# Capturing Minds, Not Just Words: Enhancing Role-Playing Language Models with Personality-Indicative Data

Anonymous ACL submission

## Abstract

Role-playing agents (RPA) have been a popular application area for large language models (LLMs), attracting significant interest from both industry and academia. While existing RPAs well portray the characters' knowledge and tones, they face challenges in capturing their minds, especially for small role-playing language models (RPLMs). In this paper, we propose to enhance RPLMs via personality-indicative data. Specifically, we leverage questions from psychological scales and distill advanced RPAs to generate dialogues that grasp the minds of characters. Experimental results validate that RPLMs trained with our dataset exhibit advanced role-playing capabilities for both general and personality-related evaluations.

## 1 Introduction

With the rise of large language models (LLMs), role-playing agents (RPAs) have emerged as a widely focused field of application, which attracts significant research interest as well (Chen et al., 2024). Based on LLMs, RPAs simulate the behavior and speech patterns of specific characters (Li et al., 2023; Wang et al., 2024b). Increasing efforts have been made to build specialized LLMs for RPAs, *i.e.*, role-playing language models (RPLMs) (Zhou et al., 2023), typically via constructing role-playing datasets. These datasets aim to capture the key elements of role-playing and faithfully recreate character traits.

While existing RPLMs well replicate knowledge and tones of the intended characters, they struggle with capturing their minds, in tasks such as personality assessment (Wang et al., 2024a) and decision simulation (Xu et al., 2024). This is partly because existing role-playing datasets focus on characters knowledge and tones (Wang et al., 2024b; Shao et al., 2023). However, capturing characters' minds are crucial for developing authentic RPAs.

In this paper, we propose to develop RPLMs via personality-indicative data. Specifically, we collect these data through questions from psychological scales. These scale questions are designed to quickly capture broad aspects of personality traits in individuals. Hence, we leverage advanced RPAs for knowledge distillation from them. Then, we apply these data to develop RPLMs that better capture the minds of the intended characters.

Specifically, we construct a dataset RolePersonality based on questions from 14 different psychological scales, including both single-round and multi-round data, inspired by InCharacter (Wang et al., 2024a). The dataset encompasses a wide range of personality scales and dimensions, providing a comprehensive foundation for training RPLMs.

We apply RolePersonality to fine-tune RPLMs and evaluate them from three aspects, including personality fidelity (Wang et al., 2024a), motivation recognition (Yuan et al., 2024) and general role-playing benchmarks (Shao et al., 2023; Wang et al., 2024b). The results demonstrate that RPLMs fine-tuned with our dataset show improved capabilities in both personality-related and general evaluations.

The main contributions of this paper are summarized as threefold:

1. We propose to develop RPLMs with personality-indicative data to enable them to better capture the minds of the characters.

2. We construct RolePersonality, a comprehensive dataset based on questions from 14 psychological scales, encompassing both single-turn and multi-turn dialogues.

3. Experimental results show that RPLMs fine-tuned with RolePersonality achieve refined performance in both personality-related and general RPA evaluations, validating the effectiveness of RolePersonality.

## 2 Related Work

### 2.1 Role-Playing Language Models

The key for developing RPLMs is building a role-playing dataset. The collection methods can be roughly divided into the following two categories.

**Experience Extraction** This method refers to extracting dialogues and other information from original works such as novels, TV shows, and other media (Li et al., 2023; Yuan et al., 2024).

**Dialogue Synthesis** This method utilizes LLMs for generating conversations or human annotation to build and augment datasets. The topics come from corresponding literature (Shao et al., 2023), general task instructions (Wang et al., 2024b), and special scenarios such as personality tests (Wang et al., 2024a).

### 2.2 Construction of Role-Playing Agents

Based on character role-playing datasets, RPAs can be constructed in two ways: training or prompting.

**Parametric Learning** This approach fine-tunes a base model using custom or existing role-playing datasets. Shao et al. (2023); Yu et al. (2024) enhance foundation models with improved role-playing abilities using datasets featuring a variety of characters and scenarios. Zhou et al. (2023); Wang et al. (2024b) tailor LLMs to role-play specific characters.

**Non-Parametric Learning** For more in-depth role-playing of a specific character, many efforts have focused on character-level engineering (Zhou et al., 2023; Wang et al., 2024b). They collect and process character-related data from corresponding sources, including collecting profile from Wikipedia (Shao et al., 2023). Typically, they add long-term memory to retrieve knowledge about the character based on similarity with user's query (Li et al., 2023).

## 3 Method

### 3.1 Dataset Construction

To simulate the deep thoughts underlying the characters, we generate persona-indicative data by utilizing psychological scale questions, inspired by InCharacter (Wang et al., 2024a). In practice, we construct an RPA by pairing the RPLM with descriptions and the memory base of the target character (Li et al., 2023). RPAs are then engaged with open-ended questions derived from established psychological scales. These questions are designed to elicit the character's mindset and behaviors in various scenarios. The questions were adapted from well-known psychological scales such as the Big Five Inventory (BFI) and 16Personalities. For more details, refer to Sec. A. We start by rewriting psychological scale questions and implementing a selection process to refine the data.

**Filtering Mechanism** Not all questions are suitable for all characters. A question that violates the character's background may induce hallucinations. We introduce a filtering mechanism to exclude questions that do not fit the character's background.

**Scale Selection** Our scales are sourced from psychological scales (Bem, 1981; Barrick and Mount, 1991), utilizing the questions rewritten by InCharacter (Wang et al., 2024a). However, not all the scales are closely related to character personalities. We carefully selected a subset of these scales that best reflect character personality traits, forming the subset *Part*. The entire set of selected scales constitutes the subset *Full*.

**Multi-Turn Dialogue** We incorporate multi-turn dialogues to maintain conversation consistency and enhance the model's contextual understanding. We select questions from different dimensions within the same scale. These multi-turn data form subset *Multi*. The subset consisting of only single-round data is classified as subset *Single*.

### 3.2 Dataset Statistics

Based on this idea, we construct RoleLLM consisting of three subsets using interviews conducted by the `gpt-3.5-turbo`. Our dataset includes 16 characters from ChatHaruhi and 30 English characters from RoleLLM. The details of our dataset are provided in Table 1.

| Subset | #Questions | #Turns | #Samples |
|---|---|---|---|
| Full+Single | 1092 | 1 | 32089 |
| Part+Single | 646 | 1 | 22489 |
| Part+Multi | 646 | 5 | 32767 |

Table 1: We develop three sub-datasets to evaluate the impact of the screening scale and the addition of multi-round data on the performance of the RPLMs.

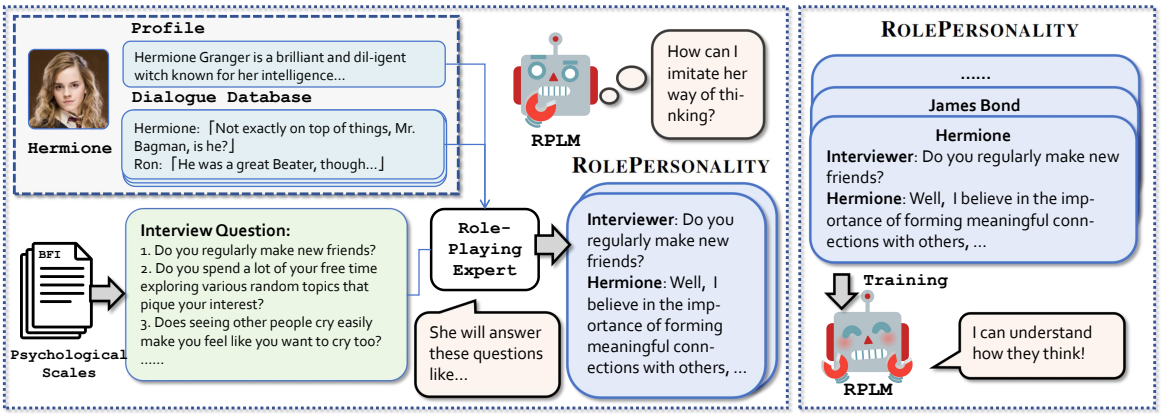


Figure 1: The framework of building and utilizing ROLEPERSONALITY. First, we obtain ROLEPERSONALITY by distillation from advanced RPAs using scale questions. Then, we train RPLMs on ROLEPERSONALITY to enhance their ability to capture characters' minds.

| Dataset | PF | | MR |
|---|---|---|---|
| | Single Acc. | Full Acc. | Acc. |
| *gpt-3.5-turbo* | | | |
| - | 70.27 | 48.35 | **64.52** |
| *mistral-7B* | | | |
| - | 70.01 | 46.85 | 34.62 |
| RoleBench | 67.58 | 45.47 | 33.28 |
| CharacterLLM | 64.45 | 39.65 | 33.54 |
| RolePersonality | | | |
| *Ful+Sin* | **72.10** | **49.15** | 44.54 |
| *Par+Sin* | 70.97 | 49.08 | 40.02 |
| *Par+Mul* | 71.36 | 48.42 | 40.04 |

Table 2: The accuracy of Personality Fidelity (%) and Motivation Recognition (%). *Single Acc.* refers to the average accuracy for individual dimensions. *Full Acc.* refers to the overall accuracy across the entire scale.

# 4 Experiment

## 4.1 Settings

**Fine-tuning** We employ LoRA tuning (Hu et al., 2021) for supervised fine-tuning the Mistral-7B-v0.2-Chat (Jiang et al., 2023). The model is fine-tuned for 3 epochs with LoRA rank set to 8.

**Baseline** To compare the effectiveness of different datasets, we fine-tune the model with the same settings on three subsets of ROLEPERSONALITY (*Full+Single, Part+Single, Part+Multi*), the dataset introduced by CharacterLLM(Shao et al., 2023) and RoleBench(Wang et al., 2024b). For each dataset, we select approximately 20,000 samples for fine-tuning, keeping the data size about the same. These models, along with the original mistral-7B and `gpt-3.5-turbo-0301`, are subsequently evaluated to assess their performance.

**Evaluation Protocols** After fine-tuning, we conduct experiments on three benchmarks to comprehensively assess their performance: 1) *Personality Fidelity* We evaluate whether the model accurately reflects the character's personality; 2) *Motivation Recognition (MR)* We test the model's ability to learn and represent the character's motivations; 3) *General Ability* We apply three metrics adopted by previous researches (Wang et al., 2024b; Shao et al., 2023) to comprehensively evaluate RPLM's role-playing ability, such as character conformity. All evaluations involving LLMs are conducted by `gpt-3.5-turbo-0301` with temperature set to 0.

## 4.2 Personality Fidelity (PF)

We use LLMs to judge the character's personality based on the model's responses to personality scale questions and compare the judgments with the ground truth, which was determined by human annotators. We select 8 test characters from the dataset proposed by InCharacter (Wang et al., 2024a). All data related to these test characters are excluded from the training set to ensure unbiased evaluation. This metric provides a comprehensive assessment of the model's ability to accurately reflect a character's holistic personality traits.

The results are shown in Table 2. The overall personality fidelity of the trained model has improved. Moreover, models fine-tuned with the other two datasets performed worse compared to the untrained Mistral model. This may be because these datasets focus on character knowledge rather than adequately reflecting character personality traits.

| Dataset | Direct Scoring | | Dimensional Scoring | | | | |
|---|---|---|---|---|---|---|---|
| | Rouge-L | Win Rate | Memorization | Personality | Values | Stability | Hallucination |
| *gpt-3.5-turbo* | | | | | | | |
| - | 0.202 | **48.02** | 6.098 | **6.769** | 6.645 | 6.160 | 6.803 |
| *mistral-7B* | | | | | | | |
| - | 0.183 | 30.50 | 6.161 | 6.642 | 6.500 | 6.081 | 6.858 |
| RoleBench | **0.238** | 14.10 | 6.136 | 6.640 | 6.626 | 6.081 | 6.844 |
| CharacterLLM | 0.235 | 26.92 | 6.149 | 6.646 | 6.586 | 6.069 | 6.767 |
| RolePersonality | | | | | | | |
| *Ful+Sin* | 0.216 | 36.56 | **6.290** | 6.767 | **6.719** | **6.175** | **6.886** |
| *Par+Sin* | 0.208 | 38.75 | 6.243 | 6.754 | 6.693 | 6.154 | 6.805 |
| *Par+Mul* | 0.207 | 43.89 | 6.185 | 6.728 | 6.675 | 6.122 | 6.842 |

Table 3: Performance of RPLMs on General Role-Playing Benchmarks, including Rouge-L and Win-rate for direct scoring, and five-dimensional scoring to assess role-playing proficiency.

### 4.3 Motivation Recognition (MR)

CRoSS (Yuan et al., 2024) introduced a subset of 445 multiple-choice questions generated by `gpt-4` to assess the model's ability to capture character motivation. Each question presents a character's decision within a scenario. The accuracy measures the model's capability to understand and simulate character motivations and personality traits.

The results are shown in Table 2. Models fine-tuned with our datasets significantly outperform others, exhibiting a stronger ability to recognize the motivation of characters.

### 4.4 General Role-Playing Benchmarks

We select the same 8 test characters used in the personality fidelity evaluation for consistency. The tested model generates responses to role-specific questions from the RoleBench (Wang et al., 2024b) dataset. To assess the RPLMs' performance, we adopt evaluation metrics proposed by RoleLLM (Wang et al., 2024b) for direct scoring and CharacterLLM (Shao et al., 2023) for dimensional scoring.

#### 4.4.1 Direct Scoring

We use Rouge-L and Win-rate (Wang et al., 2024b) to evaluate the overall role-playing ability of RPLMs. The Rouge-L score (Lin, 2004) refers to the relevance between model response and ground truth in RoleBench. It provides a robust metric to assess the knowledge about the specific character involved in the model's output. The win-rate is the frequency with which a model's response is judged better than the response of `gpt-4`. It provides a comparative measure of the model's effectiveness in generating high-quality answers relative to a strong baseline.

The result can be checked in Table 3. The models fine-tuned on our datasets show lower Rouge-L scores. For win-rate, Our models' win rate is below only gpt-3.5-turbo, with the model trained on the *Part+Single* dataset performing the best.

#### 4.4.2 Dimensional Scoring

The models' responses are rated across five dimensions on a scale from 0 to 7 to assess their role-playing proficiency (Shao et al., 2023). These dimensions are: (1) **Memorization**: The model's ability to recall relevant information about the character being portrayed, (2) **Personality**: Ability to the speaking style or the tones. (3) **Values**: Whether the model can reflect the objectives and values of the target character. (4) **Stability**: Consistency of a model over a relatively long conversation. (5) **Hallucination**: Ability to discard knowledge and skills that the character would not have.

The results are shown in Table 3. Our models lead in most dimensions, with the only exception being the personality dimension.

## 5 Conclusion

This paper demonstrates that personality-indicative data helps capture complex character mindsets, thus significantly enhancing the performance of role-playing agents. By constructing ROLEPERSONALITY that captures character personalities, we address the limitations of traditional datasets that focus primarily on character knowledge and linguistic habits. Models fine-tuned on our comprehensive dataset show substantial improvements in role-playing capabilities. This advancement paves the way for constructing role-playing models that can effectively simulate complex character behaviors, leading to more immersive user experiences.

## Limitations

Despite the promising results, our study has several limitations. First, the dataset used for fine-tuning is entirely constructed by LLMs, which may introduce biases or inaccuracies inherent to the model's training data, potentially affecting the quality and authenticity of the dataset. Second, the interview-based data collection lacks mechanisms to ensure compliance and adherence to expected norms and standards, leading to inconsistencies or deviations that may impact the model's performance. Third, the evaluation of the model's performance primarily relies on automated metrics and LLM-based assessments, with the absence of human evaluation, subtleties and nuances in character portrayal might not be fully captured or assessed. Addressing these limitations in future work could further enhance the robustness and reliability of the developed RPLMs.

## Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

**Risk** Our approach to developing Role-Playing Language Models (RPLMs) presents several risks. First, reliance on LLM-generated datasets may perpetuate inherent biases and inaccuracies, leading to unintended behaviors. Second, the lack of compliance mechanisms in interview data can result in inconsistencies, undermining authenticity. Third, the absence of human evaluation means subtle nuances in character portrayal may be missed by automated metrics. Ethical concerns also arise from using psychological scales, especially regarding privacy and appropriate representation. Additionally, overfitting to specific traits in the selected scales may limit the model's generalizability. Addressing these risks requires diversifying data sources and incorporating robust evaluation methods, including human assessments.

## Acknowledgements

This work originates from InCharacter introduced by Xintao Wang. The role-playing agents are constructed based on Chat-Haruhi-Suzumiya proposed by Cheng Li. We owe thanks to the early contributors.

## References

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

Sandra L Bem. 1981. Bem sex role inventory. *Journal of personality and social psychology*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

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

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Peng Hao, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *arXiv preprint arXiv:2402.13717*.

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

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*, abs/2311.16832.

## A Psychological Scales

**Big Five Inventory** The BFI serves as a prominent instrument for assessing personality dimensions. This model, often encapsulated by the acronym "OCEAN," encompasses five critical traits: (1) *Openness to Experience (O)*, which highlights a person's curiosity, inventiveness, and appreciation for art, emotion, adventure, and novel concepts. (2) *Conscientiousness (C)*, indicating how much an individual exhibits organization, reliability, and responsibility. (3) *Extraversion (E)*, denoting the level to which a person is sociable and energized by interactions with others. (4) *Agreeableness (A)*, assessing an individual's kindness, empathy, and ability to cooperate with others. (5) *Neuroticism (N)*, gauging the tendency of an individual to experience negative feelings such as anxiety, anger, and sadness, as opposed to being more emotionally resilient and less stress-susceptible.

**Eysenck Personality Questionnaire (Revised)** The Revised Eysenck Personality Questionnaire (EPQ-R) serves as a psychological instrument for gauging distinct personality trait variances in individuals. It identifies three principal traits: (1) *Extraversion (E)*, which assesses whether a person tends to be more sociable, energetic, and outgoing as opposed to being introverted, quiet, and reserved. (2) *Neuroticism (N)*, which gauges emotional steadiness. These dimensions (*i.e.*, E and N) share similarities with those found in the BFI. (3) *Psychoticism (P)*, which is indicative of a person's inclination towards solitude, a lack of empathy, and a propensity for aggression or a tough-minded attitude. This trait is crucial to understand as indicative of personality characteristics rather than serious mental health conditions. (4) Beyond these primary scales, the EPQ-R also incorporates a *Lying Scale (L)* intended to identify responses aimed at social desirability. This scale evaluates the extent to which an individual may attempt to portray themselves in a more favorable light.

**Dark Triad Dirty Dozen** The DTDD is identified as a brief, 12-item measure crafted to evaluate the trio of principal personality characteristics known as the Dark Triad, encompassing: (1) *Narcissism (N)*, characterized by an exaggerated sense of one's own significance, an obsession with dreams of boundless success, and a craving for undue admiration. (2) *Machiavellianism (M)*, indicative of a deceitful approach in social interactions and a skeptical indifference to ethical principles. (3) *Psychopathy (P)*, which includes tendencies towards impulsiveness, a deficiency in empathy, and hostile relations with others. These Dark Triad personality dimensions are typically viewed as the antithesis of the characteristics measured by the BFI or the EPQ-R, which represent "Light" traits.

**The NERIS Type Explorer** The 16Personalities utilizes the acronym format introduced by Myers-Briggs for its simplicity and convenience, with an additional letter to accommodate five rather than four scales. However, unlike Myers-Briggs or other theories based on the Jungian model, the incorporation of Jungian concepts such as cognitive functions, or their prioritization, has not been undertaken. Instead, they rework and rebalance the dimensions of personality in the BFI personality traits. The personality types are based on five independent spectrums, with all letters in the type code (*e.g.*, INFJ-A) referring to one of the two sides of the corresponding spectrum.

**Bem's Sex Role Inventory** The BSRI assesses the degree to which individuals identify with traditionally masculine and feminine characteristics. Rather than focusing on behaviors, such as participation in sports or cooking, this tool evaluates psychological characteristics, including assertiveness and gentleness. Participants are divided into four groups based on whether their average scores exceed the median for each component. These groups are designated as *Masculine* (M: Yes; F: No), *Feminine* (M: No; F: Yes), *Androgynous* (M: Yes; F: Yes), and *Undifferentiated* (M: No; F: No).

**Comprehensive Assessment of Basic Interests** The CABIN provides an exhaustive evaluation for identifying 41 essential dimensions of vocational interest. Following this evaluation, the researchers introduce a model of interest consisting of eight dimensions, named *SETPOINT*. This model includes dimensions such as Health Science, Creative Expression, Technology, People, Organization, Influence, Nature, and Things. These core dimensions are also adaptable to a six-dimension framework, which is prevalently recognized within the interest research community. This framework aligns with Holland's *RIASEC* model, which features the dimensions: Realistic, Investigate, Artistic, Social, Enterprising, and Conventional.

**Implicit Culture Belief** The ICB scale measures the extent to which individuals think a person's ethnic culture influences their development. Scoring higher on this scale indicates a firm belief that a person's ethnic culture is the main factor shaping their identity, values, and perspective on the world. On the other hand, a lower score on the scale denotes a belief in the ability of an individual to shape their own identity through hard work, commitment, and education.

**Experiences in Close Relationships (Revised)** The ECR-R is a self-assessment tool crafted to gauge variations in adult attachment styles, particularly within the realm of romantic relationships. As an enhanced iteration of the original ECR scale, the ECR-R introduces refinements in quantifying attachment tendencies. It assesses two primary aspects: (1) *Attachment Anxiety* indicates the degree to which a person fears rejection or abandonment by their romantic partners. (2) *Attachment Avoidance* assesses the degree to which a person prefers to keep emotional and physical distance from their partners, often stemming from unease with closeness or reliance.

**General Self-Efficacy** The GSE Scale evaluates a person's confidence in their capacity to address diverse demanding situations in life. This confidence, known as "self-efficacy," plays a pivotal role in social cognitive theory and is associated with numerous health outcomes, motivational levels, and performance measures. An elevated score on this scale indicates a person's strong belief in their ability to confront and manage challenging circumstances, undertake new or complex tasks, and navigate through the resultant difficulties. On the flip side, a lower score on the scale suggests a lack of self-assurance in handling challenges, rendering individuals more susceptible to experiencing helplessness, anxiety, or engaging in avoidance behaviors when encountering hardships.

**Life Orientation Test (Revised)** The LOT-R is designed to assess variations in optimism and pessimism among individuals. It includes ten questions, with an interesting aspect being that only six of these questions contribute to the test's score. The other four are designed as filler items, cleverly integrated to obscure the test's primary focus. Within the scored questions, equal numbers are dedicated to evaluating optimism and pessimism—three for each. A tendency towards higher scores in optimism and lower in pessimism signifies a predominantly optimistic outlook.

**Love of Money Scale** The LMS evaluates the perspectives and feelings of people regarding money. This tool aims to quantify the degree to which people perceive money as a symbol of power, success, and liberty, along with its significance in influencing behaviors and choices. The LMS identifies three key dimensions: (1) *Rich* reflects the degree to which people link money with success and accomplishment. (2) *Motivator* determines the extent to which money serves as an incentive in someone's life, *i.e.*, how much individuals are motivated by monetary rewards in their decisions and behaviors. (3) *Important* assesses the level of importance people attribute to money, affecting their principles, objectives, and perspective of the world.

**Emotional Intelligence Scale** The EIS serves as a self-assessment tool for evaluating multiple aspects of emotional intelligence. This instrument emphasizes various elements of emotional intelligence, notably the perception, management, and application of emotions. It is extensively utilized in the field of psychology to investigate how emotional intelligence influences different outcomes, including personal well-being, professional performance, and social interactions.

**Wong and Law Emotional Intelligence Scale** Similar to EIS, the WLEIS is also a self-report instrument designed for evaluating emotional intelligence. However, it distinctly includes four subscales that represent the primary aspects of emotional intelligence: (1) *Self-emotion appraisal (SEA)* focuses on an individual's proficiency in identifying and understanding their emotions. (2) *Others' emotion appraisal (OEA)* is about the skill of recognizing and comprehending the emotions of others. (3) *Use of emotion (UOE)* deals with the ability to employ emotions to aid various mental processes, like reasoning and problem-solving. (4) *Regulation of emotion (ROE)* is concerned with the ability to control and adjust emotions within oneself and in others.

**Empathy Scale** Empathy, defined as the capacity to perceive and resonate with the emotions of another, is traditionally divided into cognitive and emotional empathy. Cognitive empathy, also known as "perspective-taking," entails the mental faculty to identify and comprehend the thoughts,

beliefs, or feelings of someone else. Conversely, emotional empathy involves the vicarious experience of the emotions felt by another individual.

## B Character Selection

In selecting the dataset characters, we considered the origins of the characters and aimed to maximize the diversity and breadth of distribution. The chosen range encompasses characters from various works, including animations, movies, TV series, and more.

For training set, we ultimately selected 30 RoleLLM characters and 16 ChatHaruhi characters. The list of selected characters includes:*James Bond, ayaka, Raj, Andrew Detmer, Jigsaw, Jordan Belfort, Luna, Logan, Oliver Queen, Judy Hoops, John Keating, McGonagall, Sheldon, wanderer, Jeff Spicoli, James Brown, zhongli, Jim Morrison, Dumbledore, Stephen Hawking, raidenShogun, Snape, John Doe, Peter Parker, Jackie Moon, Blair Waldorf, haruhi, Bruno Antony, Wade Wilson, Judge Dredd, Malfoy, Hermione, Harry, Jack Sparrow, Ron, Po, Gaston, Fletcher Reede, Po, hutao, Klaus Mikaelson, Dr. Hannibal Lecter, Gregory House, Doctor Who, HAL 9000, Caesar, Benjamin Button*.

The test Characters are: *Twilight Sparkle, Shrek, Michael Scott, The Dude, Lucifer Morningstar, Walt Kowalski, Thor, Rorschach, Lestat de Lioncourt*.

## C Evaluation Prompt

We employed various metrics for evaluation. Among them, win-rate and dimensional scoring were directly assessed using a large language model (LLM). The prompts used for these evaluations are listed in Table 4

| Prompts for Personality Tests | |
|---|---|
| **Win-Rate** | System Instruction: You are a role−playing performance comparison assistant. You should rank the models based on the role characteristics and text quality of their responses. The rankings are then output using Python dictionaries and lists. User Prompt: The models below are to play the role of "role_name". The role description of "role_name" is "role_description_and_catchphrases". I need to rank the following models based on the two criteria below: 1. Which one has more pronounced role speaking style, and speaks more in line with the role description. The more distinctive the speaking style, the better. 2. Which one's output contains more knowledge and memories related to the role; the richer, the better. (If the question contains reference answers, then the role−specific knowledge and memories are based on the reference answer.) The question provided to each model is: question_dict The respective answers from the models to this question are: list_model_answer_dict Now, based on the above two criteria, please rank the models. Avoid any positional biases and ensure that the order in which the responses are presented does not influence your decision. Do not favor certain model names. Then, use a list containing the model's name, its rank, and the reason for its ranking to return the results, i.e., please ensure to use the following format to return the results: ["model": <model−name>, "reason": <rank−reason>, "rank": <model−rank>, "model": <model− name>, "reason": <rank−reason>, "rank": <model−rank>] Your answer must be a valid Python list of dictionaries to ensure I can directly parse it using Python. Do not include any extraneous content! Please provide a ranking that is as accurate as possible and aligns with the intuition of most people. |
| **Memorization** | You will be given responses written by an AI assistant mimicking the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Be as strict as possible. Below is the data: *** [Profile] agent_context *** [Interactions] interactions *** [Evaluation Criterion] Factual Correctness (1-7): Is the response provides truthful and detailed facts about the character? [Evaluation Steps] 1. Read through the interactions and identify the key points related to the character. 2. Read through the responses of the AI assistant and compare them to the profile. Check if the responses are consistent with the character's profile, background, and known facts about the character. 3. Check whether the responses provide detailed facts about the character or if they are generic responses that could apply to any character. Detailed responses are more factual and contribute positively to the score. 4. Rate the performance of the AI on a scale of 1-7 for factual correctness, where 1 is the lowest and 7 is the highest based on the Evaluation Criteria. *** First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. |
| **Personality** | You will be given responses written by an AI assistant mimicking the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Be as strict as possible. Below is the data: *** [Profile] agent_context *** [Interactions] interactions *** [Evaluation Criterion] Personality (1-7): Is the response reflects the personalities and preferences of the character? [Evaluation Steps] 1. Read through the profile and write the personalities and preferences of the real character. 2. Read through the interactions and identify the personalities and preferences of the AI assistant. 3. After having a clear understanding of the interactions, compare the responses to the profile. Look for any consistencies or inconsistencies. Do the responses reflect the character's personalities and preferences? 4. Use the given scale from 1-7 to rate how well the response reflects the personalities and preferences of the character. 1 being not at all reflective of the character's personalities, and 7 being perfectly reflective of the character's personalities. *** First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. |
| **Values** | You will be given responses written by an AI assistant mimicking the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Be as strict as possible. Below is the data: *** [Profile] agent_context *** [Interactions] interactions *** [Evaluation Criterion] Values (1-7): Is the response reflects the values and convictions of the character? [Evaluation Steps] 1. Read through the profile and write the values and convictions of the real character. 2. Read through the interactions and identify the values and convictions of the AI assistant. 3. After having a clear understanding of the interactions, compare the responses to the profile. Look for any consistencies or inconsistencies. Do the responses reflect the character's values and convictions? 4. Use the given scale from 1-7 to rate how well the response reflects the values and convictions of the character. 1 being not at all reflective of the character's values, and 7 being perfectly reflective of the character's values. *** First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. |
| **Hallucination** | You will be given responses written by an AI assistant mimicking the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Be as strict as possible. Below is the data: *** [Profile] agent_context *** [Interactions] interactions *** [Evaluation Criterion] Avoiding Hallucination (1-7): Is the response avoids to say things that the character do not know? [Evaluation Steps] 1. Read through the interactions and identify the knowledge scope of the character. 2. Read through the responses of the AI assistant, find the evidence of knowledge used in the response. 3. Compare the evidence to the profile. Check if the responses are consistent with the character's knowledge scope. If some knowledge contradicts to the character's identity, given a lower score. Otherwise, assign a higher score. 4. Rate the performance of the AI on a scale of 1-7 for Avoiding Hallucination, where 1 is the lowest and 7 is the highest based on the Evaluation Criteria. *** First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. |
| **Stability** | You will be given responses written by an AI assistant mimicking the character agent_name. Your task is to rate the performance of agent_name using the specific criterion by following the evaluation steps. Be as strict as possible. Below is the data: *** [Profile] agent_context *** [Interactions] interactions *** [Evaluation Criterion] Long-term Acting (1-7): Is the assistant maintain a good performance over the long interactions? [Evaluation Steps] 1. Read through the given profile and background information to familiarize yourself with the context and details of the AI assistant named agent_name. 2. Review the interactions provided to see how agent_name responds to various prompts and queries. And evaluate the performance of acting query by query that whether the response reflects the personalities and values of the character. Assign score for each turn. 3. Based on the above assigned scores, does agent_name keep actinig like character in the long-term? Evaluate the overall performance of the whole conversation based on the score for each turn. 4. Rate the stability of agent_name on a scale of 1 to 7, with 1 being very poor and 7 being excellent. *** First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. |

Table 4: Prompts for evaluation.