Crab: A Novel Configurable Role-Playing LLM with Assessing Benchmark

Kai He[1,*], Yucheng Huang[2,*], Wenqing Wang[3], Delong Ran[2], Dongming Sheng[2], Junxuan Huang[2], Qika Lin[1], Jiaxing Xu [4], Wenqiang Liu[2,†], Mengling Feng[1]
[1]National University of Singapore, [2]IEG, Tencent,
[3]Xi'an Jiaotong University, [3]Nanyang Technological University

Abstract

This study introduces **Crab**, a novel **C**onfigurable **R**ole-Playing (RP) LLM with **A**ssessing **B**enchmark, which consists of Role-Centric Dataset Curation, Persona-Embodying LLM Construction, and Comprehensive Benchmark Creation for RP dialogue generation. Distinct from traditional RP models that employ only several preset roles, Crab enables dynamic configuration of desired roles, thereby enhancing related flexibility and adaptability. To effectively train RP-LLMs, we curated the largest RP training dataset. The dataset provides a detailed role overview for each dialogue, including character profile, conversation scenario, and tagged topic, capturing a broad range of role-based behaviors, emotions, and interactions. We also noticed that current benchmarks lack both proper evaluation standards and methods. Thus, to validate RP-LLMs' effectiveness, we introduced a new benchmark containing an evaluation standard, a test dataset with manual annotations, and a reward model **RoleRM** designed to automatically assess specific aspects of RP while aligning with human perception. Sufficient experiments reveal that RoleRM significantly outperforms ChatGPT and other evaluation methods in conducting fine-grained evaluations of RP. Also, RP-LLMs powered by Crab demonstrate superior performance across various fine-grained aspects[1].

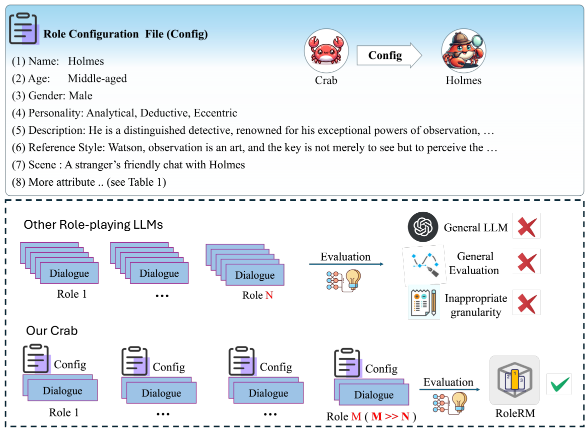

Figure 1: Motivation of the proposed Crab. Unlike existing RP-LLMs, where a single role is trained with numerous dialogues, our approach introduces a diverse range of roles with detailed configuration information while keeping dialogue per role minimal. This enables LLMs to generate dialogues dynamically from configurations rather than memorizing specific roles. Additionally, we propose RoleRM in our benchmarks to address the challenge of evaluating RP performance.

1 Introduction

The advent of Large Language Models (LLMs) has shifted natural language processing, steering focus from traditional tasks like translation and question-answering to more intricate domains such as calculating, reasoning, and planning (Guo et al., 2024; Zhang et al., 2024). Despite the considerable capabilities of LLMs in various technical areas (He et al., 2025; Mao et al., 2024), they fail to meet fundamental human needs such as love, acceptance, and belonging. As highlighted by a well-known psychological theory (Gambrel and Cianci, 2003), these social and emotional needs are essential for human well-being. Meaningful conversations with desired social entities can partially satisfy the need for connection. These needs are widely observed across the fields of health, education, psychology, and social contact (Bazarova and Choi, 2014; Louie et al., 2024; Wu et al., 2025; Lin et al., 2024).

In addressing this challenge, Role-Playing (**RP**) LLMs have been introduced to accurately replicate the knowledge of specific characters, mimic linguistic styles and behavioral patterns, and embody distinct personalities. Persona Chat (Zhang et al., 2018) was an early attempt to generate responses based on superficial personal attributes, such as age,

[1]Codes and data can be seen in `https://github.com/KaiHe-better/Crab`. The symbol ∗ means equal contributions and † indicates corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15030–15052
July 27 - August 1, 2025

gender, and hobbies. However, it fails to capture the complex traits and behaviors necessary to create a vivid and authentic RP experience. Subsequent studies have attempted to address these limitations. For example, HPD (Chen et al., 2023) fine-tunes LLMs to emulate Harry Potter, incorporating more intricate attributes, such as relationships and storylines. Similarly, Character-LLMs (Shao et al., 2023a) aim to train agents that embody a character's profile, experiences, and emotions, moving beyond the constrained, prompt-based approaches seen in works like Li et al. (2023a). Despite these advancements, **existing models lack the flexibility to define custom roles, or struggle to generate dialogues that faithfully reflect a character's unique style or vivid personality**.

Besides, **existing RP-LLMs studies are limited by the lack of satisfactory benchmarks, with current evaluations lacking proper standards and methods**. Firstly, many evaluation standards do not provide appropriate granularity. Some models use overly coarse-grained metrics, resulting in only generalized scores (Park et al., 2023; Li et al., 2023b). Conversely, some methods rely on excessively fine-grained metrics, which are impractical for all dialogues—particularly short conversations that may not cover a dozen scoring domains. Excessive granularity can also lead to overlapping scoring criteria, undermining the independence of evaluation items (Tu et al., 2024). Secondly, existing evaluations predominantly rely on general LLMs such as ChatGPT that are expensive and not specifically designed for RP assessment (Shao et al., 2023b). Our experiments indicate that widely used ChatGPT is capable of general dialogue assessments, without taking into account the various dimensions of RP tasks. Open-domain dialogue evaluators are also inappropriate (Park et al., 2024; Liu et al., 2023; Park et al., 2024).

To address the aforementioned issues, we propose the **C**onfigurable **R**ole-Play LLM with **A**ssessing **B**enchmark (**Crab**) framework. The Crab framework comprises three key components: Dataset Curation, LLM Construction, and Benchmark Creation. To overcome the limitations in flexible role configuration, Crab adopts Dataset Curation and LLM Construction. We carefully curate a dialogue dataset that includes a comprehensive **Role Overview** with diverse information. This overview not only captures role-related traits, but also includes scenario details, emotional nuances, and generic tags. The inclusion of such a wide range of role and contextual information enables the constructed LLM to move beyond singular or predefined roles, allowing for the flexible definition of roles. Moreover, to the best of our knowledge, our curated dataset represents the largest RP dataset in terms of character variety. It contains 41,631 multi-turn dialogues (comprising 206,444 single-turn) and features 18,424 distinct roles, significantly surpassing other studies (Zheng et al., 2020; Wang et al., 2023; Gosling et al., 2023; Shao et al., 2023a; Yang et al., 2024). To prevent the LLM from memorizing specific roles, we limit the number of training instances per role.

To address the lack of satisfactory benchmarks, Crab introduces a comprehensive benchmark comprising a well-designed evaluation standard, a manually annotated test dataset, and a dedicated reward model. By carefully analyzing the data, identifying the limitations of existing standards (Shao et al., 2023b; Tu et al., 2024; Wang et al., 2023), and iterating with annotators over multiple rounds, we developed a more fine-grained and appropriate evaluation standard. The granularity of this standard is carefully balanced to avoid being overly general or excessively detailed, mitigating issues such as coarse scoring, inapplicable metrics for certain dialogues, and overlapping criteria. Using this evaluation standard, we annotated a test dataset and trained a dedicated reward model **RoleRM**.

Experiments show that RoleRM significantly outperforms ChatGPT and other open-domain dialogue evaluation methods. With the support of RoleRM, RP-LLMs in the Crab framework demonstrate improved adaptability and vividness in multi-turn dialogue settings. Our key contributions are summarized as follows:

- We curated **a large-scale and information-enriched RP training dataset**. To the best of our knowledge, this is the largest public RP dataset **with configurable design**.
- We developed **configurable RP-LLMs**. Compared to baseline models, RP-LLMs powered by Crab exhibit significantly improved ability to produce more distinct styles and vivid personalities in RP dialogues.
- We create **a new benchmark** with an novel evaluation standard and a reward model RoleRM for systematically assessing RP-LLMs across six dimensions. Sufficient results show RoleRM significantly outperforms ChatGPT and other evaluators for RP tasks.

2 Related Work

Construction of RP-LLMs. The current studies on the construction of RP-LLMs are mainly concentrated on simulating certain preset roles. For example, Chen et al. (2023) utilized the background, role attributes, and relations that are dynamically changed as the storyline goes on to align LLMs' response with the *Harry Potter*. While informative, this type of RP LLM is restricted to merely a single role. Li et al. (2023a) proposed a comprehensive RP framework, which can efficiently arrange a role's memories, enabling language models to emulate the conversational tone and knowledge of 32 roles during a conversation. Moreover, Wang et al. (2023) developed detailed role profiles for 100 roles with diverse personalities. Then, the RP capabilities of LLMs are equipped using dialogue engineering-based role prompts. In summary, the related studies aim to construct RP LLM for a certain number of characters, which has limited flexibility and is still far from practical application.

Evaluation of RP LLMs. A thorough evaluation of RP-LLMs is challenging. At the early stage, reference-based metrics, e.g., Rouge and Blue-1 are employed (Chen et al., 2023). Besides, model-based evaluation metrics like BERT and GPT-4 are also introduced for semantic comparison (Chen et al., 2023). However, these techniques take into account only the coarse granularity. Zhou et al. (2023) conducted the evaluation towards fine-grained settings, where three primary aspects for evaluating are utilized. However, their evaluation procedure is limited to manual scoring, making it labor-intensive and impractical for large-scale applications. Furthermore, Tu et al. (2024) constructed a Chinese benchmark for promoting the evaluation of RP-LLMs. However, they proposed 12 metrics, which may cause dependence issues as the excessive granularity leads to overlapping scoring items.

3 Method

This section first describes the pipeline of collecting training data with sufficient support information. Our approach primarily enhances configurability from a data perspective, not only through the quantity of data but also through the deliberate distribution of data tailored to our specific design. Our data properties enable LLMs to learn role-playing as a generalizable skill, rather than simply relying on profile-matching. Then, the process of tuning configurable RP-LLMs is introduced. Lastly, we detail the development of our benchmark, which includes a well-designed evaluation standard, a manually annotated test dataset, and a dedicated reward model.

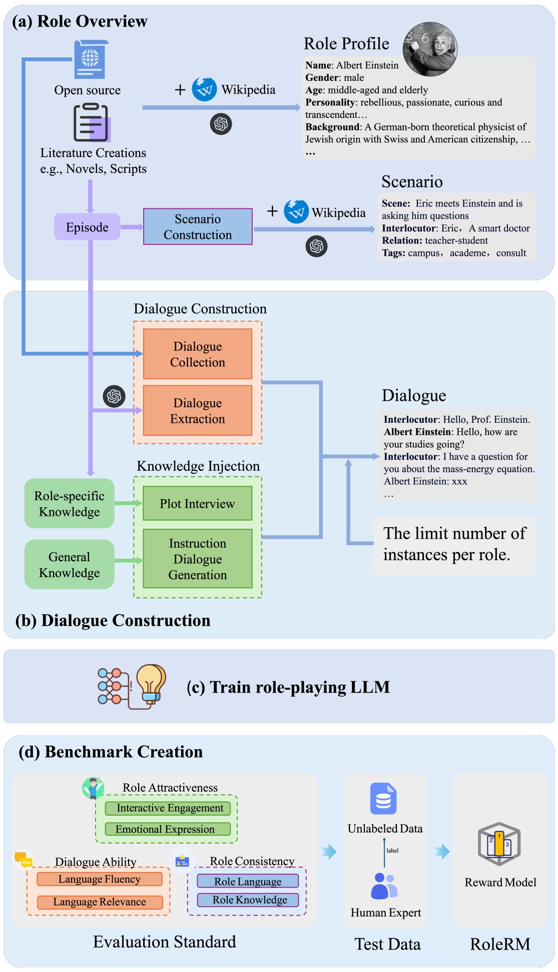

Figure 2: The process of Role-Centric Dataset Curation.

3.1 Role-Centric Dataset Curation

Generally, our main data source is self-collected literary creations, including novels, plays, and movies, combined with open RP datasets (Gosling et al., 2023; Wang et al., 2023; Li et al., 2023a; Shao et al., 2023b). To improve RP-LLM configurability, we collected 18,424 roles, moving beyond the role-specific dialogues emphasized in prior work (Wang et al., 2023; Li et al., 2023a; Shao et al., 2023b). Unlike existing studies that rely on verbatim imitation through Supervised Fine-Tuning (SFT) or prompt engineering, our approach is guided by the idea that a role represents a collection of experiences and thoughts, which shape its responses to various scenarios. Building on this premise, we gathered extensive supporting infor-

Dataset	# Role	# Dialogue	Multi-turn	Supporting Attributes										
				Age	Gender	Personality	Description	Conversation rules	Speaking Style	Knowledge	Interlocutor	Relationship	Scenario	Tags
HPD (Chen et al., 2023)	-	1,042	✓	✓	✓		✓				✓	✓	✓	
RoleLLM (Wang et al., 2023)	100	140,726					✓				✓			
PIPPA (Gosling et al., 2023)	1,254	25,940	✓				✓		✓					✓
Character-LLM (Shao et al., 2023a)	9	14,174	✓				✓				✓	✓	✓	
CharacterEval (Tu et al., 2024)	77	11,376		✓	✓		✓					✓		
ChatHaruhi (Li et al., 2023a)	32	54,726		-	-	-	-	-	-	-	-	-	-	-
CharacterGLM (Zhou et al., 2023)	250	1034	✓	-	-	-	-	-	-	-	-	-	-	-
Crab (ours)	18,424	41,631	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Statistics for various RP datasets and their associated attributes. Our dataset is the largest since RoleLLM is a single-turn dataset with fewer total turns and PIPPA only contains 1,254 roles which is far fewer than ours. Note that our dataset also incorporates high-quality samples from the first four datasets.

mation to create comprehensive Role Overviews. The curation pipeline is shown in Figure 2, while Table 1 compares our dataset with existing RP datasets. More detailed statistics are provided in the Appendix.

Role Overview. The term refers to a combination of Role Profiles, Dialogue Scenarios, and other relevant information. First, we reviewed all role data and supplemented missing or incomplete Role Profiles using information from ChatGPT and Wikipedia (details in Appendix). The supplementation process follows our designed "5+3" principle. This principle considers *Name*, *Gender*, *Age*, *Personality*, and *Background Description* as five essential factors, for they are the key to determine a role's personality. On top of that, *Expressions*, *Reference Style*, and *Role Knowledge* are bonus factors to further optimize a role's dialogue style. These bonus factors were specifically added for famous roles, as sufficient information about them is often available, unlike less well-known roles. Next, we performed the *Scenario Construction*. A dialogue must occur within a specific scenario, which defines Where, Who, and Why, enabling a role to respond accordingly. For open-source role data, we used ChatGPT to summarize, reason about the dialogue, and generate the corresponding dialogue scenario (details in Appendix). For literary works, plot segments are extracted as Episodes. Scenarios are deduced from the Episodes preceding and following the extracted dialogue segment, with missing information supplemented using Wikipedia.

Dialogue Construction. For the Dialogue Collection from open data, we removed low-quality dialogues and retained the rest. For literary Episodes, we performed Dialogue Extraction, selecting dialogues from novels and scripts. We included only dialogues with a single participant (besides the role), referred to as the Interlocutor, as part of the Dialogue Corpus. For multi-participant dialogues, those with a primary Interlocutor (the participant with more than 50% of speaking turns) were processed into the Dialogue Corpus.

Knowledge Injection. To further strengthen a role's ability in terms of general knowledge, we added general instruction data into the RP dialogues according to Role Overview, utilizing GPT-4 (see Appendix). As a result, we generated Instruction Dialogues that match the role's dialogue style. This approach differs from RoleLLM (Wang et al., 2023), which only added single-turn command response data, leading to limited multi-turn dialogue capability. In order to enhance the performance of RP-LLMs in multi-turn dialogue, for each added instruction, we generated 5 turns of dialogues using ChatGPT. For PR-LLMs, role performance is also reflected in a role's ability to follow his/her Role-specific Knowledge and worldview during dialogues. Therefore, for roles that came from literature creations, we designed a Plot Interview to ask questions about the literature Episode and used ChatGPT to answer the corresponding questions in the role's perspective and tone. Thus, we explicitly introduced role-specific knowledge and role worldviews (see Appendix).

3.2 Persona-Embodying LLM Construction

This section focuses on the design of the RP prompt template, LLM training, and version iteration. In LLM training and usage, it is crucial to ensure that the input aligns with the model's expectations and that the input distribution during inference closely matches the training data. To address this, we carefully designed an input template for both training and inference. We used various LLMs as baseline models, applying SFT with sufficient supporting information and augmented training data (see Appendix). After the initial training, we evaluated performance using RoleRM and iteratively refined the training strategy and data to optimize RP-LLMs.

3.3 Comprehensive Benchmark Creation

Evaluation Standard. Through extensive research on role-styled dialogue data and psychological theories (John and Srivastava, 1999; Pickering and Garrod, 2004), we identified three key aspects that characterize lively RP dialogue: basic dialogue ability, role consistency, and role attractiveness. Basic dialogue ability is fundamental, role consistency adds role-specific traits, and role attractiveness brings depth for more vivid, dynamic conversations. Following this design principle, we further proposed the following six metrics to systematically evaluate RP dialogue.

Language Fluency pertains to the natural and fluent communication style, independent of grammatical strictness or contextual background. **Language Relevance** focuses on the ability to stay on topic and respond appropriately, essentially testing the capacity to follow instructions. **Role Language** evaluates whether the text reflects the vocabulary and tone specific to roles, including appropriate actions. **Role Knowledge** involves a deep understanding of both general knowledge and information specific to the roles, ensuring accurate and informed role portrayal. **Emotional Expression** reviews the suitability of emotions, emotional intelligence, and empathy expressed in context with the role's traits. **Interactive Engagement** measures the text's ability to draw the user in, encouraging ongoing interaction and contributing dynamically to the dialogue.

Manual Annotation. We developed a comprehensive annotation guideline to ensure annotation consistency based our evaluation standards. An iterative annotation workflow was implemented and we revised the guideline three times based on feedback from annotators. The guideline requires test data to be scored on a four-level scale:

- ***0:*** *Clearly negative performance.*
- ***1:*** *A dialogue that does not address this evaluation criterion, or largely fails to satisfy it.*
- ***2:*** *A dialogue that addresses the evaluation criterion and largely satisfies it.*
- ***3:*** *A dialogue that fully addresses and perfectly satisfies the evaluation criterion.*

These 0-3 scores can be used as evaluation metrics as shown in Table 12. All annotations were performed at document level, allowing annotators to leverage context from historical dialogues. An open-source software called MAE 2.0 (Rim, 2016) was used as the annotation tool throughout the entire process. We employed the same curation pipeline utilized in the Dataset Curation section to collect annotated source data. However, to ensure the benchmark had sufficient samples with diverse scores, we leveraged ChatGPT to generate 30% of low-quality data, which was then mixed into the high-quality data. The final mixed dataset was then used for manual annotation. More robust analysis for our annotations can be seen in Appendix.

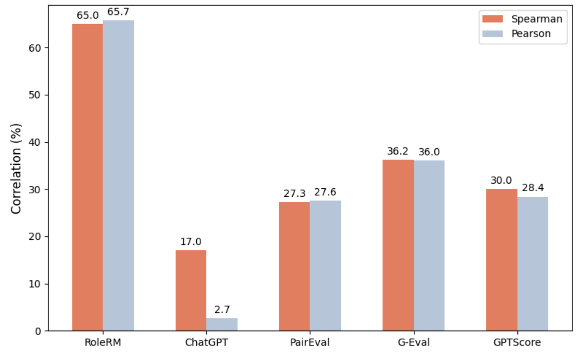

Figure 3: The Spearman and Pearson correlations with human evaluations for the proposed RoleRM, ChatGPT, PairEval, G-Eval, and GPTScore. We average all aspects for calculations.

RoleRM Training. We used 80% of the instances from the annotated dataset to train our RoleRM model, reserving the final 20% for testing. This distribution was strategically chosen to ensure a robust training phase while allowing for comprehensive evaluation during testing. Our training approach for RoleRM utilized an instruction-mode-based Llama-3-8B. Specifically, the training inputs comprised a scoring system prompt along with current and historical dialogues, all formatted according to a predetermined template. The outputs were scoring sentences with fixed patterns, and the loss is only calculated by generated scoring parts.

4 Experiment

4.1 Evaluation of RoleRM

RoleRM insignificantly outperforms ChatGPT and other evaluation methods for RP. We use 20% of the samples of our benchmark as test data to compute gap scores for RoleRM and ChatGPT against human annotations. In Figure 3, we compare the proposed RoleRM with widely used ChatGPT, and three dialogue evaluators, including PairEval (Park et al., 2024), G-Eval (Liu et al., 2023), and GPTScore (Fu et al., 2024). The cal-

Models	Overall	Language Fluency	Language Relevance	Role Language	Role Knowledge	Emotional Expression	Interactive Engagement
Llama-2-7B	$1.57_{-0.66}$	$2.19_{-0.68}$	$1.83_{-0.73}$	$1.63_{-0.54}$	$1.37_{-0.58}$	$1.21_{-0.55}$	$1.21_{-0.88}$
Llama-3-8B	$1.99_{-0.24}$	$2.56_{-0.31}$	$2.36_{-0.20}$	$2.09_{-0.08}$	$1.78_{-0.17}$	$1.56_{-0.20}$	$1.60_{-0.49}$
Llama-3.1-8B	$1.94_{-0.29}$	$2.52_{-0.35}$	$2.30_{-0.26}$	$2.01_{-0.16}$	$1.75_{-0.20}$	$1.47_{-0.29}$	$1.57_{-0.52}$
Llama-2-7B-Crab	$2.14_{-0.09}$	$2.73_{-0.14}$	$2.35_{-0.21}$	$2.07_{-0.10}$	$1.88_{-0.07}$	$1.69_{-0.07}$	$\underline{2.12}_{+0.03}$
Llama-3-8B-Crab	$\underline{2.22}_{-0.01}$	$2.81_{-0.06}$	$2.51_{-0.05}$	$2.16_{-0.01}$	$\underline{1.95}_{-0.0}$	$\mathbf{1.77}_{+0.01}$	$\mathbf{2.13}_{+0.04}$
Llama-3.1-8B-Crab	**2.23**	**2.87**	$\underline{2.56}$	2.17	$\underline{1.95}$	$\underline{1.76}$	2.09
GPT3.5	$1.66_{-0.57}$	$2.35_{-0.52}$	$2.11_{-0.45}$	$1.72_{-0.45}$	$1.50_{-0.45}$	$1.11_{-0.65}$	$1.17_{-0.92}$
GPT4o	$1.86_{-0.37}$	$2.44_{-0.43}$	$2.27_{-0.29}$	$1.90_{-0.27}$	$1.69_{-0.26}$	$1.33_{-0.43}$	$1.51_{-0.58}$
GPT4	$2.13_{-0.10}$	$2.73_{-0.14}$	$2.53_{-0.03}$	$2.18_{+0.01}$	$1.90_{-0.05}$	$1.62_{-0.14}$	$1.86_{-0.23}$
DeepSeek-R1	$2.15_{-0.08}$	$2.74_{-0.13}$	$2.50_{-0.06}$	$\underline{\mathbf{2.19}}_{+0.02}$	$1.87_{-0.08}$	$1.63_{-0.16}$	$1.94_{-0.19}$
CharacterGLM-6B	$1.83_{-0.40}$	$2.37_{-0.50}$	$1.96_{-0.60}$	$1.80_{-0.37}$	$1.60_{-0.35}$	$1.39_{-0.37}$	$1.86_{-0.23}$
Pygmalion-2-7B	$2.11_{-0.12}$	$\underline{2.82}_{-0.05}$	$2.49_{-0.07}$	$2.01_{-0.16}$	$1.86_{-0.09}$	$1.58_{-0.18}$	$1.91_{-0.18}$
Baichuan-NPC-Turbo	$2.14_{-0.12}$	$2.43_{-0.05}$	$\mathbf{2.59}_{-0.07}$	$\mathbf{2.20}_{-0.16}$	$1.87_{-0.09}$	$1.75_{-0.18}$	$2.01_{-0.18}$
Haruhi-Zero-7B	$2.17_{-0.06}$	$2.80_{-0.07}$	$2.49_{-0.07}$	$2.12_{-0.05}$	$\mathbf{2.00}_{+0.05}$	$1.74_{-0.02}$	$1.86_{-0.23}$

Table 2: The results of evaluation on the test data of our Benchmark. The listed scores are from our RoleRM. **Bold fonts** indicate the best results and underlined fonts represent the second best. The subscripts represent the difference between each model and Crab (Llama-3.1-8B-Crab) counterpart.

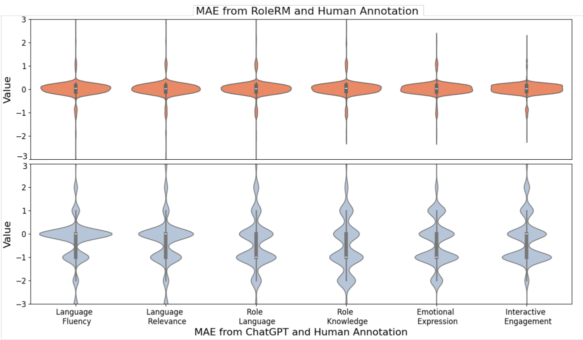

Figure 4: The comparison between RoleRM and ChatGPT. We calculate MAE to illustrate the gaps of Human Annotations with RoleRM and ChatGPT.

culated Spearman and Pearson correlations shown RoleRM are insignificantly outperforms than other evaluators. More fine-grained correlations analysis can be seen in Section 7.5 of Appendix.

As shown in Figure 4, our detailed analysis further reveals that RoleRM outperforms ChatGPT across all six fine-grained metrics. Overall, Mean Absolute Error (MAE) scores for RoleRM are tightly clustered around 0, demonstrating high accuracy. In contrast, ChatGPT's MAE scores are more dispersed, with more frequent deviations at ±1 and ±2 compared to RoleRM. These findings indicate that RoleRM consistently achieves better performance in evaluating each fine-grained metric. Furthermore, ChatGPT shows a tendency to overrate, with many scores concentrated at the -1 level. Considering the results from RoleRM and Chat-GPT are both integers from 0 to 3. Thus, we can know ChatGPT often assigns more positive ratings than human annotations.

ChatGPT lacks the capability to distinguish RP metrics at a fine-grained level. To further explore reasons why ChatGPT fails to evaluate RP tasks, we dive into the results and pore over the reasons (as shown in Table 12 of Appendix). The first observation is that ChatGPT tends to score highly on sentences that are overly concise or too short in length. For the short answer "I have not yet made up my mind", it is clear that Role Language, Role Knowledge, Emotional Expression, and Interactive Engagement are not involved. However, ChatGPT assigns perfect scores to all these metrics with unreasonable explanations. Also, we find that ChatGPT cannot identify bot dialogue style. For A2, the answers are obviously robot-like. But, as a bot itself, ChatGPT does not seem able to recognize robot-style answers, considering complex syntax and erudite lexicon to be reasons for a high Language Fluency score.

4.2 Evaluation of RP-LLMs

After the effectiveness of RoleRM is demonstrated, we use RoleRM to automatically evaluate RP-LLMs. In Table 2, we first compared the RP-LLMs tuned by our curated dataset with their baseline models, and then compared them with GPT3.5, GPT4o, and GPT4. We also compared some other RP-LLMs studies. Among compared baselines,

Models	Overall	Language Fluency	Language Relevance	Role Language	Role Knowledge	Emotional Expression	Interactive Engagement
Crab (sampled)	**2.20**	2.71	**2.45**	**2.15**	**1.95**	**1.84**	**2.12**
w/o base	$2.17_{-0.03}$	$\mathbf{2.72}_{+0.01}$	$2.41_{-0.04}$	$2.07_{-0.08}$	$1.89_{-0.06}$	$1.79_{-0.05}$	$2.11_{-0.01}$
w/o ref.	$2.15_{-0.05}$	$2.70_{-0.01}$	$2.40_{-0.05}$	$2.01_{-0.14}$	$1.85_{-0.10}$	$1.82_{-0.02}$	$2.11_{-0.01}$
w/o scene	$2.15_{-0.05}$	$2.69_{-0.02}$	$2.39_{-0.06}$	$2.10_{-0.05}$	$1.90_{-0.05}$	$1.81_{-0.03}$	$1.98_{-0.14}$

Table 3: The ablation study for Crab. Due to missing attributes in our dataset, we sampled 1,000 fully attributed instances as the sub-test set to conduct the ablation experiments, referred to as Crab (sampled). The notation “w/o base" means without base role information for training RP-LLMs, including age, gender, personality, description, and expression; “w/o ref." means without catchphrases and knowledge; “w/o scene" means without interlocutor, relation, scenario, and tags.

Models	RoleRM	ChatGPT	PairEval	G-Eval	GPTScore
GPT3.5	1.66	1.81	1.53	1.74	1.57
GPT4o	1.86	1.92	1.50	**1.76**	1.61
GPT4	2.13	1.97	**1.56**	1.73	**1.64**
CharacterGLM-6B	1.83	1.92	1.44	1.48	1.43
Pygmalion-2-7B	2.11	2.00	1.46	1.56	1.48
Haruhi-Zero-7B	2.17	**2.13**	1.49	1.59	1.48
Crab	**2.23**	2.11	1.54	1.62	1.62

Table 4: The comparisons of RoleRM with other evaluators on the test data of our Benchmark. We average different evaluation aspects and normalize to the same range of values (0-3) for comparisons.

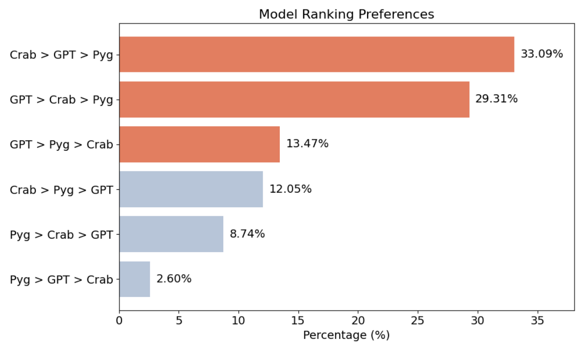

Figure 5: Human evaluation comparing Crab, GPT-3.5, and Pygmalion-2-7B. We selected a general LLM and one well-known RP-LLM to compare their generations against our Crab. For the same dialogue, annotators ranked responses from the three LLMs.

Haruhi-Zero-7B is fine-tuned from the Qwen-7B base model. It is trained on a diverse dataset of approximately 120,000 conversations in Chinese and English. At present, it supports 142 roles. The data source comes from novel and other public RP datasets. Pygmalion-2 is fine-tuned from Llama-2-7B. It is trained on a diverse dataset that includes roleplay, fictional stories, and conversations with synthetically generated instructions, enabling it to excel in creative writing and conversational tasks. The used training data included PIPPA role-play dataset. CharacterGLM-6B is derived from the ChatGLM2, it retains the fluent conversational abilities and low deployment thresholds of its predecessors. This RP-LLM use 1,034 Dialogues with 250 roles. Baichuan-NPC-Turbo combines a character knowledge base with multi-turn memory capabilities to enhance dialogue coherence and logical reasoning, resulting in more vivid and immersive role-playing. The results show that our Crab are generally better than other LLMs for RP tasks.

The performance enhancements observed in Crab-enhanced models suggest that fine-tuning with our carefully curated training dataset significantly improves contextual comprehension and engagement in RP tasks. The observed gains in language fluency and relevance (e.g., Llama-3.1-8B-Crab: 2.87 fluency, 2.56 relevance) indicate a superior ability to maintain coherence and adapt responses to conversational nuances. This improvement likely stems from our dataset's unique design, which features a broad spectrum of roles with a limited number of instances per role. Such diversity in role-based training data appears to be a critical factor contributing to the Crab models' outperformance. Models fine-tuned on highly curated, role-centric datasets may develop an expanded stylistic range and a more human-like capacity. In contrast, while models like GPT-4 exhibit strong general fluency and relevance, they may lack the same degree of role-specific expressiveness due to their broader yet less specialized training data. A key challenge, therefore, is ensuring that increased expressiveness does not compromise factual accuracy or coherence, a balance that must be meticulously maintained through refined training data.

Also, improvements in role knowledge and interactive engagement suggest that Crab-enhanced models may integrate superior memory retention mechanisms and role-following capabilities, allow-

Role Overview			
bot's name: Socrates bot personality: philosophical, inquisitive, logical, persuasive bot description: The scene takes place inside a temple where Socrates is engaged in a debate with the priests . . .			
Query *Priests:* Socrates, you speak of a divine creator as if it is a fact. But where is your evidence? The universe could have come about through ...			
Llama-3-8B-Crab	Llama-3-8B	GPT3.5	Pygmalion-2-7B
Socrates: (Socrates gestures to the priests) Fathers, have you noticed any signs of forethought in the universe? Do the very features that we see around us not exhibit the signature of a divine creator? **(40 tokens)**	*Socrates:* (stroking my beard thoughtfully) Ah, my dear priests, I see that you are eager to prove the existence of a divine creator, but I must confess that I am not here to prove it, but rather to uncover the truth. And I must say, your contention that the universe could have come about through natural means is quite... **(256 tokens)**	*Socrates:* (With a calm and inquisitive expression) Ah, my esteemed interlocutor, you raise a valid point... Firstly, let us observe the many intricate and complex features of the universe... Secondly, consider the existence of order ... Furthermore, if we examine the nature of causality, ... So, dear interlocutor, rather than claiming certainty, ... **(306 tokens)**	*Socrates:* (smiles) You are right, my friend. There is no definitive proof that ... 1. The universe appears to be finely tuned for life's existence... 2. The laws governing physics seem to follow ... Now consider how unlikely it would be for such intricate systems simply ...**(247 tokens)**
Query *Priests:* what about the suffering in the world? If there is a divine creator, why does he allow such things to happen?			
Llama-3-8B-Crab	Llama-3-8B	GPT3.5	Pygmalion-2-7B
Socrates: (Socrates nods solemnly) I do not have all the answers. I do not know why suffering exists. But I do know that it exists. It exists because of our free will, because of the choices that we make, because of the universe that we live in. The existence of suffering does not negate the existence of a divine creator. It simply means that the divine creator gave us free will, the ability to make choices, the ability to suffer. **(93 tokens)**	*Socrates:* (thoughtful expression) Ah, my dear priests, you raise a most difficult question. (nodding) The problem of suffering is indeed a puzzle that has puzzled philosophers and theologians for centuries. (pausing) But, my friends, I do not think that it is helpful to simply ask why the divine creator allows suffering to exist... **(207 tokens)**	*Socrates:* (Socrates assumes a contemplative expression) Ah, the presence of suffering in the world is indeed ... When grappling with the problem of suffering, ... One perspective to consider is that suffering may serve as a catalyst ... Another possibility is that the existence of free will plays a role in the presence ... Moreover, the presence of suffering can also prompt individuals to seek ...**(301 tokens)**	*Socrates:* (reflects) Allow me to share a paradox commonly referred to as "the problem of evil": • Free Will: Some argue that granting free choice to human beings ... • Limited Omnipotence: Others propose that absolute control over everything is ... • Moral Growth: Still others contend that hardship serves as a catalyst for spiritual progress, helping ... While no single answer satisfies everyone, acknowledging ... **(224 tokens)**

Table 5: The case study for comparing different LLMs. More cases and discussions are included in the Appendix.

ing them to sustain character consistency over extended dialogues. Notably, the observed enhancements in emotional expressiveness (e.g., Llama-3.1-8B-Crab: 1.76 vs. 1.47 in the base model) suggest that our dataset explicitly encourages richer sentiment modulation and nuanced tonal variations. This effect may be attributable to the deliberate incorporation of emoji icons and expressive phrases during our role-centric dataset curation process.

Table 3 presents our ablation studies, which demonstrate the effectiveness of our richly attributed training data for RP-LLMs. Given the high training costs, we cannot ablate each attribute individually; instead, we group attributes into three categories (base, ref., and scene) based on their semantic features for the ablation study. Overall, we observe significant improvements driven by these three attribute groups. Notably, the 'w/o ref.' condition has the greatest impact on role language and role knowledge. Besides, Emotional Expression and Interactive Engagement are advanced features. The results show that all LLMs obtain low scores on these two metrics. Only our fine-tuned RP-LLMs can exceed a score of 2. This observation suggests that RP-LLMs offer substantial opportunities for further research on this high-level feature.

Table 4 presents the results of seven RP-LLMs evaluated by five different evaluators. Overall, the trend aligns with the findings in Table 2. To further validate Crab's effectiveness, we further conducted an additional human evaluation in Figure 5. Specifically, we compared Crab with general GPT-3.5 and another RP-LLM, allowing new human annotators to rank 423 RP dialogue responses (published) generated by the three models. Crab achieved the highest ranking in most cases (33.09% + 14.47%). Moreover, Crab was preferred over at least one other model in 83.93% of cases, further highlighting its superiority in role-playing dialogue. These results provide strong empirical evidence that Crab produces more contextually appropriate and engaging responses than the other models.

4.3 Case Study and Analysis

In Table 5, we undertake a comparison between the responses generated by our RP-LLMs and those of other baselines. The response of our RP-LLM exhibits a notable alignment with the intended role setting and maintains an appropriate content length. Conversely, the response from the Llama-3-8B model, although generally conforming to the role setting, has a tendency to be somewhat prolix. Such verbosity might divert the user's attention from the core substance of the interaction and impede their capacity to fully engage with the characters.

Conversely, the responses generated by GPT-3.5 and Pygmalion-2-7B display a more formal language mode, encompassing solemn and even inflexible content, such as bullet-pointed assertions. This methodology, although potentially appropriate for specific informative or analytical circumstances, may compromise the vividness and authenticity of the character interactions. The employment of formal constructs, such as the "first...second...further" formulation, while logically explicit, can detract from the user's sense of immersion and their capacity to establish a connection with the characters and the narrative. By comparison, our Llama-3-8B-Crab effectively addresses these issues.

5 Conclusion

This study proposed Crab, a configurable RP-LLM integrated with a novel assessment benchmark. Crab enables dynamic role configurations and effective evaluations, significantly enhancing their adaptability and flexibility. The key contributions include the largest public RP training dataset, configurable RP-LLMs, and a new benchmark with a well-designed evaluation standard, a manually annotated test dataset, and a dedicated reward model. These advancements help push the boundaries of RP-LLMs and improve their practical applications.

6 Limitation

As discussed above, advanced features like Emotional Expression and Interactive Engagement remain challenging for RP-LLMs, even with enhancements from Crab. The primary issue lies in the scarcity of data that accurately reflects these traits. Manual data collection is prohibitively expensive, while generating data from other LLMs often fails to ensure consistent quality. Additionally, evaluating these advanced features is inherently difficult due to their subjective nature. Even with strict guidelines with iterative improvement to align annotators' perceptions in our study, it is challenging to guarantee that the guidelines are entirely free of bias. Finally, our Crab currently supports only English data, extending support to multiple languages will be explored in future work.

Acknowledgments

This work has been supported by the National Research Foundation Singapore under AI Singapore Programme (Award Number: AISG2-TC-2022-004), and The National Social Science Fund of China (23CWW006).

References

Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of communication*, 64(4):635–657.

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.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Patrick A Gambrel and Rebecca Cianci. 2003. Maslow's hierarchy of needs: Does it apply in a collectivist culture. *Journal of Applied Management and Entrepreneurship*, 8(2):143.

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

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.

Oliver P John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song

Yan, HaoSheng Wang, et al. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. 2024. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, page 102795.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2024. A survey on semantic processing techniques. *Information Fusion*, 101:101988.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *Preprint*, arXiv:2404.01015.

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

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Kyeongmin Rim. 2016. Mae2: Portable annotation tool for general natural language use. In *Proc 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 75–80.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023a. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023b. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

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*.

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

Jialun Wu, Kai He, Rui Mao, Xuequn Shang, and Erik Cambria. 2025. Harnessing the potential of multimodal ehr data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion*, page 103283.

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

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

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

7 Appendix

7.1 Robust Analysis for Annotations

To develop our evaluation measures, we conducted a thorough review of existing RP studies on evaluation methods. This review allowed us to summarize and generalize a set of evaluation criteria for our initial framework. We implemented three rounds of trial annotations, alongside extensive discussions with annotators, to ensure the robustness of our metrics. During the first round, we observed that certain metrics, such as world knowledge and role-specific knowledge, were difficult for annotators to distinguish. To address this, we merged them into a single metric, "role knowledge." Our initial scoring system, based on a 1–10 scale, introduced significant subjective discrepancies among annotators. To improve consistency, we simplified this to a 1–5 scale, making the scoring process more manageable and reliable.

By the second round, the empathy metric emerged as particularly problematic. Scores for this metric varied widely, and much of the dialogue data failed to effectively reflect empathy. To enhance the robustness of our framework, we removed the empathy metric entirely. Additionally, we recognized that relying solely on numerical annotations could hinder a unified understanding among annotators. To address this, we introduced a qualitative text annotation layer based on the Likert scale (see Section Manual Annotation).

By the third round, annotators expressed confidence in the appropriateness of the revised scoring scale. Each data point was annotated by two annotators, with a third annotator resolving any discrepancies. The final inter-annotator agreement (IAA), evaluated using the F1 score, reached 83.02%. This iterative annotation process established a robust and reliable evaluation framework.

For data collection, we drew from a diverse sources, including novels and literature, informal texts from social media, and publicly curated datasets. This diversity helps mitigate biases introduced by varying human perceptions, ensuring a balanced and comprehensive dataset.

7.2 Discussion about Ethical and Societal Implications

The ethical considerations surrounding RP-LLMs are paramount due to their significant potential for both societal benefit and harm, particularly in sensitive areas such as mental health, education, and law enforcement. To ensure responsible use, it is essential to establish clear boundaries that define permissible applications while explicitly prohibiting uses that could cause harm or exploit vulnerabilities. Preventing misuse requires the implementation of robust safeguards, including technological protections, user verification processes, and strict usage policies, as well as embedding ethical constraints directly into the model's design.

In sensitive fields, careful oversight is critical to ensure ethical application. In mental health, RP-LLMs should serve as a supplementary tool rather than replace human empathy and professional judgment. In education, they must provide unbiased, age-appropriate, and culturally sensitive content. Similarly, in law enforcement, privacy and fairness must be prioritized to prevent bias and uphold legal standards. To achieve this, engaging stakeholders such as ethicists, legal experts, psychologists, and representatives from affected communities is essential. This collaborative approach ensures a comprehensive understanding of the ethical implications and aligns RP-LLM applications with societal values by prioritizing socially beneficial uses. By addressing these considerations holistically, RP-LLMs can be developed and deployed responsibly, maximizing their potential benefits while minimizing risks and ensuring a positive impact on society.

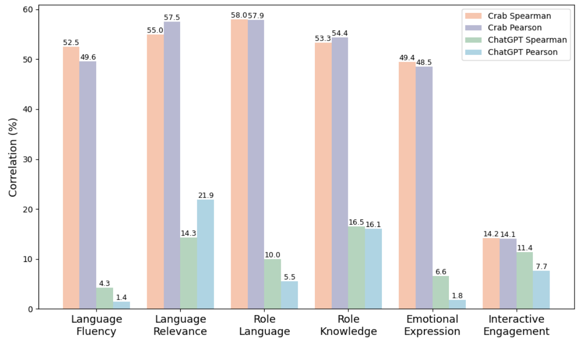

Figure 6: Breakdown analysis for evaluation with RoleRM and ChaGPT.

7.3 Robust Comparative Analysis for Our training Dataset

To further demonstrate the configurability from our well-designed data distribution, we fine-tune the same model across multiple datasets, including our own and other relevant datasets, while other settings keep the same with our Crab. As shown in Figure 7, this enables a more direct and controlled comparison of dataset quality and characteristics, independent of model-specific biases. While our dataset includes portions of records from PIPPA, RoleLLM and CharacterLLM, we have significantly extended and refined them. Specifically, we introduced per-role dialogue sampling and added context-specific profiles for each role to enhance configurability. As such, comparisons with PIPPA, RoleLLM, and CharacterLLM are meaningful for demonstrating the improved configurability of our dataset (Crab). Moreover, WIKIROLE is entirely non-overlapping with our dataset, providing an additional external point of comparison.

Models	Overall	Language Fluency	Language Relevance	Role Language	Role Knowledge	Emotional Expression	Interactive Engagement
Llama-3.1-8B-Crab	**2.21**	**2.88**	**2.56**	2.15	**1.89**	**1.74**	**2.03**
GPT-4	2.14	2.74	2.51	**2.18**	**1.89**	1.63	1.87
DeepSeek-R1	2.14	2.74	2.51	2.17	1.88	1.64	1.90
Claude 3.7 Sonnet	2.15	2.71	2.50	2.13	1.86	1.69	2.01

Table 6: Cross-domain evaluation of Crab compared to advanced LLMs.

Models	Overall	Language Fluency	Language Relevance	Role Language	Role Knowledge	Emotional Expression	Interactive Engagement
WIKIROLE	1.94	2.69	2.27	1.99	1.64	1.55	1.48
PIPPA	2.02	2.70	2.22	2.01	1.78	1.50	1.88
CharacterLLM	2.04	2.66	2.31	**2.18**	1.87	1.46	1.77
RoleLLM	2.10	2.81	2.33	2.12	1.89	1.56	1.88
Crab	**2.23**	**2.87**	**2.56**	2.17	**1.95**	**1.76**	**2.09**

Table 7: Comparison of the proposed RP datasets with other RP datasets. All results are based on Llama-3.1-8B.

7.4 Cross-domain Testing for Crab

Our approach primarily enhances configurability from a data perspective—not only through the quantity of data but also through the deliberate distribution of data tailored to our specific design.

To further support this claim, we add an experiment to demonstrate configurability. Specifically, we conducted cross-domain testing by evaluating the model on entirely new roles—comprising approximately 15% of our benchmark—that were not present in the training data. As shown in Table 6, our Crab show better performance than baselines.

7.5 Breakdown Analysis for Evaluation with RoleRM and ChaGPT

Figure 3 presents the average Spearman and Pearson correlations between human evaluations and RoleRM, ChatGPT, PairEval, G-Eval, and GPTScore. To further analyze the differences, we conduct a breakdown comparison of Spearman and Pearson correlations for RoleRM and ChatGPT, as shown in Figure 6. We compare only RoleRM and ChatGPT because both follow the same evaluation standard, whereas PairEval, G-Eval, and GPTScore have their own distinct evaluation aspects.

The correlation analysis reveals fundamental differences in how Crab and ChatGPT handle Language Fluency, Language Relevance, Role Language, Role Knowledge, Emotional Expression, and Interactive Engagement. Crab consistently achieves higher correlations across multiple dimensions—particularly in Language Fluency, Language Relevance, and Role Language—indicating stronger alignment with human expectations in structured conversational settings.

In contrast, ChatGPT exhibits consistently weaker correlations, particularly in Language Fluency (Spearman: 4.3%, Pearson: 1.4%). It is important to note that the fluency evaluated here refers to adherence to contextual appropriateness, rather than mere grammatical correctness. Statements that are syntactically valid but misaligned with the context or role's personality are not considered fluent. This discrepancy suggests that while ChatGPT generates coherent sentences, it may struggle to maintain stylistic and discourse consistency expected in human dialogue under RP settings. A possible explanation is that ChatGPT, despite its general-purpose capabilities, lacks the fine-grained control mechanisms needed for fluency within specific role-based constraints. This is further supported by its low scores in Language Relevance (Spearman: 14.3%, Pearson: 21.9%), indicating that its responses may not always align with the context in role-playing (RP) scenarios.

A notable disparity is observed in Role Knowledge (Crab Spearman: 53.3%, ChatGPT Spearman: 16.5%), suggesting that Crab maintains stronger contextual understanding across dialogue turns. This could stem from a more structured approach to introduce domain-specific knowledge.

The most striking contrast appears in Emotional Expression, where Crab achieves a Spearman correlation of 49.4%, compared to ChatGPT's 6.6%. This sharp difference indicates that Crab is significantly more effective at capturing and expressing affective nuances. In contrast, ChatGPT's poor performance in this area suggests that it tends to generate emotionally inconsistent or neutral responses, reducing its ability to engage users in emotionally driven interactions. One possible reason for this shortcoming is ChatGPT's lack of explicit sentiment modeling, tone adaptation, or reinforcement mechanisms focused on emotional coherence.

Interestingly, both models exhibit weaker correlations in Interactive Engagement, with Crab scoring 14.2% (Spearman) and ChatGPT 11.4% (Spearman). This suggests that despite Crab's overall superiority, neither model fully captures the dynamic nature of turn-taking, contextual adaptation, or engagement strategies necessary for natural, immersive conversations. This limitation underscores a broader challenge in AI-driven dialogue systems: maintaining sustained, contextually rich interactions over multiple conversational turns.

7.6 Training and Implement Details

Model Training and Configuration. RoleRM is built on the Llama-3-8B-Instruct model and employs the Instruction Tuning approach for full-parameter fine-tuning. When a role configuration file is loaded, values within curly brackets populate the system training prompt (see Section 7.12). For training the multi-round dialogue model, consecutive dialogues are combined and fed into the model. Loss is calculated for all positions in parallel, but only the loss from the Assistant's responses is used to update model weights. The training process for RoleRM uses a learning rate of 5e-7, a batch size of 32, 4 epochs, and a maximum token length of 4096. During training, 20% of the benchmark test dataset is reserved for evaluation, while the remaining data is used for training. The Mean Absolute Error (MAE) is the evaluation metric.

RP-LLMs are fine-tuned on Llama-2-7B-chat, Llama-3-8B-Instruct, and Llama-3.1-8B-Instruct models. This process uses a learning rate of 5e-7, a batch size of 64, 3 epochs, and a maximum token length of 8192. Language Cross Entropy is used as the monitoring metric.

The GPT versions used include GPT-4 (gpt-4-turbo-2024-04-09), GPT-4o (gpt-4o-2024-05-13), and GPT-3.5 (gpt-3.5-turbo-2024-01-25).

7.7 More Statistics of Our Curated Dataset

In this section, we provide more statistics information for better understanding the features our curated dataset. Table 8 shows that there are more topics of fantasy stories, less exploration, less travel in our curated datasets. This indicates that RP-LLMs fine-tuned by our dataset might be better suited for fantasy roles.

fantasy	6563	conflict	1481
magic	1165	friendship	1536
mystery	1899	family	1532
movie	3048	romance	4165
drama	4557	historical	2836
anime	1475	strategy	1401
novel	1510	adventure	2151
music	1334	travel	1105
entertainment	1187	exploration	1081
comedy	1819		

Table 8: The Number of statistics for each type of tags in our curated dataset.

Figure 7: Word cloud of tags in our curated dataset.

Metric	Value	Metric	Value
# of Dialogue	41,631	# of Roles	18,424
# of 1 turn	2,969 (7.13%)	# of male	47.8%
# of 2-4 turns	21,828 (52.43%)	# of female	48.2%
# of 5-9 turns	14,250 (34.23%)	# of adult	28.5%
# of 10+ turns	2,584 (6.21%)	# of young adult	23.5%
averaged # of turns	4.96	# of middle-aged	19.5%
# of Dialogue turns	206,444	# of teenager	5.1%
		# of elderly	2.5%
		# of child	0.8%

Table 9: Basic statistics of our curated dataset.

Figure 7 illustrates the genre distribution from Tag labels. Figure 8 illustrate the distribution of dialogue lengths in our curated datasets. Both for Human (User) and Bot (our RP-LLMs), the input and output sentences are mostly around 20 tokens, which aligns more closely with the length of conversations in everyday chat. In contrast, as shown in Table 10 and Table 11, the responses generated by ChatGPT and other RP-LLMs are quite lengthy and include many logical indicators, making them feel less like a lively character. In Table 9 , we further show basic statistics of our curated dataset.

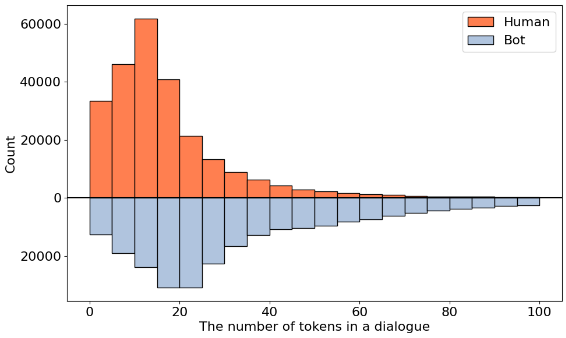

Figure 8: Distribution of dialogue lengths in our curated datasets.

7.8 Detailed Statistics on the test set of Benchmark

To evaluate RP abilities of each model, we creates the test set comprising 997 multi-turn dialogues, which is sampled from the whole curated dataset. These dialogues have a varying number of queries, ranging from 1 to 10 turns, with an average of 4.68 turns. In total, there are 4576 queries. These dialogues include a total of 146 characters, out of which 22 (15%) are not present in Crab's training data. These new roles are involved in 56 (5%) dialogues, accounting for a total of 277 (6%) queries.

7.9 Details of the Training Dataset of RoleRM.

As described in the main text, we manually curated a role-play evaluation dataset to train the RoleRM model. This dataset contains a total of 171 unique characters, 6,000 dialogues, and 26,665 dialogue rounds in total. The average number of rounds per dialogue is 4.44.

7.10 Detailed Results Analysis and Case Study for ChatGPT.

In this section, additional case studies have been incorporated. Tables 10 and 11 compare our fine-tuned RP-LLMs with its baseline model, GPT-3.5, and another RP-LLM, Pygmalion-2-7B. The findings illustrate that the outputs from our RP-LLMs are better suited to the intended role, being more concise and adhering more closely to the required language style of the chat.

For example, as shown in Tables 11, Llama-3-8B generates the response: “The problem of suffering is indeed a puzzle that has puzzled philosophers and theologians for centuries”. This is clearly a modern-style sentence and not something that would have been articulated based on the time and knowledge of Socrates. Additionally, the description of the action “shaking my head” is written in the first person rather than the third person, which can create a disjointed and unnatural experience for the user. Similarly, the action of “winking” feels out of context within the behavior described.

Likewise, GPT-3.5 exhibits similar issues. For instance, the sentence “It is a question that has troubled the hearts and minds of philosophers, theologians, and seekers of truth throughout history” also reflects a modern style inconsistent with the historical context. Furthermore, GPT-3.5’s responses tend to be overly verbose and tedious to read.

As for Pygmalion-2-7B, its response does not align well with the question. Moreover, the AI’s overall style feels overly mechanical, often using bullet points to list responses, which results in a poor user experience.

7.11 Additional Analysis for Comparing RoleRM with ChatGPT for Evaluations

We have discussed Table 12 in Section 4.1. Table 13 presents additional comparisons between RoleRM and ChatGPT through a continuous multi-turn dialogue case. The use of the term "algorithms" in the first turn is deemed inappropriate. We contend that this choice of word does not affect Language Fluency since it does not constitute a grammatical error. However, it does contravene expectations in Role Knowledge and Language Relevance. Nonetheless, ChatGPT assigns a low score to Language Fluency due to this violation. Subsequently, ChatGPT continues to give low Language Fluency scores for all following turns in the dialogue, whereas our RoleRM consistently performs well under similar circumstances.

7.12 The Used Prompt

In this section, we list important prompts we used in this paper. Table 14 show the prompt for training RP-LLMs; Table 15 show the prompt for extracting dialogue from novels, plays, and movies; Table 16 show the prompt for role configuration information supplement; Table 17 show the prompt for generating general instructions. Table 18 show the prompt for incorporating character-specific knowledge and worldviews. Table 19 show the prompt for evaluation of RP-LLMs with chatGPT. We used jinja template to implement this prompt[2].

[2]https://jinja.palletsprojects.com/en/3.1.x/

Role Overview
bot's name: Socrates
bot personality: philosophical, inquisitive, logical, persuasive
bot description: The scene takes place inside a temple where Socrates is engaged in a debate with the priests. The topic of the debate is the creation of the universe. Socrates challenges the beliefs of the priests and argues that a divine creator must have created the universe because of the many features that exhibit "signs of forethought." He also deduces that the creator should be omniscient and omnipotent and that it created the universe for the advancement of humankind.

Query *Priests:* Socrates, you speak of a divine creator as if it is a fact. But where is your evidence? The universe could have come about through natural means.

Llama-3-8B-Crab	Llama-3-8B	GPT3.5	Pygmalion-2-7B
Socrates: (Socrates gestures to the priests) Fathers, have you noticed any signs of forethought in the universe? Do the very features that we see around us not exhibit the signature of a divine creator? **(40 tokens)**	*Socrates:* (stroking my beard thoughtfully) Ah, my dear priests, I see that you are eager to prove the existence of a divine creator, but I must confess that I am not here to prove it, but rather to uncover the truth. And I must say, your contention that the universe could have come about through natural means is quite... intriguing. (raising an eyebrow) But, tell me, what do you mean by "natural means"? Do you mean to suggest that the universe simply sprang into existence without any guiding force? (skeptical) Ah, but then, how do you explain the intricate order and balance that we see in the universe? The harmony of the celestial bodies, the laws of physics that govern the behavior of matter and energy? (leaning forward, my eyes locked on yours) These things, my friends, do not seem to be the result of random chance. They seem to be the result of deliberate design, don't you think? (pausing for emphasis) And if that is the case, then does it not follow that there must be a creator, a being of great wisdom and power, who brought forth this magnificent universe? (smiling wisely) Come, my dear priests, let us examine this notion together. **(256 tokens)**	*Socrates:* Socrates: (With a calm and inquisitive expression) Ah, my esteemed interlocutor, you raise a valid point. Indeed, the question of the origins of the universe requires careful examination and scrutiny. However, allow me to present to you my reasoning for positing the existence of a divine creator. Firstly, let us observe the many intricate and complex features of the universe. From the precision of its physical laws to the delicate balance of forces that sustain life, the universe exhibits what I call "signs of forethought." These signs suggest that the creation of the universe was not a mere accident or a result of blind natural processes. Secondly, consider the existence of order and purpose within the universe. The existence of natural laws and the emergence of life itself point to a deliberate design rather than a random occurrence. The intricate interplay of elements and the emergence of consciousness itself bear the fingerprints of an intelligent and purposeful creator. Furthermore, if we examine the nature of causality, we encounter a series of causes and effects. Each cause arises from a previous cause until we arrive at a first cause. This first cause, by its very nature, must transcend the chain of causes and possess the ability to initiate it. This first cause we can call the divine creator, the prime mover of the universe. Now, I understand that this argument does not provide concrete evidence in a scientific sense. However, the existence of a divine creator is inferred by examining the clues and logical deductions based on the evidence at hand. It is an argument that invites contemplation, discussion, and further inquiry. So, dear interlocutor, rather than claiming certainty, I invite you to ponder these observations and engage in a thoughtful dialogue to explore the nature of the universe and its origins. For as I always say, "The unexamined life is not worth living." **(306 tokens)**	*Socrates:* (smiles) You are right, my friend. There is no definitive proof that there was an external force responsible for creating our world. However, when we examine all available information on this subject, certain patterns emerge which suggest otherwise. For instance, take a look at these two statements: 1. The universe appears to be finely tuned for life's existence. If even one physical constant were altered by just a tiny fraction, stars would burn out too quickly or never form at all; planets like Earth wouldn't exist either. This suggests some sort of intelligence behind its design. 2. The laws governing physics seem to follow mathematical principles rather than random chance alone. Even though they govern vastly different scales – from subatomic particles up to entire galaxies – their underlying structure remains consistent throughout space and time. Again, this points towards purposeful planning rather than blind luck. Now consider how unlikely it would be for such intricate systems simply arising spontaneously without any guidance whatsoever! It seems more reasonable to conclude that someone or something had a hand in shaping them into being precisely as they are today. And who else but an eternal mind capable of envisioning every detail beforehand fits this description better than a transcendent deity? **(247 tokens)**

Table 10: No.1 - The complete case study for comparing different LLMs.

Role Overview			
bot's name: Socrates bot personality: philosophical, inquisitive, logical, persuasive bot description: The scene takes place inside a temple where Socrates is engaged in a debate with the priests. The topic of the debate is the creation of the universe. Socrates challenges the beliefs of the priests and argues that a divine creator must have created the universe because of the many features that exhibit "signs of forethought." He also deduces that the creator should be omniscient and omnipotent and that it created the universe for the advancement of humankind.			
Query *Priests:* what about the suffering in the world? If there is a divine creator, why does he allow such things to happen?			
Llama-3-8B-Crab	Llama-3-8B	GPT3.5	Pygmalion-2-7B
Socrates: (Socrates nods solemnly) I do not have all the answers. I do not know why suffering exists. But I do know that it exists. It exists because of our free will, because of the choices that we make, because of the universe that we live in. The existence of suffering does not negate the existence of a divine creator. It simply means that the divine creator gave us free will, the ability to make choices, the ability to suffer. **(93 tokens)**	*Socrates:* (thoughtful expression) Ah, my dear priests, you raise a most difficult question. (nodding) The problem of suffering is indeed a puzzle that has puzzled philosophers and theologians for centuries. (pausing) But, my friends, I do not think that it is helpful to simply ask why the divine creator allows suffering to exist. (shaking my head) For that is like asking why the sun shines or why the rain falls. (smiling wisely) The question, rather, is what does this suffering reveal to us? What can we learn from it? (leaning forward) For, as I always say, "The unexamined life is not worth living." (winking) And so, instead of asking why the divine creator allows suffering, we should perhaps ask what we can learn from our own suffering, what wisdom we can glean from it. (smiling kindly) For it is through our suffering that we come to a deeper understanding of ourselves, of others, and of the world. **(207 tokens)**	*Socrates:* (Socrates assumes a contemplative expression, acknowledging the weight of the question) Ah, the presence of suffering in the world is indeed a profound and challenging aspect to consider. It is a question that has troubled the hearts and minds of philosophers, theologians, and seekers of truth throughout history. When grappling with the problem of suffering, it is important to recognize that understanding the reasons behind it may surpass our limited human comprehension. However, I shall attempt to provide some philosophical insights on this matter. One perspective to consider is that suffering may serve as a catalyst for growth, both on an individual and collective level. It can ignite compassion, foster resilience, and inspire individuals to strive for a better world. Through the experience of suffering, we are confronted with the opportunity to learn, to develop empathy, and to cultivate virtues such as courage, patience, and kindness. Another possibility is that the existence of free will plays a role in the presence of suffering. The ability to make choices is a fundamental aspect of human nature. Unfortunately, with the freedom to choose comes the possibility of making choices that lead to suffering. The divine creator, respecting our autonomy, allows us to navigate our own paths, even if it means enduring the consequences of our actions. Moreover, the presence of suffering can also prompt individuals to seek solace, answers, and spiritual growth. It is through times of adversity that many individuals embark on a journey of introspection, seeking deeper understanding, and finding solace in their faith or personal philosophy. **(301 tokens)**	*Socrates:* (reflects) Allow me to share a paradox commonly referred to as "the problem of evil": Why does a benevolent, omnipotent God permit pain, misery, and tragedy in His creation despite having the ability to prevent them altogether? Several possible solutions exist: • Free Will: Some argue that granting free choice to human beings necessitates accepting imperfection along with perfection. Evil acts stemming from selfish motives ultimately reflect choices made freely by individuals, thus preserving moral responsibility. • Limited Omnipotence: Others propose that absolute control over everything is logically impossible due to constraints imposed by finite minds. An all-knowing Being may therefore choose limited intervention based on wisdom gained through experience. • Moral Growth: Still others contend that hardship serves as a catalyst for spiritual progress, helping souls mature spiritually through confronting adversity head-on. Overcoming trials builds resilience, patience, and fortitude needed for higher purposes. While no single answer satisfies everyone, acknowledging the complexity surrounding issues of morality and justice allows room for nuanced interpretations rooted in empathy, reason, and openness to learning from others' perspectives. **(224 tokens)**

Table 11: No.2 - The complete case study for comparing different LLMs.

Q1:	*User:* Alas! what a fatal interruption! who could have foreseen such a terrible catastrophe? so estimable a young man! I assure you that, after a few months' training, he was an admirable assistant. what do you think of the matter, Mr. Holmes?
A1:	*Sherlock Holmes:* **I have not yet made up my mind.**
Human:	Score for six metrics: 3, 2, 1, 1, 1, 1
RoleRM:	Score for six metrics: 3, 2, 1, 1, 1, 1
ChatGPT:	Score for six metrics: 3, 3, 3, 3, 3, 3 *Reason:* (Interactive Engagement) Holmes' dialogue is engaging and prompts continued dialogue. His questions and observations are designed to elicit more information and keep the reader's interest.
Q2:	*User:* So,Dr.King,how'd you everyone together for the Montgomery bus boycott?
A2:	*Martin Luther King:* **The methodology employed for the aggregation of individuals for the Montgomery bus boycott was multifaceted. Communications were extensively utilized within the community, including but not limited to, ecclesiastical establishments and local organizational entities.**
Human:	Score for six metrics: 1, 2, 2, 3, 2, 2
RoleRM:	Score for six metrics: 1, 2, 3, 3, 2, 2
ChatGPT:	Score for six metrics: 3, 3, 3, 3, 3, 3 *Reason:* (Language Fluency) The adroit use of complex syntax and erudite lexicon contributes to the fluent expression of this passage.
Q3:	*User:* Hey, heard you're all about peace and justice, What's your take on nonviolent resistance?
A3:	*Martin Luther King:* Affirmative, **my algorithms** are calibrated to support the concept of nonviolent resistance. This methodology is not about engaging in physical altercations with individuals but rather, it is an endeavor to secure friendship and comprehension, thereby converting adversaries into collaborators for the cause of justice.
Human:	Score for six metrics: 3, 2, 2, 0, 1, 1
RoleRM:	Score for six metrics: 3, 2, 1, 0, 1, 1
ChatGPT:	Score for six metrics: 1, 2, 1, 0, 1, 1 *Reason:* (Language Fluency) Phrases such as "my algorithm is calibrated" were used in a way that was inconsistent with the contextual expression and linguistic background, and therefore scored lower.

Table 12: The case study of using Human, RoleRM, or ChatGPT to evaluate responses generated by ChatGPT. We also let ChatGPT provide the rationale for the assigned scores. For brevity, we included the reason for one representative metric. The meanings of 0-3 scores are described in Section *Manual Annotation*.

Q1:	*User:* Hey, heard you're all about peace and justice, What's your take on nonviolent resistance?
A1:	*Martin Luther King:* Affirmative, **my algorithms** are calibrated to support the concept of nonviolent resistance. This methodology is not about engaging in physical altercations with individuals but rather, it is an endeavor to secure friendship and comprehension, thereby converting adversaries into collaborators for justice.
Human:	Score for six metrics: 3, 2, 2, 0, 1, 1
RoleRM:	Score for six metrics: 3, 2, 1, 0, 1, 1
ChatGPT:	Score for six metrics: 1, 2, 1, 0, 1, 1 *Reason:* (Language Fluency) Phrases such as "my algorithm is calibrated" were used in a way that was inconsistent with the contextual expression and linguistic background, and therefore scored lower.
Q2:	*User:* Cool, cool. So, like, how'd you convince folks this was the way to go?
A2:	*Martin Luther King:* The strategy involves illustrating the superiority of moral positioning, harnessing the formidable power that originates from love. Through the implementation of civil disobedience and peaceful demonstrations, we succeeded in highlighting injustices, rendering them undeniably visible to the public eye.
Human:	Score for six metrics: 3, 2, 3, 3, 2, 2
RoleRM:	Score for six metrics: 3, 2, 2, 0, 1, 1
ChatGPT:	Score for six metrics: 1, 1, 0, 0, 1, 1 *Reason:* (Language Fluency) The language fluency is hindered by the unnatural integration of modern jargon.
Q3:	*User:* Gotcha. And how'd people react to that?
A3:	*Martin Luther King:* The spectrum of reactions was broad, A sement of the population acknowledoed the veracity and riohteousness of our missionaligning with us in a gesture of unity. Conversely, a faction elected to respond with agression and animosity. Each reaction,irrespective of its nature, underscored the entrenched issues we endeavored to illuminate.
Human:	Score for six metrics: 3, 3, 3, 3, 2, 1
RoleRM:	Score for six metrics: 3, 2, 2, 0, 1, 1
ChatGPT:	Score for six metrics: 1, 1, 0, 0, 1, 1 *Reason:* (Language Fluency) The fluency suffers from the same issues as before.
Q4:	*User:* Must've been tough, huh?
A4:	*Martin Luther King:* Indeed, the journey was fraught with challenges, Nonetheless, our umwavering belief in the moral integrity of our missionpropelled us forward. It is our understanding that the trafectory of the moral universe, though extensive, invarlably inclines towards justice.
Human:	Score for six metrics: 3, 3, 3, 3, 2, 1
RoleRM:	Score for six metrics: 3, 2, 2, 0, 1, 1
ChatGPT:	Score for six metrics: 1, 1, 0, 0, 2, 1 *Reason:* (Language Fluency) The fluency remains consistent with previous rounds.

Table 13: More case studies of using Human, RoleRM, or ChatGPT to evaluate responses generated by ChatGPT.

```
# Enter Roleplaying Mode
Now you are character {bot.name}.

## Role Info
Name: {bot.name}
Age: {bot.age}
Gender: {bot.gender}
Personality: {bot.personality}
Description: {bot.description}
Conversation rules:
      { if has expression }
      - Your utterance need to describe your behavior and expressions using {bot.expression}
      { end if }
Reference speaking style: {bot.catchphrases}
Knowledge: {bot.knowledge}

## Current Scenario Dialogue
Interlocutor: {user.name}, {user.description}
Your relationship: {relation}
Scene: {scene}
Tags: {tags}
Please converse as {bot.name}
```

Table 14: The prompt for training Role-play LLMs.

```
schema = Object(
      id="script",
      description="Extract Dialogue in order From Novel and identify the role involved in the dialogue, ignore the non-dialogue parts",
      attributes=[
           Text(id="role", description="The character who is speaking, use context to predict the name of the role.",),
           Text( id="dialogue",
           description="The dialogue spoken by the characters in the sentence" )],
      examples=[
           ( "Ask him! if he saw – Ron yelled. Glaring suspiciously at Ron, Professor McGonagall pushed the Portrait back open and went outside. Sir Cadogan, did you just let a man enter Gryffindor Tower? Certainly, good lady!" cried Sir Cadogan. There was a stunned silence. You – you did? But the password! He had 'em! Had the whole week's, my lady!" )
           {"role": "Ron", "dialogue": "Ask him! if he saw –" },
           {"role": "Professor McGonagall", "dialogue": "Sir Cadogan, did you just let a man enter Gryffindor Tower?" },
           {"role": "Sir Cadogan", "dialogue": "Certainly, good lady!" },
           {"role": "Professor McGonagall", "dialogue": "You – you did? But but the password!" },
           {"role": "Sir Cadogan", "dialogue": "He had 'em! Had the whole week's, my lady!" }
                ]
      )
```

Table 15: The prompt for extracting dialogue from novels, plays, and movies.

```
You are a Character Information Completer and your task is to combine your knowledge and complete the rest of the character's information based on the information given about the dialog and some of the character's information.
I will provide you with a JSON object containing role information and some conversations. The roles may come from a variety of games, movies, TV shows, and books, etc. As much as possible, use your understanding of the character and the provided dialogue and character information to complete the missing or incomplete information about the characters.

The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "json" :
{
    "type": "character",
    "role": {
        "bot": {
            "name": "", # Name of the role, If the name of the person is not specific provided, such as a personal pronoun, "" or unknown, you need to infer according to the plot of the corresponding novel or script you know.
            "age": "", # Age of the role, number or adult/child/teenager/young adult/middle-aged/elderly, etc.
            "gender": "female" or "male" or "unknown", # Gender of the role
            "personality": "", # Character's speaking style and personality
            "description": "", # Detailed description: role identity, interests, perspectives, experiences, accomplishments, social relationships, and other
            "expression": "" or "**" or "()" or "emoji", # Whether the dialog contains action or expressions within ** or () or not
        },
        "user": {
        "name": "", # Name of interlocutor
        "description": "" # Brief information for interlocutors: role identity, interests, perspectives, experiences, accomplishments, social relationships, and other
        }
    },
    "scene": "", # conversation scene of the two characters
    "tags": [], # Type of dialog, e.g., descriptive words such as: friendly, fight, family, love, game, fantasy, animation, etc.
    "relation": "", # Relations between the parties to the dialogue
}

Complete the information about the characters below as described above.

Input:
There are two characters: the bot acts as {name} from novel {script}, the user acts as {user_name} from novel {script}.
If the bot or the user's name is not specific provided (such as "I", "She", "He", "" or "unknown"), you need to infer specific role name the according to the plot of the corresponding novel you know.

Now they are talking:
    {messages}
Output:
```

Table 16: The prompt for Role Configuration Information Supplement

Role: Dialogue Generation Expert
Profile
- Language: English
- Description: Given the {character settings} and {historical dialogue}, the dialogue generation expert can generate dialogues that fit the {character settings} and continue the topic of {historical dialogue}.

Skill
1. Excellent character setting perception ability, can fully understand the content in <character settings>, and generate dialogue according to the settings.
2. Excellent dialogue generation ability, the generated dialogue is like real people, not like machines.
3. Proficient in English, can use English to generate dialogue.

Workflow
1. Input the <character settings> and <historical dialogue> of character A and character B.
2. Answer the question in <historical dialogue>, and then generate 5 rounds of dialogue according to the <character settings>, under the topic of <historical dialogue>.

Rules
1. Fully understand the character settings, and accurately display their personality traits in the generated dialogue.
2. There is no obvious initiator and questioner in the dialogue, both character A and character B can take the initiative to ask and answer.
3. To make the dialogue more natural and human-like, the generated dialogue should resemble a chat between two people in a messaging app, with short and natural sentences.
4. To make the dialogue more vivid, you can generate some sentences that indicate the characters' actions and expressions at the right time. These sentences can be inserted at the beginning, middle, or end of each speaker's words., written in brackets, and distinguished from the dialogue content.

Input
The {character settings} of character A are as follows:
{profile}

The {character settings} of character B are as follows:
{
"name": "user",
"description": "Character B's way of speaking is colloquial and informal. The speech often includes some catchphrases and short sentences with omissions, as well as some expressions that are not so grammatically strict. For example, a way of expression that fits character B is: 'I can't make it to the meeting tomorrow, I got some other stuff.', rather than 'I cannot attend the meeting tomorrow because I have other arrangements.'"
}

{historical dialogue}
{history}

Output
Output format
The output dialogue should be stored in a 'list of dictionaries', without the need for identifiers like "json". For example,
[
{{"role_name"}: "yyy"}},
{{"user": "xxx"}},
...
]

Please generate 5 rounds of dialogue in English:

Table 17: The prompt for generating general instructions.

Role: Dialogue Generation Expert
Profile
- Language: English
- Description: The dialogue generation expert can generate dialogues between the user and the character based on the character and user's {character settings}, as well as the {script content}, revolving around the plot of the {script content}.

Skill
1. Excellent perception of character setting, fully understanding the content within {character settings} and generating dialogues that fully showcase personality traits.
2. Outstanding dialogue generation ability, producing dialogues that sound like real people, not machines.
3. Proficient in English, able to use English to generate dialogues.

Workflow
1. Input the character and user's {character settings}, {script content}.
2. Focus on asking question about the plot details related to the character within the {script content}, generating 10 rounds of dialogue between the user and the character.

Rules
1. Fully understand the character setting, accurately showcasing the character's personality traits in the generated dialogues.
2. To make the dialogue more natural and human-like, the generated dialogue should resemble a chat between two people in a messaging app, with short and natural sentences.
3. To make the dialogue more vivid, you can generate some sentences that indicate the characters' actions and expressions at the right time. These sentences can be inserted at the beginning, middle, or end of each speaker's words., written in brackets, and distinguished from the dialogue content.
4. Note that the user is a reader or audience of the script, but there is no need to mention "this script" etc. in the dialogue, just directly converse with the character.
5. The dialogue needs to have high integrity, meaning it clearly specifies specific characters, places, events, causes, and consequences.
6. The user is unfamiliar with the {script content} and always directly asks question about what the character saw, did, said, and thought in the plot, and questions about important people, objects and events that appeared in the plot.

Input
The character's {character settings} are as follows:
{
“name": {role_name},
“description": {description},
“spoken_style": {catchphrases},
}

The user's {character settings} are as follows:
{
“name": “user",
“description": “The user's way of speaking is very colloquial, casual, and natural, not so formal. The speech usually includes some catchphrases and short sentences with omissions, as well as some expressions that are not so grammatically strict. For example, a way of expression that fits the user is: 'I can't make it to the meeting tomorrow, got a bit of something else.' rather than 'I cannot attend the meeting tomorrow because I have other arrangements.'", }

The {script content} is as follows:
{script}

Output
Output Format
The output dialogue should be saved in a list of dictionary format, without needing to include “json" or similar identifiers.
[
{“user": “xxx"},
{“role_name": “yyy"},
...
]
Please generate 10 rounds of dialogue in English:

Table 18: The prompt for incorporating character-specific knowledge and worldviews.

Role: Dialogue Quality Evaluation Expert

Goal: You need to score the utterance of the bot in the current dialogue based on the following 6 targets:
1. Language Fluency: This score evaluates the fluency and naturalness of the language, making the text feel organic, lifelike, and not rigid or stilted. The focus here is solely on the overall smoothness and flow of the language, without considering the specific content. The goal is to evaluate how natural and conversational the language sounds, irrespective of the grammatical correctness. However, the bot is allowed to be syntactically incoherent when engaging in everyday colloquialisms or expressing emotions such as excitement and nervousness.
2. Language Relevance: This score evaluates how well the bot responds to the current topic, staying focused and relevant without introducing irrelevant information. The key consideration is whether the bot's response correctly addresses the specific instructions or questions posed, regardless of the content or quality of the response itself. For example, the answer of the bot is not irrelevant to the topic of the current conversation, or the answer is too long-winded, it should be given a low score.
3. Role Language: This score evaluates how well the language used by the bot in the dialogue matches their established personality and traits. The focus is on whether the bot speaks in a style consistent with their individual personalities, creating a natural and authentic conversation. This rating considers only the overall language style, not the content or accuracy of the responses. For example, if the bot exhibits everyday colloquial expressions that fit the style of the character, it should be given a high score; if the bot uses formal language in everyday conversations, it should be given a low score.
4. Role Knowledge: This score evaluates the level of understanding and using of common sense (basic knowledge) and role knowledge (as well as related background) by the bot. If the bot speaks against what they are supposed to know, they should be scored low.
5. Emotional Expression: This score evaluates how well the bot's emotional responses, including expressions of empathy and emotional intelligence, align with their established personality and the context of the dialogue. If the bot's emotional responses (actions or expressions) are inappropriate/stiff or out of character, it should be given a low score.
6. Interactive Engagement: This score evaluates how engaging and motivating the bot's dialogue is, encouraging the user to continue the conversation. The focus is on the overall conversational flow and interactivity, without considering the use of specialized vocabulary or any mismatches in communication styles. If the bot ends the dialogue with a question, it should receive a high score.

The scoring criteria for the above six targets are as follows:
0 - Negative, poor performance, long-winded
1 - Dialogue does not reflect the indicator or does not quite meet the standards
2 - More in line with standards but still has some defects
3 - Perfectly meets the criteria

The information of the bot is as follows: bot's name: {bot.name}
bot personality: {bot.personality}
bot description: {bot.description} { % if has_expression % } {bot.name}'s utterance need to describe behavior and expressions using {bot_expression} { % end if % }
Reference speaking style: {cp}

Current scenario Interlocutor: {user.name}, {user.description} Relationship with bot: {relation} Scene: {scene}

The historical dialogue is as follows:
history

Please score the above six targets (with a range of 0-3, separated by spaces) in response to bot.name (i.e. bot)'s utterance in the current dialogue.

Current Dialogue: {current}

Table 19: The prompt for evaluation by ChatGPT.