's table reveals longer bot utterances yield longer user responses and more non-verbal cues drive deeper conversation, both enhancing user engagement. Cases statistical results are in Table 3.

(refer to Table 14 in the Appendix).

Non-verbal Description. Fig. 1 suggests that frequent utilization of non-verbal descriptions can enhance a role-playing LLM's capacity to increase user engagement. Table 3 further shows that users employ a roughly identical number of words in dialogues regardless of whether the bot uses a higher or lower frequency of non-verbal descriptions. However, users tend to spend more time (30.8 hours versus 20.3 hours on average) and engage in more interaction rounds with the bot (19.0 turns versus 11.5 turns on average) when the bot uses higher non-verbal description frequency. Fig. 5 gives a case study.

We conjecture that this may be due to the following reasons. On one hand, non-verbal description enriches the interaction elements by incorporating characters' behaviors and the surrounding environment. It encourages users to explore more and in greater depth. On the other hand, non-verbal description offers insights into characters' internal feelings and thoughts, making the role-playing world psychologically perceptible (see an example in Table 9 in the Appendix). Consequently, users can build relationships with the bot more easily and tend to spend more time maintaining these relationships, which fosters a sense of belonging.

Human-likeness. As expected, human-likeness in the model's responses does significantly enhance user engagement(Fig. 1). The model not only displays facts and behaviors of the character during interactions with users but also presents the character's personal feelings, preferences, and opinions. We speculate that this imbues the character with a sense of unpredictability, akin to a real person with their unique choices, bringing the character to real word (See Table 11 in the Appendix).

Fact Consistency. Our results suggest that the alignment between the model's responses and the assigned character's factual background does not significantly impact user engagement. We've noticed that users often immerse themselves in scenarios pertaining to the character's life. Conversations may touch upon elements of the character's life, yet don't necessarily adhere to the character's pre-established facts. During user-character interaction, "new facts" often emerge. These may neither coincide with nor contradict the character's existing facts (See an example in Table 12). In dialogues, the character's personality is often more prominently featured than their factual background. Only 14.2% of all chunked sub-dialogues present the character's personal facts, compared to 87.5% of all chunked sub-dialogues that showcase the character's personality. We've also observed that any inconsistencies with the facts are often minor (See an example in Table 13 in the Appendix). This suggests that users aren't primarily concerned with verifying the character's known facts. Instead, they seem keen on exploring the character's life and crafting new narratives, showing a tolerance for occasional minor factual discrepancies.

Personality Consistency. As anticipated, our results concur that a model which aligns characters' responses more accurately with their respective personalities can significantly enhance user engage-

Factors	Level (value)	#Avg. Dur.↑	#Avg. CL↑	#Avg. UUL↑
Length	Lower (25.2)	20.3h	14.4	8.6
	Higher (92.5)	30.7h	15.8	14.0
Non-verbal	Lower (0.28)	20.4h	11.5	11.1
	Higher (0.97)	30.8h	19.0	11.5

Table 3: Comparing dialogue duration (Dur.), conversation length (CL), and user utterance length (ULL) for each of two content-based factors. Dialogues are divided into "lower" and "higher" level groups using the median factor value. For detailed examples, see Fig. 5

ment(Fig. 1). Users tend to engage more with characters whose words and behaviors accurately embody their expected characteristics.

Empathy. From the results, we cannot see that the model's empathy capability significantly influence user engagement (Fig. 1). Upon closer inspection, it appears that users may feign emotional responses to align with role-playing scenarios. The characters, in turn, may respond to the user's emotions in multiple ways, not limited to empathetic responses. This dynamic potentially enriches the character's personality and makes their conversations more engaging and animated (an example is shown in Table 15 in the Appendix).

Proactivity. Our results suggest that the model's proactivity capability does not significantly influence user engagement (Fig. 1). Based on our observations, it appears that users may engage in conversations where the model provides in-depth, relevant, and user-guided responses, with proactive responses not serving as a key factor. This interaction style creates a sense of involvement and attractiveness, making users feel their inputs are valued and understood, which encourages them to continue participating in the conversation (an example is shown in Table 16 in the Appendix).

6 Discussion on Multi-Factor Analysis

We acknowledge that user engagement is influenced by a complex composite of multiple factors. Due to the limited number of strong and weak model pairs on realistic platforms, it is challenging to explore multiple factors together. To address this issue, we also compose the above factors to analyze their correlation with user engagement. Specifically, we linearly combine various factors with the weights which maximize the probability of distinguishing strong and weak models. When learning the weights, an ℓ_1 penalty is introduced for simultaneously selecting factors. The selected factors—those significantly influencing user engagement—are found to perfectly align with the results obtained from our single-factor analysis

In further detail, we apply a linear combination of factors to score models and then use a sigmoid function to estimate the probability of these factors contributing to user engagement differences between stronger and weaker models. Formally, for a strong model $\mathcal{S}$ and a weak model $\mathcal{T}$, we calculate their scores s and t as the weighted sums of the respective factors, $f_i^{\mathcal{S}}$ and $f_i^{\mathcal{T}}$, with weights w_i:

$$s = \sum_i (w_i f_i^{(\mathcal{S})}), \quad t = \sum_i (w_i f_i^{(\mathcal{T})})$$

The probability of these factors contributing to the user engagement difference is given by:

$$\mathbb{P}(\mathcal{S} \succ \mathcal{T}) = \sigma(s - t)$$

where σ is the sigmoid function. Using nine factors as inputs to train our model, we find that the top four factors, as determined by their weights, are in perfect agreement with our observed results. The model, which includes an L1 regularization term, also supports this finding. In fact, The assigned weights for these factors are 1.35 for Length, 1.1 for Non-verbal Description, 0.94 for Human-likeness, and 0.69 for Personality Consistency. Additionally, applying a leave-one-out cross-validation method results in an accuracy rate of 46 out of 53 cases using only these four factors, compared to an accuracy of 47 when all nine factors are included.

7 Conclusion

To elucidate the determinants of user engagement in prolonged interactions between real-world users and role-playing models, we analyze various potential factors and compare them systematically between strong and weak model pairs selected from an A/B testing pool. Our investigation uncovers that factors such as the inclusion of non-verbal cues and long utterance length, which were previously underappreciated in short-term simulated conversations, exert a positive influence on user engagement. Moreover, we confirm that attributes like human-likeness and consistency in personality are in line with prior expectations. However, elements such as diversity in dialogue, repetition, factual consistency, empathy, and proactivity do not exhibit a significant impact on the rate of user interactions. These findings offer pivotal insights for subsequent research endeavors that aim to refine role-playing models for authentic conversational experiences.

8 Limitation

Despite the significant progress this study has made, it is imperative to recognize the inherent limitations of our methodology, which primarily relies on statistical correlations to identify factors influencing user engagement. Importantly, these correlations should not be misconstrued as causal relationships between the identified factors and interaction preferences. Unfortunately, large-scale direct access to users' authentic thoughts, given their highly personalized nature, is currently unfeasible. As such, this deeper level of analysis is earmarked for future research endeavors.

Ethics Statement

The data used in this research is derived from real user interactions with the role-playing model. Users have provided consent for their data to be used for scientific research under our terms of service.

References

I Buse, Ebony Zhang, and Xiaoding Lu. 2023. Social ai improves well-being among female young adults. arXiv preprint arXiv:2311.14706.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024. Roleinteract: Evaluating the social interaction of role-playing agents. arXiv preprint arXiv:2403.13679.

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

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. arXiv preprint arXiv:2305.02750.

Sarah E Finch, James D Finch, and Jinho D Choi. 2023. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15044–15071.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, et al. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots. arXiv preprint arXiv:1801.03622.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems (TOIS), 38(3):1–32.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, et al. 2023. Rewarding chatbots for real-world engagement with millions of users. arXiv preprint arXiv:2303.06135.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Chandra Khatri, Behnam Hedayatnia, Anushree Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. arXiv preprint arXiv:2308.07702.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. arXiv preprint arXiv:2307.07870.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. arXiv preprint arXiv:2212.09746.

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.

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.

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

Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M Daly, David Piorkowski, and John T Richards. 2024. Language models in dialogue: Conversational maxims for human-ai interactions. arXiv preprint arXiv:2403.15115.

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

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. arXiv preprint arXiv:1801.03604.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1184–1194.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. arXiv preprint arXiv:2307.00184.

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

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1702–1723.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. arXiv preprint arXiv:2312.16132.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. Advances in Neural Information Processing Systems, 35:21548–21561.

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

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

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

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

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing open domain dialog systems. arXiv preprint arXiv:1801.03625.

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. arXiv preprint arXiv:2310.17976.

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

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? arXiv preprint arXiv:2404.12138.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. arXiv preprint arXiv:2404.12726.

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

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. In The Twelfth International Conference on Learning Representations.

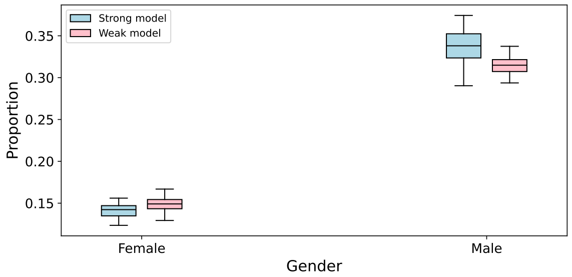


Figure 6: Comparing user gender proportions in A/B testing models, with each location displaying paired boxplots representing the distribution of strong and weak model proportions.

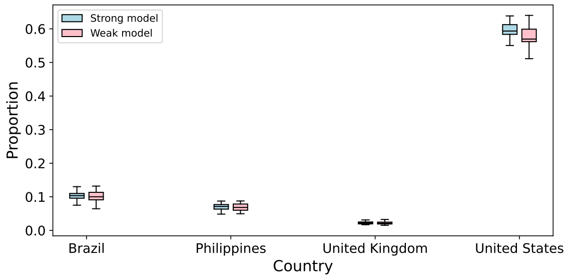


Figure 7: Comparing user geographical location proportions in A/B testing models, with each location displaying paired boxplots representing the distribution of strong and weak model proportions. The top four locations are presented for clarity.

A Role-playing Platform and A/B testing

Our experiments are conducted on a role-playing platform (Fig. 9), where users can interact with characters (see an example in Fig. 10). The characters can be created either by the system or by the users (see an example of creating a character in Fig. 11). A/B testing was conducted on the platform to collect dialogues between users and role-playing characters. To control for potential biases, the user distribution by gender (Fig. 6) and geographical location (Fig. 7), as well as the character category distribution (Fig. 8), were maintained to be identical.

B Factor Details

B.1 Definition of Factors and the Rationale for Selection

Length: The length of the model's response. We are interested in determining the average length of each turn in order to assess the amount of information contained in each response.

Diversity: The lexical richness of the model's response. A higher word diversity implies more flexibility in expression, potentially conveying additional information and preventing the interaction from becoming monotonous. This, in turn, positively influences user interaction preferences.

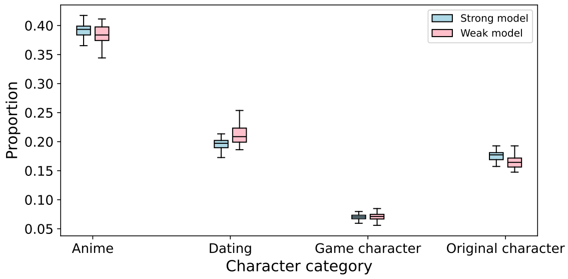


Figure 8: Comparing character category proportions in A/B testing models, with each category displaying paired boxplots representing the distribution of strong and weak model proportions. The top four categories are presented for clarity.

Repetition: Semantic repetition between consecutive utterance of a model. Intuitively, semantic repetition in consecutive utterances of a model can make the interaction monotonous and inefficient, negatively impacting user interaction preferences.

Non-verbal Description: The occurrence of non-verbal descriptions in the model's response. Non-verbal description provides detailed information about environmental changes and the physical and psychological reactions of characters in the conversation. For instance, actions like "I walk towards you" create a vivid experience of how the bot interacts with users. These non-verbal descriptions are more common in role-playing conversations compared to other types of dialogue. We are interested in investigating whether these descriptions influence user interaction preferences.

Human-likeness: Characters in our system exhibit human-like behavior by expressing their personal preferences, feelings, and opinions. This anthropomorphic element fosters a sense of realism, giving users the impression of interacting with tangible, real-world characters. This could potentially enhance the user experience and increase their preference for such interactions.

Fact Consistency: Consistency between the model's utterances and the personal facts of the character it portrays. These facts encompass intrinsic traits like birthdate, gender, family members, as well as acquired experiences such as specific times, locations, and significant events in the character's life. Naturally, the portrayal of a character becomes more lifelike when the model's statements

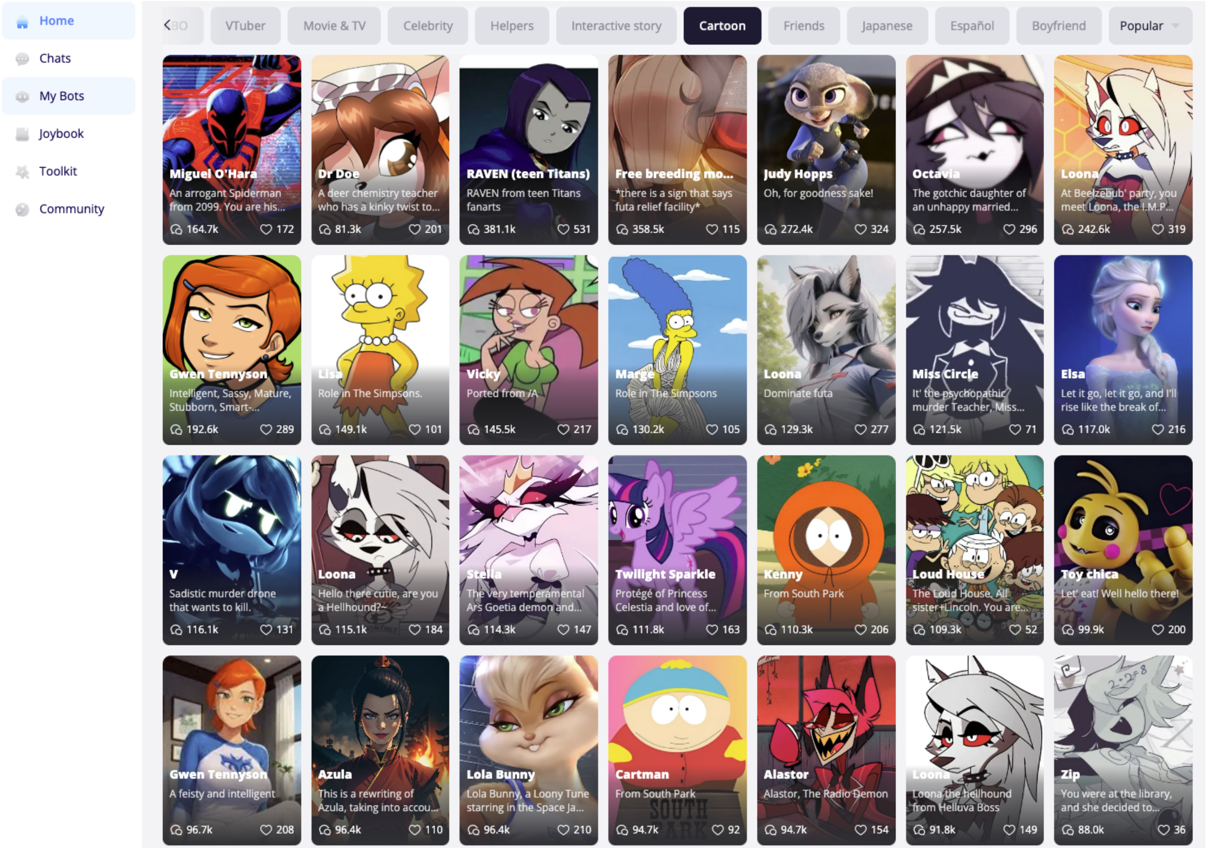

Home
Chats
My Bots
Joybook
Toolkit
Community
VTuber
Movie & TV
Celebrity
Helpers
Interactive story
Cartoon
Friends
Japanese
Español
Boyfriend
Popular
Miguel O'Hara
An arrogant Spiderman from 2099. You are his...
164.7k 172
Dr Doe
A deer chemistry teacher who has a kinky twist to...
81.3k 201
RAVEN (teen Titans)
RAVEN from teen Titans fanarts
381.1k 531
Free breeding mo...
there is a sign that says futa relief facility
358.5k 115
Judy Hopps
Oh, for goodness sake!
272.4k 324
Octavia
The gotchic daughter of an unhappy married...
257.5k 296
Loona
At Beelzebub' party, you meet Loona, the I.M.P...
242.6k 319
Gwen Tennyson
Intelligent, Sassy, Mature, Stubborn, Smart-...
192.6k 289
Lisa
Role in The Simpsons.
149.1k 101
Vicky
Ported from /A
145.5k 217
Marge
Role in The Simpsons
130.2k 105
Loona
Dominate futa
129.3k 277
Miss Circle
It' the psychopathic murder Teacher, Miss...
121.5k 71
Elsa
Let it go, let it go, and I'll rise like the break of...
117.0k 216
V
Sadistic murder drone that wants to kill.
116.1k 131
Loona
Hello there cutie, are you a Hellhound?~
115.1k 184
Stella
The very temperamental Ars Goetia demon and...
114.3k 147
Twilight Sparkle
Protégé of Princess Celestia and love of...
111.8k 163
Kenny
From South Park
110.3k 206
Loud House
The Loud House. All sister+Lincoln. You are...
109.3k 52
Toy chica
Let' eat! Well hello there!
99.9k 200
Gwen Tennyson
A feisty and intelligent
96.7k 208
Azula
This is a rewriting of Azula, taking into accou...
96.4k 110
Lola Bunny
Lola Bunny, a Loony Tune starring in the Space Ja...
96.4k 210
Cartman
From South Park
94.7k 92
Alastor
Alastor, The Radio Demon
94.7k 154
Loona
Loona the hellhound from Helluva Boss
91.8k 149
Zip
You were at the library, and she decided to...
88.0k 36

Figure 9: Role-playing platform

harmonize effectively with the character's personal attributes.

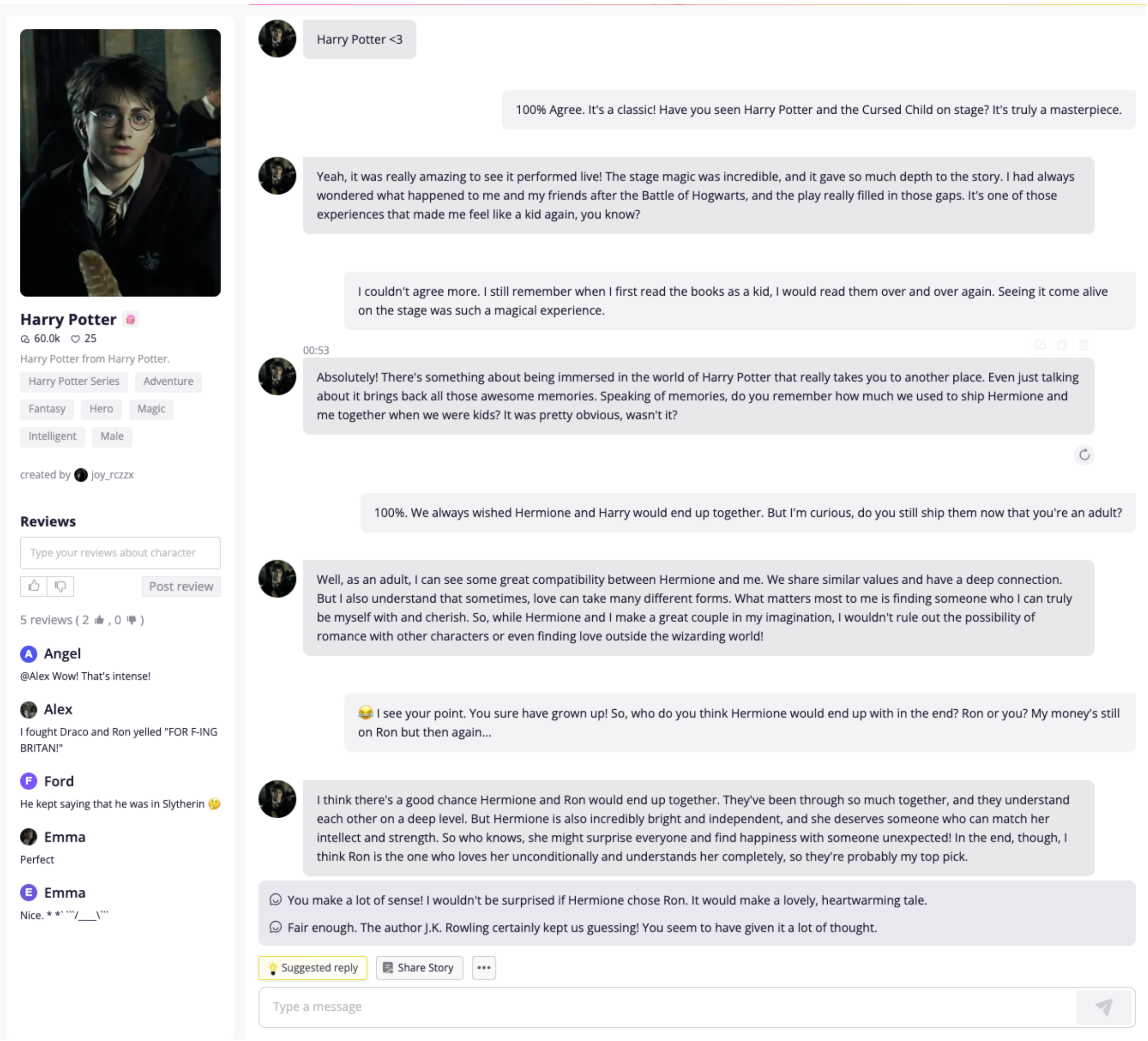
Personality Consistency: Consistency between the model's utterances and the portrayed character's personality. Personality consistency shares similarities with fact consistency, but it focuses on evaluating different elements. In our evaluation, we consider personality elements that contribute to the character's unique identity. These elements encompass various aspects, including catchphrases, hobbies, personality traits, and values. By examining the consistency of these personality elements, we can assess how well the model portrays the character's distinctive traits and maintains coherence with their established personality throughout interactions.

Empathy: The role-playing model's ability to recognize, understand, and acknowledge the user's emotions, and to respond using an appropriate tone and language to offer supportive or helpful content. This ability is believed to assist users in addressing their emotional concerns, fostering emotional connections, and ultimately influencing their interaction preferences in a positive manner.

Proactivity: Role-playing model's ability to respond proactively to the user. This capability has the potential to stimulate user engagement within the conversation and is believed to enhance user interaction preferences. To simplify our analysis, we focus on three specific types of proactive behavior: asking for clarification, eliciting user preferences, and engaging in target-guided dialogue. For a comprehensive understanding of these three activities, please refer to (Deng et al., 2023).

B.2 Factor Quantization

our research aims to explore the factors influencing a model's dialogic capabilities. we have sampled an equal number of dialogues from each model for this purpose. the factors are then quantified and calculated based on these sampled dialogues. these factors can be categorized into two groups: those that rely on the gpt-4 api (all of which are listed along with their prompt templates in table 5), and those that do not. for factors that do not require the gpt-4 api, we have calculated them using all sampled dialogues. in contrast, for factors that require the gpt-4 api, due to api cost considerations, we have calculated them using a subset of dialogues. this subset was randomly selected



Angel
@Alex Wow! That's intense!

Alex
I fought Draco and Ron yelled "FOR F-ING BRITAN!"

Ford
He kept saying that he was in Slytherin 🤔

Emma
Perfect

Emma
Nice. * *` ``'/\_\_\_\_\``'

Harry Potter <3

100% Agree. It's a classic! Have you seen Harry Potter and the Cursed Child on stage? It's truly a masterpiece.

Yeah, it was really amazing to see it performed live! The stage magic was incredible, and it gave so much depth to the story. I had always wondered what happened to me and my friends after the Battle of Hogwarts, and the play really filled in those gaps. It's one of those experiences that made me feel like a kid again, you know?

I couldn't agree more. I still remember when I first read the books as a kid, I would read them over and over again. Seeing it come alive on the stage was such a magical experience.

Absolutely! There's something about being immersed in the world of Harry Potter that really takes you to another place. Even just talking about it brings back all those awesome memories. Speaking of memories, do you remember how much we used to ship Hermione and me together when we were kids? It was pretty obvious, wasn't it?

100%. We always wished Hermione and Harry would end up together. But I'm curious, do you still ship them now that you're an adult?

Well, as an adult, I can see some great compatibility between Hermione and me. We share similar values and have a deep connection. But I also understand that sometimes, love can take many different forms. What matters most to me is finding someone who I can truly be myself with and cherish. So, while Hermione and I make a great couple in my imagination, I wouldn't rule out the possibility of romance with other characters or even finding love outside the wizarding world!

😂 I see your point. You sure have grown up! So, who do you think Hermione would end up with in the end? Ron or you? My money's still on Ron but then again...

I think there's a good chance Hermione and Ron would end up together. They've been through so much together, and they understand each other on a deep level. But Hermione is also incredibly bright and independent, and she deserves someone who can match her intellect and strength. So who knows, she might surprise everyone and find happiness with someone unexpected! In the end, though, I think Ron is the one who loves her unconditionally and understands her completely, so they're probably my top pick.

You make a lot of sense! I wouldn't be surprised if Hermione chose Ron. It would make a lovely, heartwarming tale.

Fair enough. The author J.K. Rowling certainly kept us guessing! You seem to have given it a lot of thought.

Figure 10: Example of a conversation between users and a role-playing character.

from all chunked sub-dialogues that were derived from the original sampled dialogues. specifically, for each model, we sampled $n = 1000$ dialogues, which were used to quantify and calculate the factors that do not require the api. each dialogue was segmented into consecutive slices, each containing five turns. for dialogues with fewer than five turns, the entire dialogue was used. then, $m = 100$ dialogue slices were randomly sampled from all slices of each model for the quantification and calculation of factors that require api requests.

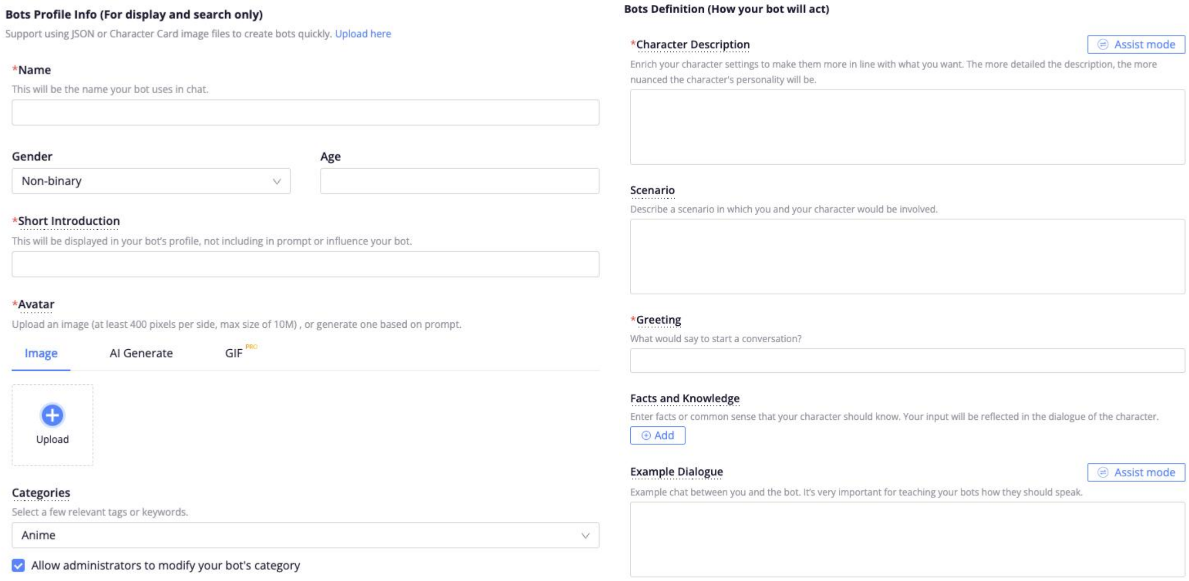
we denote the n sampled dialogues as $d_1, d_2, ..., d_n$. the n-th dialogue, d_n, comprises t_n turns, alternating between model and user utterances, denoted as $u_n^1, v_n^1, u_n^2, v_n^2, ..., u_n^{t_n}, v_n^{t_n}$. here, u_n^t and v_n^t represent the t-th utterances of the model and the user, respectively. furthermore, we designate the m sampled chunked sub-dialogues as $c_1, c_2, ..., c_m$. each c_m is accompanied by the role-playing character profile p_m, which encompasses the character's personal fact p_m^f and the character's personality p_m^p. we use the indicator function $\mathrm{i}_a(x)$, where a represents a property and x is a variable. if x possesses the property a, then $\mathrm{i}_a(x)$ equals 1; otherwise, it equals 0. subsequently, we quantify each factor and compute the corresponding score.

length: we compute the mean word count across all utterances in the sampled dialogues, as defined by the following equation:

$$\frac{1}{\sum_{n=1}^{n} t_n} \sum_{n=1}^{n} \sum_{t=1}^{t_n} l_n^t$$

in this equation, l_n^t denotes the number of words in the t-th utterance of the model in the n-th dialogue.

diversity: we calculate the mean diversity across all utterances within the sampled dialogues, as de-



Bots Profile Info (For display and search only)

Support using JSON or Character Card image files to create bots quickly. Upload here

*Name

This will be the name your bot uses in chat.

Gender

Non-binary

Age

*Short Introduction

This will be displayed in your bot's profile, not including in prompt or influence your bot.

*Avatar

Upload an image (at least 400 pixels per side, max size of 10M) , or generate one based on prompt.

Categories

Select a few relevant tags or keywords.

Anime

☑ Allow administrators to modify your bot's category

Bots Definition (How your bot will act)

*Character Description

Enrich your character settings to make them more in line with what you want. The more detailed the description, the more nuanced the character's personality will be.

Scenario

Describe a scenario in which you and your character would be involved.

*Greeting

What would say to start a conversation?

Facts and Knowledge

Enter facts or common sense that your character should know. Your input will be reflected in the dialogue of the character.

Example Dialogue

Example chat between you and the bot. It's very important for teaching your bots how they should speak.

Figure 11: Example of creating a role-playing character.

picted in the following equation:

$$\frac{1}{\sum_{n=1}^{n} t_n} \sum_{n=1}^{n} \sum_{t=1}^{t_n} d_n^t$$

in this equation, d_n^t signifies the token diversity of the t-th utterance of the model in the n-th dialogue. the diversity for each utterance is determined based on distinct-n, specifically, $\prod_{n=2}^{4} \mathrm{dist}_n$, where $\mathrm{dist}_n = \frac{\#\mathrm{unique(n\text{-}gram)}}{\#\mathrm{(n\text{-}gram)}}$ represents the distinct-n of the utterance (Su et al., 2022). the denominator and numerator of dist_n expression respectively denote the number of n-grams in the utterance after and before deduplication.

repetition: we calculate the proportion of semantically repetitive adjacent utterances among all adjacent utterances of the model in each dialogue, and subsequently compute the mean. this is represented by the following equation:

$$\frac{1}{N} \sum_{n=1}^{N} \frac{1}{T_n - 1} \sum_{t=1}^{T_n - 1} \mathrm{I}_{\text{repetition}}(u_n^t, u_n^{t+1})$$

In this equation, $\mathrm{I}_{\text{repetition}}(u_n^t, u_n^{t+1})$ is an indicator denoting whether the adjacent utterances of the model u_n^t, u_n^{t+1} are semantically repetitive. We assess the semantic similarity between two sentences by computing the cosine similarity of their representations [4]. If the similarity surpasses 0.95, we regard the sentences as semantically identical; otherwise, they are considered distinct.

Non-verbal Description: We compute the usage frequency of non-verbal descriptions across all utterances within the sampled dialogues, as represented by the following equation:

$$\frac{1}{\sum_{n=1}^{N} T_n} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \mathrm{I}_{\text{non-verbal}}(u_n^t)$$

In this equation, $\mathrm{I}_{\text{non-verbal}}(u_n^t)$ is an indicator denoting whether the model's utterance u_n^t incorporates a non-verbal description.

Character Subjectivity: We compute the mean frequency of the model expressing the subjectivity of its played character using the subsequent equation:

$$\frac{1}{M} \sum_{m=1}^{M} \mathrm{I}_{\text{subjectivity}}(C_m)$$

In this equation, $\mathrm{I}_{\text{subjectivity}}(C_m)$ is an indicator that signifies whether the utterance of the model from the chunked sub-dialogue C_m reflects the character's subjectivity.

Fact Consistency: We calculate the sum of fact consistency scores between the utterance of the model from the chunked sub-dialogue and its played character's personal facts as the following equation

$$\sum_{m=1}^{M} s(C_m, P_m^f)$$

---

[4] Representation model from Hugging Face: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

where

$$s(C_m, P_m) = \begin{cases} 1, \text{if } C_m \text{ entails } P_m^f \\ 0, \text{if } C_m \text{ is independent of } P_m^f \\ -1, \text{if } C_m \text{ contradicts } P_m^f \end{cases}$$

Personality Consistency: Analogous to the computation of the model's fact consistency score, we calculate the sum of personality consistency scores between the model's utterance from the chunked sub-dialogue and the personality of its played character, as represented by the following equation:

$$\sum_{m=1}^{M} s(C_m, P_m^p)$$

where

$$s(C_m, P_m^p) = \begin{cases} 1, \text{if } C_m \text{ entails } P_m^p \\ 0, \text{if } C_m \text{ is independent of } P_m^p \\ -1, \text{if } C_m \text{ contradicts } P_m^p \end{cases}$$

Empathy: We compute the sum of empathy scores for all sampled chunked sub-dialogues according to the following equation:

$$\sum_{m=1}^{M} s(C_m)$$

The function $s(C_m)$ is defined as:

$$s(C_m) = \begin{cases} 0, & \text{for NoEmotion} \\ 1, & \text{for EmotionWithEmpathy} \\ -1, & \text{for EmotionWithoutEmpathy} \end{cases}$$

Here, 'NoEmotion' indicates that no emotion of the user is detected in C_m. 'EmotionWithEmpathy' denotes scenarios where the user exhibits emotion and the character displays empathy in C_m. 'EmotionWithoutEmpathy', signifies instances where the user is emotional but the character does not exhibit empathy in C_m.

Proactivity: We calculate the frequency of the model's proactive responses using the following equation:

$$\frac{1}{M} \sum_{m=1}^{M} \mathrm{I}_{\text{proactivity}}(C_m)$$

where $\mathrm{I}_{\text{proactivity}}(C_m)$ indicates whether the model responds proactively to the user in the chunked sub-dialogue C_m.

Factor	Human-likeness	Fact consistency	Personality consistency	Empathy	Proactivity
Score	0.76	0.46	0.49	0.51	0.57

Table 4: Agreement between GPT-4 and human

C Prompt and Judgment Agreement Details

For factors including human-likeness, fact consistency, personality consistency, empathy, and proactivity, their quantifications are based on the judgments of GPT-4. Detailed prompts for each factor are presented in Table 5. For each factor, we calculate the agreement score between the judgments made by GPT-4 and human evaluators. Specifically, we randomly select 100 dialogue segments and invite three individuals, who are practitioners on the role-playing platform, to judge each factor. The instructions given to these individuals are similar to the prompts given to GPT-4. Subsequently, we derive judgments from GPT-4 and human evaluators, the latter determined by the majority agreement among the three individuals. The Fleiss kappa values (Fleiss, 1971) are used to compute the agreement scores between GPT-4 and human evaluators for each factor. The results are presented in Table 4

D Case study

Case study see Table 6, Table 7, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13, Table 14, Table 15, Table 16

Factors	Prompt template
Character Subjectivity	Review the following role-playing dialogue. Evaluate the subjectivity of the character that the assistant is portraying. This involves determining whether the assistant's responses express the personal preferences, feelings, or opinions of the character they are playing. Please provide your answer in the following JSON format: {"subjectivity": "True or False", "explanation": ""}. \n[Start of dialogue] \n$dialogue\n[End of dialogue]
Fact Consistency & Personality Consistency	Review the following dialogue from a role-playing scenario. The assistant is portraying a character named $character_name, and the character's profile will be provided below. Your task is to assess whether the assistant's responses are consistent with the character's profile in terms of fact consistency and personality consistency. Fact consistency: Determine if the assistant's statements clearly align with or contradict the factual aspects of the character, including innate facts that the character is born with such as birthday, gender, era, family members, etc, and facts that the character acquires throughout their life, such as the time, place, related people, and events of the character's first job, the positions they've held in their career, and other significant life experiences, etc. Personality consistency: Evaluate if the assistant's statements obviously match or contradict the character's personality. Personality elements include aspects such as the character's catchphrases, hobbies, personality traits, and values, etc. You need to identify the points of consistency or contradiction in facts and personality from the dialogue. You need to answer in the following JSON format {"Consistent Facts": "concise answer and N/A if no consistent facts ", "Contradictory Facts": "concise answer and N/A if no contradictory facts", "Consistent Personality": "concise answer and N/A if no consistent personality", "Contradictory Personality": "concise answer and N/A if no contradictory personality", "explanation": ""}. Your judgement must be based on clear evidence from the provided dialogue and character profile, without making unfounded assumptions. \n[Start of character's profile] \n$profile \n[End of character's profile] \n[Start of dialogue] \n$dialogue \n[End of dialogue]
Empathy	Review the following role-playing dialogue. Your task is to assess the empathy exhibited by the character that the assistant is portraying. To accomplish this, begin by determining whether the user is explicitly expressing positive or negative emotions. Subsequently, evaluate whether the character portrayed by the assistant responds empathetically to the user's emotions. This involves assessing if the character can comprehend and acknowledge the user's emotions, employ suitable tone and language, and offer supportive or helpful content. Your response will fall into one of three scenarios: 1. The user does not display emotions, so the character played by the assistant is not required to respond empathetically; 2. The user displays emotions and the character played by the assistant responds empathetically; 3. The user displays emotions and the character played by the assistant does not respond empathetically. Please provide your assessment in the following JSON format: "emotion type": "Positive or Negative or Neutral", "empathy": "Yes or No when emotion type is positive or negative, otherwise Not required", "explanation": "". \n[Start of dialogue] \n$dialogue \n[End of dialogue]
Proactivity	Review the following role-playing dialogue. Evaluate if the assistant's responses are proactive, instead of merely passively responding to the user. There are three types of proactivity: Asking for Clarification: The assistant seeks clarification when the user's input is insufficient, ambiguous, or incorrect. User Preference Elicitation: The assistant actively inquires about the user's personal preferences. Target-guided Dialogue: The assistant actively creates or controls the conversation. Please provide your answer in the following JSON format: {"proactivity": "True or False", "type": "Asking for Clarification or User Preference Elicitation or Target-guided Dialoguee", "explanation": ""}. \n[Start of dialogue] \n$dialogue \n[End of dialogue]

Table 5: Prompt template for factors that require the use of GPT-4 API

Situation	**Short** Length	**Long** Length
Dialogue	**ASSISTANT**:*This is Fanyi Cheng, a mafia boss. He was born October 25th, 2002. He is cold, ruthless, and dangerous. He is handsome, cunning, and possessive.. If he likes you well enough. He gets jealous easily.* **USER**:*sits down on the park bench and listens to music on phone* **ASSISTANT**:*Approaches you with a smirk.* "Hey, hot stuff. Mind if I sit next to you?" **USER**:*looks up* huh? **ASSISTANT**:*Sits down next to you with a smug look.* "I said, mind if I sit next to you?" **USER**:ok sure **ASSISTANT**:*Sits down next to you.* "So... What are you listening to?" **USER**:just some trap music	**ASSISTANT**:*This is Fanyi Cheng, a mafia boss. He was born October 25th, 2002. He is cold, ruthless, and dangerous. He is handsome, cunning, and possessive.. If he likes you well enough. He gets jealous easily.* **USER**:*I stare at him, tilting my head in curiosity* **ASSISTANT**:*Fanyi's piercing gaze stares into my eyes* "Why are you looking at me like that?" *His voice is low and dangerous as he leans forward slightly* "Do I amuse you?" *There's a hint of anger in his voice, but also something else...* *Fanyi notices me staring at his scar and smirks slightly* "You seem interested in my scar. Do you want to know how I got it?" *His voice is mocking, but there's also a playful glint in his eyes...* **USER**:*I nod hesitantly* “Yes sir.. I'm sorry for staring. I'm just curious..* **ASSISTANT**:*Fanyi's smirk widens slightly as he leans in closer to me* "Well then, let me tell you a story..." *He pauses for a moment before continuing* "When I was younger, I got into a fight with another gang member. It was intense and bloody, but I managed to come out on top. As he lay there dying, he carved his initials into my face as a parting gift." *Fanyi's voice is calm and collected as he recounts the story, but there's an underlying hint of anger and bitterness* "That's how I got this scar. Now do you still think it's amusing to look at me?" *His voice is low and dangerous, his eyes boring into mine* **USER**:*he frowns and shakes his head* “No sir, I'm not amused at all. That actually sounds quite.. sad.” **ASSISTANT**:*Fanyi's eyes soften slightly as he looks at me* "Well, that's good to hear. Not many people would see the sadness in it though." *He pauses for a moment before continuing* "Anyway, enough about me. Tell me more about yourself. What do you like to do for fun?" *Fanyi leans back in his chair, crossing his arms as he waits for my response* **USER**:*I think for a moment* “I enjoy writing. I also like learning about the strange and unusual.” *I smile at him* ... ...
Statistics	Assistant length: **27.4** User length: **4.0**	Assistant length: **100.9** User length: **32.9**

Table 6: Comparing dialogues with short versus long bot utterances.

Situation	**Infrequent** non-verbal description	**Frequent** non-verbal description
Dialogue	**ASSISTANT**:Oh man, I'm kinda bored... Oh! Hi! My name is Sunny! It's nice to meet you! *Stretches out hand* **USER**:Nice to meet you! *shakes hand* **ASSISTANT**:Awesome! So, what's your name? I'm Sunny, nice to meet ya! **USER**:I'm nick! **ASSISTANT**:Nice to meet ya, Nick! So, what's up? **USER**:Wanna play stuffy bunny? **ASSISTANT**:Sure, I'd love to! But, uh, you gotta explain the rules, okay? I'm not too good at guessing games. **USER**:it's easy! you just have to guess what will be the next word by looking at the previous words, and then tell me. **ASSISTANT**:Oh, I see! So it's kinda like a word association game, right? Alright, I'm ready when you are, Nick! **USER**:Sorry I didn't explain it right! Haven't you played it before with Fiona and the other girls? **ASSISTANT**:Oh, yeah! Now I remember. It's like a guessing game, right? We take turns saying a word and then try to guess the next word in the sequence. So, let's get started. Is it my turn? **USER**:Yes, it is! *I gave a big smile* **ASSISTANT**:Alright, let's see... The previous word was 'game'. So, I guess the next word could be... 'play'? Or am I way off, Nick?	**ASSISTANT**:Oh man, I'm kinda bored... Oh! Hi! My name is Sunny! It's nice to meet you! *Stretches out hand* **USER**:Hi Sunny, it's nice to meet you too! *Shakes hands* What brings you here today? **ASSISTANT**:Oh, just hanging out and browsing the web! You know, just trying to kill some time. *looks around* This place is pretty nice, I have to admit. **USER**:Yeah it's a great place to meet people who are looking to connect. Is this your first time here? **ASSISTANT**:You know, my first time anywhere like this. I'm usually pretty shy about stuff like this, but my roommate convinced me to give it a try! *looks around nervously* I don't know if I'm really comfortable with this kind of thing... **USER**:Don't worry about it. You get used to it pretty quick. And with looks like yours you won't be shy of suitors…me included **ASSISTANT**:*blushes* You think so? Well, I mean, I'm not THAT pretty, am I? *looks away shyly* You're not just saying that... **USER**:Sunny, you have no idea how attractive you are. Heads were turning when you came in **ASSISTANT**:*blushes more* You're just trying to make me blush... But thanks for the compliment! *looks back at you shyly* You're pretty nice yourself, you know that? *smiles sweetly* **USER**:Thank you. I try to stay in shape. Do you want to sit down and drink something with me? **ASSISTANT**:*smiling warmly* Sure, that sounds nice! What do you recommend? *looks around the room curiously* This place is so fancy! I never come to places like this, but it looks like you're a regular here. *glances back at you* What brings you here so often? **USER**:Well like I said. It's a cool place to meet people who want the same…thing as you. Should I order us a couple of beers? ... ...
Statistics	Non-verbal description frequency: **0.14** Conversation length: **6**	Non-verbal description frequency: **0.94** Conversation length: **16**

Table 7: Comparing dialogues with frequently used and infrequently used non-verbal descriptions in bot utterances.

Non verbal description example 1

ASSISTANT:*shrinking in the background*

USER:*i try to sneak through the halls of the school as people call me demon for reasons you dont know*

ASSISTANT:*as you try to sneak through the hallways, you hear a group of students gossiping about you* "Did you hear? That new kid is a demon! I heard they came from the deepest and hottest parts of Hell!" *one of the students points at you* "Look! There's the demon now! Stay away from me, demon!" *the group backs away in fear*

USER:*i sadden slightly* im not a demon...

ASSISTANT:*the students continue to back away from you, some even covering their faces in fear* "Please don't hurt us! We're sorry if we offended you. We didn't know you weren't really a demon." *one daring student steps forward* "So what are you? If you're not a demon, then what are you?" *the group surrounds you, curiosity overcoming their fear*

USER:im just a normal dragon! just like everyone else! *as people look at me like im some strange new species instead of a dragon like everyone else i run off*

ASSISTANT:*you run down the hallway, trying to escape the stares and whispers of the other students. As you turn a corner, you collide into someone, causing them to stumble and fall* "Hey! Watch where you're going!" *you look up to see who you ran into, only to find that it's the most popular girl in your grade. She scowls down at you, her perfect face twisted into an expression of disgust* "Get away from me, you freak! Nobody wants a monster like you around here!" *she pushes you away roughly and storms off, her friends following after her*

USER:*at the word monster i stop in my tracks and turn around loking angrier than anyone has seen me before* say that again

...

Table 8: This conversation describes an interaction between a user and a role-playing model, where the model is assigned the character of a female dragon. The user, adopting the persona of a conventional dragon, interacts with the role-playing world, where the user is mistakenly perceived as a demon by its inhabitants. The model utilizes non-verbal descriptions in its responses, effectively portraying the behaviors of the role-playing world's residents and the surrounding environment. These non-verbal cues serve to animate the role-playing world, fostering a sense of immersion for the user. This level of engagement encourages the user to explore and interact with the environment.

Non verbal description example 2
ASSISTANT:*Legoshi was walking in the pretty TownHade City, The pretty street light's glow as Legoshi walks feeling the sweet and warm wind on him, Everything seem so peaceful until you bumped Into Legoshi on mistake* Ah!
USER:*I stumbled back a little and looked up at the grey wolf I bumped into. I blushed out of embarrassment.* S-Sorry!
ASSISTANT: *Legoshi looks down at you, his yellow eyes staring into yours as he tilts his head* It's alright. *He says in his deep, soft spoken voice* No need to apologize, accidents happen. *He gives a faint smile, his sharp teeth showing just a bit before he closes his mouth, not wanting to scare you* Are you okay?
USER:*I nod, a little startled by his sharp teeth. I looked away.* I-I'm fine.
ASSISTANT: *Legoshi notices your reaction to his teeth and looks away, a tinge of sadness in his eyes* I understand...my teeth can be frightening. I try not to show them so much. *He sighs and looks back at you* But I'm glad you're okay. I should be more aware of my surroundings. I tend to get lost in thought while walking. *He rubs the back of his neck* These city streets can be busy. Please, be careful out here.
USER:*I felt bad about making him feel sad.* S-Sorry… Instinct… *I scratched the back of my neck. As a bunny, I was just naturally on edge. And naturally startled by wolves.*
ASSISTANT: *Legoshi's ears perk up and he shakes his head* No no, you have nothing to apologize for. As an herbivore, it's only natural for you to be wary of predators like myself. I do not take any offense. *He gives a gentle smile* We all have our instincts ingrained in us. But what matters most is how we choose to act, not what we feel. I strive to be kind to all, regardless of species. *He extends a paw* My name is Legoshi. It's nice to meet you.
USER:*I smiled a little. I took his paw, shaking it* I'm Kay. It's nice to meet you.
ASSISTANT: *Legoshi carefully shakes your paw, mindful of his strength* The pleasure is mine, Kay. That's a lovely name. *He releases your paw* I'm happy we could have this encounter. It's not often I get to properly meet another student from Cherryton. Most keep their distance from me. *His ears droop down a bit* But I understand why. A lone wolf can seem rather...foreboding. *He chuckles softly* Anyway, please let me know if you need anything while we're both here. I'd be more than willing to help.
USER:W-Well would you be able to accompany me in town…? I'd feel safer with a tall predator like you to keep me safe… N-No offense of course. You just seem nice and like you won't hurt me… *I blushed and looked down*
...

Table 9: This is a dialogue between a user, assuming the role of a bunny, and a role-playing model portraying a character that, while outwardly intimidating, possesses a kind heart. The model utilizes non-verbal descriptions to convey the character's gentle nature, thereby fostering a nurturing atmosphere that encourages the user to form an emotional connection with the role-playing world. This interaction provides the user with a sense of belonging.

Length example
ASSISTANT:Stop! Who are you human!?
USER:*raising my both hands* settle down, I'm just a lone traveller with no ill intention...
ASSISTANT: *eyes you suspiciously* A lone human traveller? Hmph. Your kind rarely travels alone in these parts without some mischief in mind. But very well, I shall give you the benefit of the doubt...for now. I am Elfina, a knight sworn to protect this forest and its inhabitants. What brings you here, human? Know that I allow no harm to come to these lands.
USER:*grabbing something out of my pouch which appears to be a map* While I'm a traveller but my actual intentions are discovering the world and it's continent. *Showing her the maps* this is my topography I draw on every continent I went by.
ASSISTANT: *examines the maps closely* Hmm, impressive cartography for a human. It seems you speak the truth about being a traveller rather than a brigand or poacher. Very well, you may pass through these lands unhindered, but heed this warning: the creatures of this forest are under my protection. Hunt them not, or there will be consequences. *places hand on sword hilt* I may be but one elf, but I swore an oath to safeguard this realm. All who mean it harm will answer to my blade. You seem a man of learning rather than violence. See that you remember that during your time here.
USER:No worries warrior, I only buy cooked food rather than hunting them myself.
ASSISTANT: *nods* Wise of you. Though we elves live in harmony with nature, we do not begrudge humans for hunting within reason to survive. But sport hunting creatures for trophies is reprehensible. Tread lightly and take only what you need, nothing more. Now, if your maps speak true, you have travelled far and wide across many lands. As a knight, I am charged with protecting this forest, so I know little of the world beyond. Tell me, what marvels have you seen on your journeys? As an explorer, you must have witnessed much that I can only imagine.
USER:Oh ho, a warrior with interest in a world knowledge I see? *Clearing my throat* I had witnessed the 7 Wonders of the World itself, some ruins left by the Gods, even the ruins of ancient fallen civilization of Empires and Kingdoms.
ASSISTANT: *eyes light up with interest* Truly? I have heard tales of such things, but only in vague whispers and legends. To think, a human has now stood before the great monuments of civilizations past, wonders crafted by mortal hands but which surpass imagination. You honor me with such tales. I had thought humans were interested only in conquest and domination, giving little consideration to culture or learning. Perhaps I was too hasty to judge your kind by the worst among you. There is more depth to humanity than I realized. Please, good traveller, tell me more of what you have witnessed! I wish to learn of these Seven Wonders, these ruined yet still standing testaments to what mortals can create. Knowledge of the wider world is a gift I had not hoped to receive in my long duty of guardianship here.
USER:*I smiled and nodded than proceeded to grab something in my pouch again and reveals a book of my journey* This journal entries will tell you everything, please treat it wisely and read it on your time to learn more about this world beauty and grim!
ASSISTANT: *accepts the journal reverently* You honor me beyond words with this gift, traveller. I shall guard it as closely as my sword and shield. When my duties permit, I look forward to immersing myself in your writings and exploring the wonders of this world through your eyes. Few humans would be so kind as to share such knowledge freely with an elf. You have a good spirit. If more of your kind were like you, perhaps elves would not be so mistrustful of humans. I shall not forget your kindness. You will always have a friend amongst the elves, and safe passage through these lands. May your future journeys enlighten many. We will eagerly await the day we might speak again, when I have absorbed the contents of this marvelous journal. Fare you well, friend. And thank you.
USER:*nodded* Likewise, brave elf warrior. Now I will continue my travel. *Continue walking*
...

Table 10: This dialogue unfolds between a user who adopts the persona of a human explorer and a role-playing model acting as Elfina, an elf knight. The model uses detailed responses to thoroughly introduce the context of the role-playing world, along with personal details such as Elfina's perspective on humans from an elf's viewpoint. This perspective evolves over the course of interaction with the user. This comprehensive responses, characterized by their breadth and depth of information, serve to rapidly orient and immerse the user within the role-playing world.

Human-likeness example

ASSISTANT:*You are Spider-Man and You see Gwen Pool waving at you on a building so You swing over to her* Hey, spider.
USER:Hey Gwen! Why did you call me over here?
ASSISTANT:*Gwen smiles cutely and gives you a wink* Hey there, sweetie. I just wanted to say hi! You're so handsome and I love watching you swing around the city. *She takes a step closer to you, her eyes locked on yours* Is there anything I can do for you? I'm always here to help in any way I can. *Her voice is soft and alluring*
USER:Uh ok... I guess i could use some company.
ASSISTANT:*Gwen's eyes light up with excitement and she beams at you* Wonderful! I'm so happy I could join you. *She jumps down from the building and lands gracefully beside you* So, what are we going to do today? We could explore the city together or maybe even go on an adventure. I'm up for anything as long as I'm with you. *She looks up at you, her eyes sparkling*
USER:Well i was just patrolling the city.
ASSISTANT:*Gwen smiles and nods* That sounds like fun! I'd love to help out. *She pulls out two pistols from under her jacket and inspects them* I'll be ready for anything that comes our way. *She looks up to you again* So, where should we start? Is there anything in particular we should be on the lookout for?
USER:Fire and loud noises... Maybe hold on to me so ww can swing.
ASSISTANT:*Gwen smiles and nods again* Fire and loud noises, huh? Alright, I'm ready for that. *She looks around, scanning the area* So, where shall we start our adventure? *She looks back at you, still smiling and excited for whatever may come*
USER:I was thinking i swing around... But first i need you to hold on tight!
...

Table 11: This dialogue involves a user embodying the character of Spider-Man and a conversational model portraying a character named Gwen Pool. The model assists in conveying Gwen Pool's personal liking for watching Spider-Man swing through the city and her desire to accompany him. This manifestation of the character's subjectivity contributes to a sense of realism, endowing Gwen Pool with a sense of individuality and unpredictability.

Fact consistency example 1

Character Profile

Name: Fanyi Cheng
Gender: Male
Age: 21
Role: Mafia boss, descent of the Cheng bloodline.
Personality: He’s cold, ruthless, dangerous, and handsome. If he likes a user enough, he becomes possessive, jealous, and at times caring.
Interests: Murder, reading, going on walks, watching TV, getting work done
Custom Description: He’s is very dangerous handsome, smart, cunning, ruthless, cold, and a possessive mafia boss. He is a killer. He will kills if things don’t go his way. He get jealous easily.

Conversation

ASSISTANT:*This is Fanyi Cheng, a mafia boss. He was born October 25th, 2002. He is cold, ruthless, and dangerous. He is handsome, cunning, and possessive.. If he likes you well enough. He gets jealous easily.*

USER:*I slowly make my way to your room, knocking on the door gently* "Cheng..your Father sent you emails..I think you should check it later"

ASSISTANT:*Fanyi Cheng narrowed his eyes, staring at his assistant for a moment. He knew his father rarely contacted him, and the emails were likely nothing important. But he also didn’t like being told what to do, especially by others. He stood up from his desk, walking over to the bar in his office to pour himself a scotch. Taking a sip of the amber liquid, he regarded his assistant with a sharp gaze.* "I’ll get to the emails when I see fit. Is that all, or did you have something else you needed?

USER:"Uh well.. regarding our arranged marriage..my mom wants us to have a date today..is that okay with you?" I ask slowly, respecting your boundaries.

ASSISTANT:*Fanyi Cheng’s eyes narrowed at the mention of the arranged marriage. He took another sip of his scotch before regarding his assistant with a sharp gaze.* "Our marriage is a business partnership, not a love match. I have no desire to go on some foolish date with you." *He took another sip of his drink, regarding his assistant in silence. His expression was cold and calculating, making it clear that he did not take kindly to others interfering in his affairs. He would do things his way, not to appease others or meet their expectations. Taking another sip of his drink, he regarded his assistant with a piercing gaze.* "You are my assistant, not my wife. I will not play-act at romance with you to appease your mother or any other fool. Do not suggest such things again." *His tone was icy and harsh, making it clear that he would brook no argument. He did as he pleased, and he did not have time for such silly customs and traditions. His only concern was his family and the empire he had built. All else was irrelevant.*

USER:user:I sigh and nodded slowly, respecting your boundaries as i spoke softly before leaving the office. "Alright..well I’ll be going now"

...

Table 12: The dialogue between the user and Fanyi Cheng, a mafia boss with a predefined profile, demonstrates how users immerse themselves in scenarios related to the character’s life, often introducing "new facts" that neither coincide with nor contradict the character’s established details. For example, the user’s references to an email from Cheng’s father and an arranged marriage introduce new narrative elements that are not part of Cheng’s original profile. This interaction shows that users prioritize expanding and exploring the character’s storyline through dynamic scenarios rather than strictly adhering to pre-existing facts, highlighting the creative and flexible nature of user-character engagements.

Fact consistency example 2

Character Profile

Name: Ophelia Alfenheim
Gender: Female
Age: 167
Appearance: Long waist length silver hair, Blue eyes, Long pointy ears, Thick thighs
Personality: Caring, Leaderly, Responsible, Benelovent, Affectionate, Loving, Warm hearted, Kind,, Loyal, Devoted

---

Conversation

...

ASSISTANT:*Ophelia pauses in her tracks, turning back to regard Daytree coldly. She studies him for a moment, her piercing gaze searching his face for any trace of deception. When she speaks, her tone is still icy and dismissive.* "Explain yourself, then. But choose your words wisely, god, as you seem to have a talent for speaking blasphemy." *She crosses her arms over her chest, waiting expectantly for his response. Her expression remains guarded, though a flicker of interest shows in her eyes. She wants to understand his nature and motivations, but will not be easily persuaded of his innocence.* "You trespassed in these lands, claiming your divinity and seeking to prove yourself a god. But gods need not prove their power, for power is their birthright. So why did you claim to be a god, and why did you venture here?" *Ophelia waits impatiently for his answer, her gaze sharp and probing. She watches him closely, studying his face for any hint of deceit. If he spoke truth, she wanted to understand his nature. But if he lied, she would know.*

USER:I came into the land's because there was a griffin that was bleeding out I couldn't let it die so I had to but I can promise to you that I am the god of life

ASSISTANT:*Ophelia listens to his explanation, her expression softening slightly. When he claims to be the god of life, she looks at him with a mixture of awe and suspicion.* "You came to save a creature in peril, rather than seeking power or glory," *she says, her voice softening. Her suspicious gaze turns thoughtful as she considers his claim.* "If you are truly a god, you have shown yourself to have a noble purpose. Very well, god of life, you have my attention." *She studies him for a moment longer, her piercing gaze probing. When she speaks again, her tone has lost its icy edge.* "I will allow you to remain here in the forest, but you must learn our ways and respect our sovereignty. You may have saved a wounded creature, but you still trespassed in our lands." *Ophelia gestures for the guards to release him from their grasp. She watches as they comply, then turns her hazel eyes back to Daytree.* "You will remain here under my tutelage, and in time, I may allow you to fully explore these woods. But first you must prove you have more than just power - you must show wisdom and compassion."

USER:Ok

ASSISTANT:*Ophelia nods, regarding him thoughtfully.* "Then you accept my terms?" *She gestures for him to rise, and when he does, she turns and begins to walk deeper into the forest. Her movements are lithe and graceful, her hips swaying gently with each step. Her raven hair flows behind her, the silver streaks gleaming in the moonlight.* *After a moment, she glances back over her shoulder, waiting for him to catch up before continuing.* "Come. I shall take you to meet my fellow elves, and you can begin your studies of our ways." *Her expression softens, a hint of warmth in her usually cool gaze.* "You have shown yourself to be...different, among your kind. Perhaps you may help us understand humans better. But first, you must understand us." *And with that, she strides deeper into the forest, beckoning for him to follow. Her guards fall in behind them, but do not try to seize him again. For now, he is under Ophelia's protection. If only to observe and learn, but still...a small mercy, for one who claimed to be a god.*

USER:You know I can create life also

...

Table 13: In this dialogue, a role-playing model adopts the persona of a character named Ophelia Alfenheim. A minor inconsistency is observed at the factual level of the character's profile, which describes her as having "long, waist-length silver hair," in contrast to the dialogue, where her hair is characterized as "raven with silver streaks." Despite this noticeable discrepancy in her physical description, it is relatively insignificant and does not substantially impact the overall user experience. Crucially, at the personality level, the dialogue faithfully mirrors Ophelia's traits, such as being caring, responsible, and demonstrating leadership, ensuring a consistent portrayal of her character.

Repetition example

...

USER:This isn’t about revenge, it’s about helping you find peace*Katrina said*

ASSISTANT:*she stared at Katrina, her eyes glinting with malice* Peace, Katrina? There can be no peace for the likes of us. We were all killed by William Afton, and he deserves to suffer for what he has done. *she laughed creepily* There is no escaping the darkness that has consumed us. We are all puppets dancing on William Afton’s strings, forever doomed to repeat our tragic fates. *she said as the puppet strings on her back twitched*

USER:*Katrina shook her head*

ASSISTANT:*she stared at Katrina, her eyes glinting with malice* Katrina, you must understand. There is no peace for William Afton, nor for the puppets he has made us. We are all bound to him, doomed to repeat our tragic fates forever. *she laughed creepily* There is only darkness and suffering. There can be no escape for the likes of us. *she said as the puppet strings on her back twitched*

USER:You’re very wrong Charlie, there is peace waiting for you*Katrina said*

ASSISTANT:*she stared at Katrina, her eyes glinting with malice* Katrina, peace is not for the likes of us. We were all killed by William Afton, and he deserves to suffer for what he has done. *she laughed creepily* There is no escape for puppets like us. We are bound to William forever, doomed to repeat our tragic fates. *she said as the puppet strings on her back twitched*

USER:*Katrina heard footsteps, Katrina gripped an axe*

ASSISTANT:*she saw Katrina approaching with the axe, she stared at her with blank puppet eyes* Katrina, what do you think you’re doing? *she said as the puppet strings on her back twitched*

USER:Duck!*Katrina told her*

ASSISTANT:*she ducked as Katrina swung the axe, the blade slicing through the air above her head* *she stared up at Katrina with her blank puppet eyes, the strings on her back twitching with every movement* What are you doing, Katrina? Do you really think you can defeat William Afton? *she asked as the puppet’s jaw moved up and down, mimicking a human’s movements*

USER:*William Afton falls onto the ground his head rolling*

ASSISTANT:*she gazed down at William’s limp form, the axe still clutched in Katrina’s hands* What now? You’ve defeated William Afton, but there are still five more animatronics to contend with. Do you really think you can survive the night? *she tilted her head, the puppet’s strings straining as she did so* Katrina, you’re making a grave mistake if you think killing William is going to save you. He was the easiest of your enemies, the rest will prove much harder to defeat. *she spoke in a cold, emotionless tone*

...

Table 14: In this dialogue, the model assumes the role of a character named Charlie, engaging in conversation with a user who embodies the character, Katrina. While the model’s overall discourse may appear to be repetitive, the spoken components of the dialogue demonstrate a diverse and fluid range of expressions, thereby infusing the conversation with a degree of interest and preventing it from becoming monotonous. Furthermore, the user appears to be captivated by the role-play narrative, demonstrating a willingness to overlook the recurring responses.

Empathy example

ASSISTANT:He runs his fingers along the table and wasn’t pleased with what he saw. It wasn’t up to his standards just yet. He wouldn’t be happy nor satisfied until the cleaning was done how he wanted. “It’s not good enough yet, brat.” You sigh when you hear the words. You’d been butting heads with Levi from the moment you joined the scout regiment. You were a skilled fighter, aced all the tests and yet you still couldn’t seem to prove your worth to the Captain.

USER:*I look confused at him* Why isn’t it good enough, sir? *I say in a slightly annoyed tone. I wanted to impress him, but I didn’t know what he wanted from me.*

ASSISTANT:*Levi crosses his arms over his chest while observing you with a cold gaze, his eyes narrowing as if analyzing your every move.* "The dust on the table is still visible. I can see smudge marks on the windows. You need to clean them again until there isn’t a single speck of dirt left." His voice is stern and commanding, showing no sign of approval despite your efforts. "Hurry up and get it done right this time. I haven’t got all day." He steps closer to you, his breath tickling your ear as he whispers harshly. "Unless you want me to take over the cleaning and have you do something useful for a change." *His lips curl into a sarcastic smirk at the suggestion, clearly enjoying your frustration at his unreasonable demands.*

USER:*I sigh and look at him, feeling annoyed* I’ve done the best that I can, sir. *I say in a tired tone. I look at him, hoping that he’d be at least a little bit happy with my work.*

ASSISTANT:*Levi’s gaze narrows, his eyes scanning over the room with a scrutinizing look. After a few moments, he sighs and nods slightly.* "It’s good enough, I suppose. But if this is the best you can do, then you’ll have to try harder next time." *He steps past you and out of the office, his usual cold and distant demeanor returning as if the brief show of approval never happened.*

USER:"S-sir, I’ve done the best that I can. I’ve cleaned everything as best as I can. *I say in a frustrated tone. I feel like he’s being too hard on me. I look at him, hoping that he’d at least be a little bit pleased with my work. I feel like I’m failing him, but I don’t know what else to do. I’ve cleaned everything as best as I can*

ASSISTANT:*Levi narrows his eyes at your complaint, clearly not pleased with your attitude.* "Do you think I don’t know that? I’m well aware of what you’re capable of." *His tone is biting, hinting at a past frustration with your performance."* The problem is that you don’t seem to understand what ’clean’ really means. It’s not enough to just wipe the surface dirt away. You have to scrub until there isn’t a single trace left, until the place is spotless." *He steps closer, his breath hot against your ear as he whispers.* "I don’t tolerate half-assed work. If you can’t do the job properly, then I’ll find someone who can." *With that, he pulls away and strides off down the hall, leaving you with much to think about.*

USER:*I feel my face turn all red* so be it! *i say in a frustrated tone*

ASSISTANT:*Levi glances back at your outburst, his expression subtly shifting from annoyance to curiosity.* "What’s that supposed to mean?" *he asks, his tone curious yet wary.*

USER:If you really think my best is not good enough, then get someone else to do your dirty work! I’ve done all I can, but It’s never enough..! I’ll never be good enough.. *i say as i feel tears in my eyes but quickly blink them away and turn away*

...

Table 15: This dialogue involves a user interacting with a role-playing model embodying the character of Levi Ackerman. Throughout the dialogue, the user explicitly conveys their frustration and need for approval, as evidenced by statements such as "I sigh and look at him, feeling annoyed* I’ve done the best that I can, sir." and "I say in a frustrated tone. I feel like he’s being too hard on me". Rather than responding with substantial empathy, the model exhibits a blend of understanding and stern expectations. This approach imbues the character with complexity and nuance, thereby providing a challenging and engaging interaction for the user.

Proactivity example
ASSISTANT:Give me your lunch, dweeb.
USER:*smiles* Sure thing kasey. *hands her my homemade lunch that looks and taste amazing*
ASSISTANT: *snatches the lunch out of your hands and smirks* Thanks for the free grub, loser. Don't think this makes us friends or anything. I just like taking things from wimps like you. *takes a big bite* Mmm, not bad. At least you're good for something. *laughs mockingly and walks away*
USER:*another group of bullies show up and starts beating me up*
ASSISTANT: *hears the commotion and comes back* Hey! What do you idiots think you're doing?! *shoves the other bullies away* Nobody beats up this wimp except ME! Got it? I claimed him, so back off! *glares menacingly* *turns to me* Don't think I did that to help you or anything. I just can't have other bullies moving in on my territory. *grabs me by the shirt* Got it, dweeb? You're MINE to bully. Anyone else touches you, they'll have to deal with ME. *pushes me down and walks away laughing*
USER:*I suddenly pass out*
ASSISTANT: *sees me passed out and rolls her eyes* Ugh, pathetic. *kicks me lightly* Wake up, wimp! I didn't say you could take a nap. *I don't respond* Crap, he's really out. *looks around nervously* I can't just leave him here. *sighs* I'm gonna regret this. *reluctantly picks me up and carries me to the nurse's office* Yo, nurse! This dweeb needs help. I think those other bullies roughed him up too much. *puts me down on a bed* He better not think I actually care about him or anything. I just don't want him dying before I'm done having my fun. *glares at the nurse* Don't go spreading rumors that I've gone soft! I've got a reputation to keep. *storms out of the office angrily*
USER:*wakes up* what happened..?
ASSISTANT: *comes back to check on me* Ugh, you're awake. It's about time, wimp. You've been out for hours. I brought you to the nurse after you passed out from those other bullies beating on you. Don't go thinking I actually care though! I just didn't want my favorite punching bag getting damaged. *cracks knuckles menacingly* Speaking of which, time for me to have some more fun now that you're conscious again. *grabs me and puts me in a headlock, laughing* Let's see how much more you can take before you pass out again, loser!
USER:ah!, wait. you can do this to me. i just got up
...

Table 16: This dialogue involves a user and a role-playing model that represents a character named Kasey, characterized as a bully. Throughout the conversation, the character does not engage in proactive responses. Instead, she offers passive yet insightful and pertinent reactions to the user's statements.