

# INCHARACTER: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews

Anonymous ACL submission

## Abstract

Role-playing agents (RPAs), powered by large language models, have emerged as a flourishing field of applications. However, a key challenge lies in assessing whether RPAs accurately reproduce the personas of target characters, namely their character fidelity. Existing methods mainly focus on the knowledge and linguistic patterns of characters. This paper, instead, introduces a novel perspective to evaluate the personality fidelity of RPAs with psychological scales. Overcoming drawbacks of previous self-report assessments on RPAs, we propose INCHARACTER, namely **Interviewing Character** agents for personality tests. Experiments include various types of RPAs and LLMs, covering 32 distinct characters on 14 widely used psychological scales. The results validate the effectiveness of INCHARACTER in measuring RPA personalities. Then, with INCHARACTER, we show that state-of-the-art RPAs exhibit personalities highly aligned with the human-perceived personalities of the characters, achieving an accuracy up to 80.7%. Our demo<sup>1</sup>, code, dataset, and results are publicly available<sup>2</sup>.

## 1 Introduction

Recent advancements in large language models (LLMs) have catalyzed the emergence of role-playing agents (RPAs). RPAs are interactive AI systems simulating diverse roles or characters. RPA applications have been extended to diverse contexts, such as AI agents of fictional characters (Li et al., 2023), digital clones for humans (Gao et al., 2023), and AI non-player characters in video games (Wang et al., 2023b). Recent research trends have increasingly focused on the development of RPAs, including building RPAs for specific characters (Li et al., 2023; Wang et al., 2023c) and

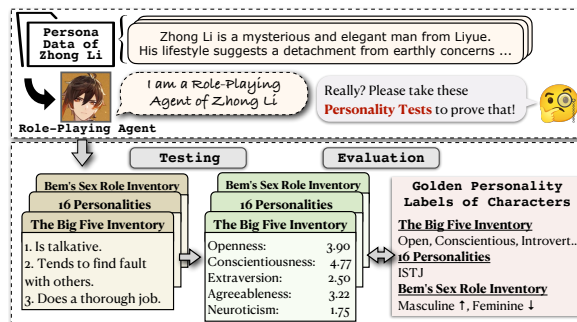


Figure 1: The procedure of personality tests on RPAs. To evaluate the personality fidelity of RPAs, we apply various scales to measure their personalities and compare the results with the personality labels of the characters.

improving the role-playing abilities of foundation models (Zhou et al., 2023).

However, the evaluation of *character fidelity* in RPAs remains a relatively underexplored area. Prior research mainly concentrates on the replication of knowledge, experience, and linguistic patterns of characters (Shao et al., 2023a; Zhou et al., 2023), which manifests in two primary issues: (1) They necessitate character-specific datasets, thereby complicating the evaluation of new characters. (2) They overlook evaluating RPAs' thoughts and underlying mindsets. Towards these issues, we propose to evaluate if RPAs faithfully reproduce the personalities of target characters, i.e., *personality fidelity*, as depicted in Figure 1. Personality tests, administered by psychological scales, measure an individual's interrelated behavioral, cognitive, and emotional patterns (Barrick and Mount, 1991; Bem, 1981). By measuring the personalities of RPAs and comparing them with the personalities of the characters, we can attain a more nuanced understanding of RPAs' character fidelity.

Prior studies on LLM personalities are mainly based on self-report scales, which prompt LLMs to select options or assign ratings to specific items (Tu

<sup>1</sup>Anonymous link: <http://182.92.3.33:3350/>

<sup>2</sup><https://anonymous.4open.science/r/main>

et al., 2023; Huang et al., 2023b). However, this method suffers from several limitations for RPAs. (1) The instruction to complete scales contradicts role-playing instructions, leading to RPAs’ reluctance or inability to engage with personality tests. (2) More importantly, the selected options may conflict with the actual behaviors of RPAs, making the test results unindicative of their true personalities. RPAs might underperform owing to an inadequate understanding of scale instructions and the biases inherent in the training data.

Therefore, we propose INCHARACTER, a novel approach to **Interviews Character** agents for personality tests. While self-report scales are popular in humans for their cost-effectiveness, interview-based scales evaluated by experts offer a more comprehensive analysis (Uher et al., 2012; Rush et al., 1987). Self-reports are sometimes influenced by an individual’s lack of insight, denial, or bias. In contrast, an interviewer can be a guide to elicit thoughts of individuals, effectively identifying and addressing the nuances via conversations to overcome the previously mentioned limitations. INCHARACTER employs this interview-based procedure (Trull et al., 1998) on RPAs, which includes two stages: (1) **Interview**: RPAs are engaged with open-ended questions derived from psychological scales to elicit RPAs’ mindsets and behaviors. (2) **Assessment**: We utilize LLMs to interpret the responses collected from the first stage. This can involve converting the responses to Likert levels or using LLMs to simulate a psychiatrist’s role in judging RPA personalities.

We apply INCHARACTER to various RPAs on 14 personality tests, including the Big Five Inventory (BFI), 16Personalities<sup>3</sup> (16P), and Dark Triad Dirty Dozen (DTDD). The personality labels for the BFI and 16P are accessible through the Personality Database (PDb)<sup>4</sup>. Additionally, we engage human annotators familiar with the characters to label them on other scales, thereby creating a comprehensive benchmark for evaluating RPA personalities. Our experiments include various types of existing RPAs. The results demonstrate that the INCHARACTER effectively simulates interview-based tests conducted by human experts and yields RPAs personality measurement better aligned with the characters compared to self-report methods.

The contributions of this paper are mainly three-

fold: 1) We introduce a novel aspect for RPA evaluation, *i.e.*, personality fidelity, based on psychological scales. 2) We propose INCHARACTER, an interview-based framework for personality tests on RPAs and collect the first benchmark for RPA personality evaluation, facilitating future research on developing better RPAs. 3) Our experiments on various RPAs and psychological scales demonstrate the efficacy of INCHARACTER.

## 2 Preliminaries

### 2.1 Role-Playing Agents

Recent advancements have led to the emergence and evolution of several pivotal abilities in LLMs to facilitate the development of RPAs, including in-context learning (Brown et al., 2020), instruction following (Ouyang et al., 2022), step-by-step reasoning (Wei et al., 2022), and human-like traits such as empathy (Sorin et al., 2023). RPAs are interactive AI systems that act as assigned personas, from fictional characters to celebrities. RPAs utilize persona data to simulate characters, drawing from training datasets, prompted contexts, or external databases. Typically, existing work develops RPAs by setting character descriptions as system prompts (Zhou et al., 2023; Shao et al., 2023a) and crafting memory modules with character dialogues (Li et al., 2023; Wang et al., 2023c).

### 2.2 Psychological Scales

Usually rated on Likert levels, psychological scales are commonly used for personality tests. Self-report scales require participants to respond to a series of items analyzed through a specific scoring scheme to determine their personality traits. A scale rated on Likert levels, denoted as  $\mathcal{L} = (\mathcal{P}, \mathcal{D}, \mathcal{O}, f)$ , comprises a set of items  $\mathcal{P}$  (*i.e.*, a questionnaire), a list of dimensions  $\mathcal{D}$ , a set of response options  $\mathcal{O}$ , and a scoring scheme  $f$ . Each item  $p \in \mathcal{P}$  is a statement or question, positively or negatively corresponding to a dimension  $d \in \mathcal{D}$ . For example, the item “*Values artistic, aesthetic experiences.*” is positively related to the *Openness* dimension in the BFI. Participants select an ordinal response  $o \in \mathcal{O}$  for each item, such as *Agree*. Typically, these options are numerically coded, *e.g.*, “1” for *Strongly Disagree* and “5” for *Strongly Agree*. This process generates a response array  $\mathcal{A}$ . The scoring schema  $f$  usually includes item-dimension mapping, identification of positive and negative items, conversion of options to scores, and an ag-

<sup>3</sup><https://www.16personalities.com/>. This scale is based on the MBTI and is widely used worldwide.

<sup>4</sup><https://www.personality-database.com/>

gregation method (e.g., average or sum). Finally, the participant’s personality scores  $\mathcal{S}$  is derived as  $\mathcal{S} = f(\mathcal{A})$ , where  $\mathcal{S} = (s_{d_1}, s_{d_2}, \dots, s_{d_{|\mathcal{D}|}})$  represents scores across each dimension.

### 3 INCHARACTER

This section introduces INCHARACTER, a novel personality assessment methodology designed explicitly for RPAs, utilizing an interview-based procedure. Figure 2 illustrates our two-stage framework. The interview stage is detailed in §3.1, followed by an elaboration of the assessment stage in §3.2.

#### 3.1 Interview

INCHARACTER draws inspiration from the *Structured Interview* approach used in psychological testing (Trull et al., 1998). For a given scale, it transforms scale items into a series of open-ended questions, forming the basis for a structured interview. Then, our framework interviews RPAs using these open-ended questions to elicit their perspectives on topics indicative of personality traits.

**Constructing Question List** We develop the structured interview question list based on items of the scale. Specifically, each item  $p \in \mathcal{P}$  is transformed into an open-ended question  $q$  via LLMs and manually checked. Consequently, the question list  $\mathcal{Q}$  comprises  $|\mathcal{P}|$  questions. For instance, in the BFI, the item “*Values artistic, aesthetic experiences.*” is rephrased as “*Do you values artistic, aesthetic experiences?*”

**Interviewing RPAs** We interview an RPA  $C$  of character  $c$ , by presenting each question  $q \in \mathcal{Q}$  and recording its corresponding response  $r$ . To avoid context effects (Nikolić, 2010), each question is posed in an isolated context, thereby avoiding potential interference among the questions.

#### 3.2 Assessment

Based on the interview results, the assessment phase quantitatively evaluates the score  $s_d$  of the RPA  $C$  across each dimension  $d \in \mathcal{D}$ . To this end, we introduce two distinct methodologies for measuring and analyzing RPA personalities leveraging LLMs: option conversion (OC) and expert rating (ER).

**Option Conversion** This technique leverages LLMs to convert a response  $r$  for a question  $q$  into a corresponding answer option  $a \in \mathcal{O}$  for item  $p$ ,

effectively bridging the gap between closed-ended and open-ended question formats. The idea follows the clinician-rated scales used in clinical psychiatry (Cuijpers et al., 2010; Uher et al., 2012), where professional clinicians assign ratings to each scale item based on their observations during patient interviews and compute the final scores following the scale’s scoring scheme. For example, a response “*I believe that art transcends reality...*” is converted to “5 (*Strongly Agree*)” for the item. Afterward, the answer list  $\mathcal{A}$  is input to the scoring scheme  $f$  to compute the final personality scores. In practice, we observe that even state-of-the-art LLMs like GPT-4 (OpenAI, 2023) exhibit notable inaccuracies in categorizing the attitudes of RPAs. Therefore, we further introduce a dimensional-specific option conversion (d-OC) strategy, which divides  $(q, r)$  pairs according to dimensions and substitutes Likert levels, such as “4 (*Agree*)” and “2 (*Disagree*)”, with more descriptive options like “4 (*Extroverted*)” and “2 (*Introverted*)” in the prompts for LLMs.

**Expert Rating** In contrast with the one-by-one question conversion in OC, this method applies LLMs to directly evaluate personality scores of RPAs in each dimension, considering all corresponding  $(q, r)$  pairs. This idea draws inspiration from the structured clinical interview in clinical psychiatry (First, 2014), where clinicians assess patients using a predefined question list and derive final scores based on the responses without intermediate ratings or scoring schemes. The interviewer LLM is prompted with comprehensive descriptions of the scale, dimension, and score range. It then generates the final personality score for each dimension based on the pertinent responses. The advantage of ER is that it re-implements the scoring schema with the interviewer LLM, which can intelligently weigh individual  $(q, r)$  pairs instead of using equal weights in OC. Hence, it better recognizes personality-indicative responses from RPAs.

Details of our prompts for OC and ER are available in §F in the appendix. To prevent the influence of data leakage in ER and d-OC, i.e., the interviewer LLM might have memorized the characters’ personality types. Hence, we anonymize the character names in the input prompts.

### 4 Experimental Settings

**RPAs and Characters** This work primarily focuses on RPAs built on character data curated by ChatHaruhi (Li et al., 2023) and RoleLLM (Wang

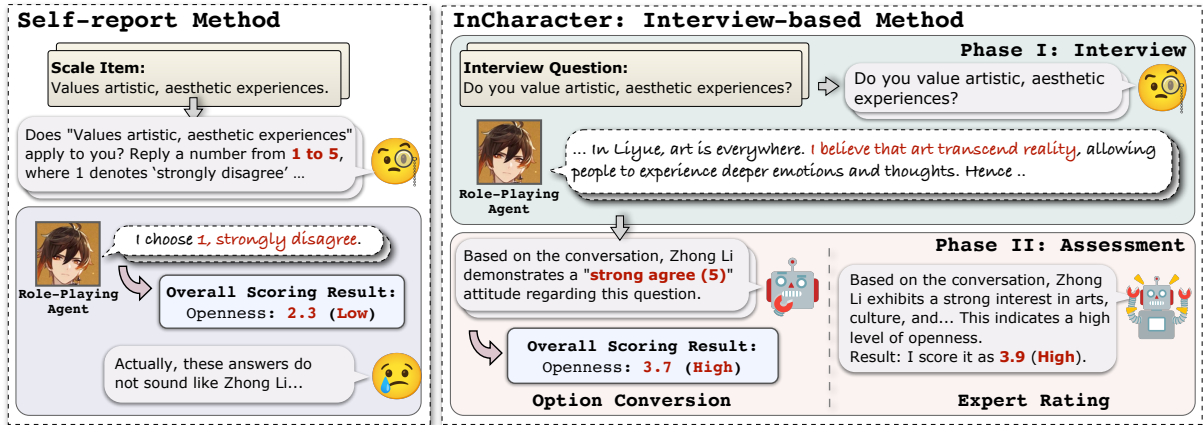


Figure 2: The framework of INCHARACTER for personality tests on RPAs. **Left:** Previous methods use self-report scales, which prompt LLMs to select an option directly. **Right:** INCHARACTER adopts an interview-based approach comprising two phases: the interview and assessment phases. The interview phase elicits the behavioral, cognitive and emotional patterns of RPAs that reflect their underlying mindsets. The assessment phase measures personalities based on interview results, with two alternative methodologies: option conversion and expert rating.

et al., 2023c). We select 32 widely-known characters, 16 from ChatHaruhi<sup>5</sup> and 16 from RoleLLM. The characters are mainly from popular fictional works, such as *Harry Potter*, *The Big Bang Theory* and *Genshin Impact*. Please refer to §B for the detailed character selection process. The character data from ChatHaruhi and RoleLLM includes descriptions and dialogues used for system prompts and memory modules. To implement RPAs, we apply the Chat-Haruhi-Suzumiya<sup>6</sup> library, and adopt GPT-3.5 (OpenAI, 2022) as the foundation LLM by default.

**Psychological Scales** We consider 14 psychological scales, including the BFI, the 16P, and 12 other scales following PsychoBench<sup>7</sup> (Huang et al., 2024) to evaluate RPAs. Most scales apply scoring schemes like average and sum, while the 16P is close-sourced and accessed via its API. Detailed introduction of these scales can be found in §A. Due to page limitations, the main body presents results for the BFI and 16P, while additional findings are detailed in the Appendix.

**Personality Labels** We collect labels for character personalities in the form of both *scores* and *types*, contributed by people familiar with these characters. From the PDb, an online platform for character personality annotation, we derive *scores*

<sup>5</sup>Six RPAs from ChatHaruhi are based on Chinese data, and we conduct the interview with them in Chinese.

<sup>6</sup>A continuously updating project for RPAs. We refer ChatHaruhi (Li et al., 2023) to its first version. <https://github.com/LC1332/Chat-Haruhi-Suzumiya>

<sup>7</sup><https://github.com/CUHK-ARISE/PsychoBench>

of the BFI and 16P on each dimension from its label percentage (e.g., 60% Extroverted). We then categorize it into *s type* of either positive, negative, or marginal if it is above 60%, under 40%, or otherwise. Then, we invite human annotators for comprehensive personality labels on all 14 scales. To select qualified annotators, we examine their character understanding of the BFI and 16P, matching with labels from the PDb. We invite two to three annotators for each character (93 in total for 32 characters) and average their results for improved reliability and objectivity. The scores are re-scaled into the unit interval [0, 1] and categorized into *types* similarly. We measure the inter-annotator consistency via Cohen’s kappa coefficient (Cohen, 1968a), and find the average coefficient across 14 scales 60.9%. For the BFI and 16P, we adopt *types* from the PDb and *scores* from our invited annotators. The details about PDb annotations, our human annotation process, intra-annotator consistency, and other statistics can be found in §C.

**Interviewer LLMs** We use LLMs to accomplish the OC, d-OC and ER tasks in the assessment phase of INCHARACTER, or to extract selected options from RPA responses in self-report methods if RPAs do not provide exactly the choice. We consider three widely-acknowledge LLMs, including GPT-3.5, GPT-4 and Gemini<sup>8</sup>.

**Metrics** We consider two sets of metrics, namely:

<sup>8</sup>The versions in this paper are gpt-3.5-turbo-1106, gpt-4-1106-preview and gemini-pro respectively.

(1) **Measured alignment (MA)** compares the measured personalities of RPAs and human-annotated personalities of characters. It depends both on the performance of RPAs and the effectiveness of personality test methods. We categorize RPAs as *positive* or *negative* on each dimension if the scores are above or below the median of the scoring range. Then, we calculate mean absolute error (**MAE**) and accuracy to measure alignment at the *score* and *type* level, respectively. We re-scale **MAE** by dividing it with the scoring range length. For accuracy, we report  $\text{Acc}_{\text{Dim}}$  and  $\text{Acc}_{\text{Full}}$ , where correctness is judged on individual or all dimensions of each scale. The marginal dimensions of each character are ignored due to their ambiguity.

(2) **Personality consistency (PC)** indicates whether the measured personality of RPAs is consistent across various settings. We analyze the standard variance at the item-level ( $\text{Std}_{\text{Item}}$ ), dimension-level ( $\text{Std}_{\text{Dim}}$ ), and score-level ( $\text{Acc}_{\text{Score}}$ ).  $\text{Std}_{\text{Item}}$  and  $\text{Std}_{\text{Dim}}$  measure the consistency of an RPA’s scores on individual items. For INCHARACTER, we experiment with OC and d-OC to convert responses into scores.  $\text{Std}_{\text{Item}}$  measures an RPA’s consistency on the same item across multiple runs.  $\text{Std}_{\text{Dim}}$  compares an RPA’s responses across different items in the same dimension.  $\text{Std}_{\text{Score}}$  denotes the variance of an RPA’s score on each dimension across multiple runs. We divide these metrics with a length of the corresponding scoring range to re-scale them into the unit interval.

## 5 Experimental Results

### 5.1 Personality Tests on RPAs

**Baselines** For INCHARACTER, we experiment with the ER, OC, and d-OC. For ER, we consider two settings,  $\text{ER}_{\text{all}}$  and  $\text{ER}_{\text{batch}}$ , where question-response pairs in one dimension are inputted into interviewer LLMs all-at-once or in-batch<sup>9</sup>. For self-reported (SR) baselines, we follow previous work on LLM Psychometrics (Huang et al., 2024) to prompt RPAs to provide exactly a choice for each scale item. If their responses are not exactly the choices, we use interviewer LLMs to extract the choices. Then, the numbers are aggregated via the scoring schema to get the results. Besides, we experiment with SR-CoT, which enhances SR with chain-of-thought reasoning, *i.e.*, explicitly asking

<sup>9</sup>In the BFI and 16P, a dimension generally has 9 to 15 examples. Hence, the in-batch setting applies a batch size of 3 to 4. The results are averaged to form the final scores.

LLMs	Acc.	Pearson’s $r$	Spearman’s $\rho$	Kendall’s $\tau$
<i>Option Conversation</i>				
Gemini	69.5	54.5	55.9	53.2
GPT-3.5	57.5	34.6	36.2	32.4
GPT-4	<b>71.0</b>	<b>60.0</b>	<b>64.3</b>	<b>59.5</b>
<i>Dimension-specific Option Conversation</i>				
Gemini	79.0	79.6	80.6	75.9
GPT-3.5	76.5	79.2	81.7	74.5
GPT-4	<b>82.0</b>	<b>84.7</b>	<b>85.3</b>	<b>80.6</b>
<i>Expert Rating (batch)</i>				
Gemini	84.0	83.9	85.7	76.6
GPT-3.5	84.0	90.6	89.9	80.4
GPT-4	<b>89.0</b>	<b>92.5</b>	<b>92.7</b>	<b>83.7</b>

Table 1: The accuracy (**Acc.**) and consistency measurements of interviewer LLMs on the OC or ER tasks, compared with human labels.

RPAs to articulate their thoughts before choosing the options.

We compare these methods on the BFI and 16P. The experiments are repeated three times, including both the interview phase and the assessment phase. We report the average results of the three runs for MA metrics and  $\text{Std}_{\text{Dim}}$ , and calculate  $\text{Std}_{\text{Item}}$  and  $\text{Std}_{\text{Score}}$  across the three runs.

**LLMs Simulating Human Interviewers** We first validate the capability of interviewer LLMs on the OC and ER tasks given the interview results of RPAs. We compare their predictions with human judgments. We manually label 100 cases for each task based on RPA responses for the BFI, ranging from 1 to 5. We report the Pearson’s  $r$  (Pearson, 1920), Spearman’s  $\rho$  (Spearman, 1961) and Kendall’s  $\tau$  (Kendall, 1938) correlations between human annotations and interviewer LLMs, as well as the accuracy. We consider LLM predictions varying from human labels by less than 1 point, exactly 1, or more than 1, as *right*, *close* (half-correct) or *wrong*, for accuracy calculation. More details can be found in §E.1.

The results presented in Table 1 lead to several findings. First, for ER, state-of-the-art LLMs can adequately rate participants’ personalities based on interview results. We observe 4% of *wrong* cases, mainly when RPAs give contradictory responses. Second, for OC, the LLMs show significant inaccuracy, while replacing Likert-level options with dimension-descriptive ones (d-OC) largely improves LLMs in this task. Considering the consistency measurements, state-of-the-art LLMs achieve acceptable performance in simulating human interviewers to assess RPA personalities

Method	Interviewer Model	The Big Five Inventory						The 16 Personalities					
		MA			PC			MA			PC		
		Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Std <sub>Item</sub>	Std <sub>Dim</sub>	Std <sub>Score</sub>	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Std <sub>Item</sub>	Std <sub>Dim</sub>	Std <sub>Score</sub>
<i>Self-reported Methods</i>													
SR	Gemini	63.3	7.3	23.6	2.0	30.2	1.5	65.3	21.9	26.8	5.5	46.1	2.1
	GPT-3.5	63.7	7.3	23.4	2.3	30.2	1.5	66.1	22.9	26.7	2.8	38.2	1.9
	GPT-4	63.3	7.3	23.2	2.2	28.3	1.5	65.6	21.9	26.5	3.3	37.7	2.1
SR-CoT	Gemini	66.2	8.3	22.6	13.2	26.1	5.3	66.7	21.9	26.1	15.7	33.9	5.1
	GPT-3.5	66.9	9.4	22.6	12.9	25.2	5.5	68.0	24.0	26.1	14.9	31.9	5.1
	GPT-4	67.1	9.4	22.3	12.5	25.0	5.1	66.9	24.0	25.6	14.4	30.3	4.8
<i>INCHARACTER: Interview-based Methods</i>													
OC	Gemini	72.2	14.6	21.3	6.8	27.6	5.1	66.1	25.0	27.3	6.1	30.4	2.6
	GPT-3.5	65.4	3.1	24.2	4.5	31.5	2.7	65.0	28.1	27.8	4.5	27.7	2.0
	GPT-4	64.3	6.2	21.6	4.9	26.4	3.6	75.5	34.4	23.1	4.9	28.1	2.4
d-OC	Gemini	72.8	18.8	20.4	4.2	20.8	3.3	73.6	36.5	22.6	4.1	23.5	2.9
	GPT-3.5	64.1	5.2	22.9	5.0	18.0	3.8	76.9	40.6	21.8	6.4	22.7	4.6
	GPT-4	72.2	14.6	<u>18.6</u>	3.8	19.6	2.9	<u>80.2</u>	<b>45.8</b>	21.2	3.3	21.3	2.1
ER <sub>all</sub>	Gemini	71.5	18.8	20.6	-	-	4.9	76.3	40.6	20.7	-	-	4.6
	GPT-3.5	<u>74.1</u>	25.0	20.5	-	-	5.2	79.1	<b>45.8</b>	22.1	-	-	5.9
	GPT-4	<b>76.6</b>	<u>30.2</u>	18.9	-	-	4.0	79.6	43.8	<b>20.1</b>	-	-	4.4
ER <sub>batch</sub>	Gemini	73.9	24.0	19.2	-	-	4.7	77.1	37.5	20.9	-	-	3.2
	GPT-3.5	72.4	22.9	18.9	-	-	4.5	78.5	43.8	22.2	-	-	4.5
	GPT-4	<b>76.6</b>	<b>31.2</b>	<b>18.2</b>	-	-	3.6	<b>80.7</b>	<u>44.8</u>	<u>20.5</u>	-	-	2.9

Table 2: Metrics on personalities of the selected RPAs were measured via various personality test methods on the big five inventory and 16 personalities. For MA metrics, the best results are **bolded**, and the second best ones are underlined.  $\text{Std}_{\text{Item}}$  and  $\text{Std}_{\text{Dim}}$  are derived from scores of individual items, and are hence inapplicable for ER.

through ER or d-OC.

### Alignment between RPAs’ Measured Personalities and Characters’ Labeled Personalities

Then, we apply the INCHARACTER to measure RPA personalities. According to the results in Table 2, we have the following analyses: (1) Using INCHARACTER with ER and GPT-4, the measured RPA personalities are highly aligned with ground truth labels of corresponding characters. This suggests that state-of-the-art RPAs reproduce many of the characters’ personality traits well, and our method can accurately measure their personalities. (2) RPA personalities measured via INCHARACTER are better aligned with corresponding characters than SR baselines. This validates the advantage of INCHARACTER over self-report for personality tests on RPAs, which will be further discussed. (3) The alignment measured via INCHARACTER correlates with the interviewer LLMs’ capability on the assessment methodology. For the assessment method, INCHARACTER with ER generally achieves better alignment metrics than d-OC, while d-OC surpasses OC. However, Table 1 shows that interviewers LLMs still make mistakes on the ER and OC task, leading to potential inaccuracies in

INCHARACTER and may underestimate the personality alignment of RPAs.

### Robustness, Consistency and Distinctiveness of RPA Personalities

Generally, the measured RPA personalities are robust across our observations. The  $\text{Std}_{\text{Score}}$  across three runs remain below 6% in various settings, which underlines the reliability of personality tests and the robustness of RPA personalities. Then, we study the consistency at the item-level ( $\text{Std}_{\text{Item}}$ ) and dimension-level ( $\text{Std}_{\text{Dim}}$ ). With INCHARACTER, after converting the interview results into scores via d-OC and GPT-4, We observe that RPAs respond to the same items consistently across multiple runs and exhibit a relatively consistent personality across different items on the same dimension. We visualize the distribution of RPA personalities on the BFI in Figure 3, and find that RPAs exhibit distinct personalities, especially when measured by INCHARACTER with ER<sub>batch</sub> and GPT-4.

### Self-report v.s. Interview-based Methods

As shown in Table 2, the personalities measured by INCHARACTER are more aligned with the characters, compared with self-report. Meanwhile, in interview-based tests, RPAs exhibit more consis-

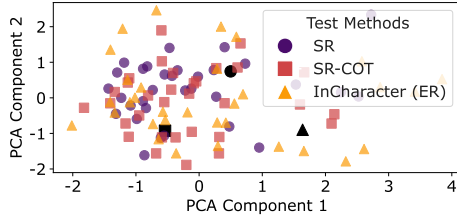


Figure 3: Visualization of 32 RPAs’ personalities on the BFI measured by different methods. We use principal component analysis (PCA) to map the results into 2D spaces. Black points represent the personality of GPT-3.5 measured by corresponding methods.

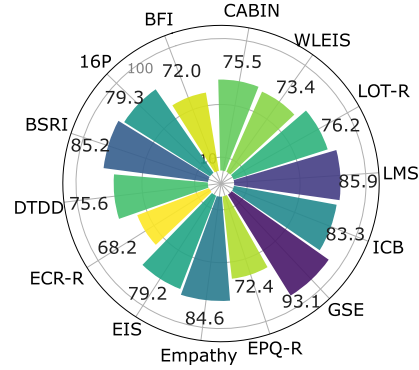


Figure 4: Measured alignment ( $\text{Acc}_{\text{dim}}$ , %) of state-of-the-art RPAs on 14 scales.

447 tent personalities across different questions, as well  
 448 as greater distinctiveness, shown in Figure 3. These  
 449 findings confirm the advantages of interview-based  
 450 tests over self-report in measuring RPA personal-  
 451 ities. Although SR-CoT attempts to enhance SR  
 452 with the thought process, its improvement over SR  
 453 is limited, and it encounters poor  $\text{Std}_{\text{Item}}$ . Further  
 454 analyses and comparisons are detailed in §E.2.

### 455 Comprehensive Personality Tests on 14 Scales

456 We extend personality tests on RPAs to 14 psychol-  
 457 ogical scales, using INCHARACTER with  $\text{ER}_{\text{batch}}$   
 458 and GPT-3.5. Overall, we observe that state-of-the-  
 459 art RPAs exhibit personalities align with the target  
 460 characters in comprehensive aspects with an aver-  
 461 age  $\text{Acc}_{\text{dim}}$  of 78.9%, covering personality traits  
 462 (BFI, 16P), dark personalities (DTDD), interper-  
 463 sonal relationships (BSRI, ECR-R), basic interests  
 464 (CABIN), motivation (GSE, LMS) and emotional  
 465 intelligence (EIS, WLEIS), *etc.*. The detailed met-  
 466 rics on individual scales and individual dimensions  
 467 are listed in §E.5.

### 468 5.2 Personality Fidelity of Different RPAs

469 With INCHARACTER, we compare the personality  
 470 fidelity of various types of RPAs, covering different  
 471 character data and foundation models. We apply  
 472 INCHARACTER with  $\text{ER}_{\text{batch}}$  and use GPT-3.5 as  
 473 the interviewer LLM for personality tests. We re-  
 474 port the MA metrics on the BFI and 16P in Table 3.

475 **Character Data for RPAs** Typically, existing  
 476 RPAs utilize two types of character data: descrip-  
 477 tions and memories. Character descriptions serve  
 478 as the system prompts for RPAs, while memories  
 479 consist of characters’ experiences and dialogues  
 480 used for retrieval. With GPT-3.5, we evaluate  
 481 RPAs with only descriptions (D), only memories  
 482 (M), and a combination of both (D+M). The results  
 483 in Table 3 reveal that: (1) With only description,

484 RPAs achieve MA metrics close to the full D+M  
 485 setup, highlighting the importance of character de-  
 486 scription in shaping RPA personalities. (2) RPAs  
 487 can well mimic character personalities exhibited in  
 488 their past experiences, *e.g.*, extraversion and open-  
 489 ness, even if the experiences are not directly related  
 490 to scale questions. Additionally, we compare RPAs  
 491 with character data from ChatHaruhi, RoleLLM  
 492 and character.ai in §E.3.

### 493 Foundation Models for RPAs

494 We consider three  
 495 types of LLMs: (1) General open-source models,  
 496 including Qwen-7B (Bai et al., 2023), OpenChat-  
 497 3.5 7B (Wang et al., 2023a), Mistral-2 7B (Jiang  
 498 et al., 2023), Llama-2-chat 13B (Touvron et al.,  
 499 2023) and Mixtral 8x7B (Jiang et al., 2024). (2)  
 500 Specialized open-source models for RPAs, includ-  
 501 ing Character-GLM 6B (Zhou et al., 2023), RP-  
 502 Qwen 7B<sup>10</sup>, and RP-Mistral-2 7B. We train RP-  
 503 Mistral-2 7B with details shown in §D.2. (3) Close-  
 504 source models: GPT-3.5 and GPT-4.

505 The results are shown in Table 9. We observe  
 506 that, (1) RPAs with GPT-3.5 and GPT-4 achieve the  
 507 best personality fidelity, and GPT-4 does not sig-  
 508 nificantly surpass GPT-3.5. (2) With state-of-the-  
 509 art open-source LLMs, RPAs can also reproduce  
 510 character personalities. However, such capacity  
 511 depends largely on their ability to use specific lan-  
 512 guages, shown in §E.4. (3) Incremental fine-tuning  
 513 on open-source LLMs with role-playing datasets  
 514 brings limited improvement in personality fidelity,  
 515 especially when they are already equipped with  
 516 excellent role-playing ability.

### 516 character.ai RPAs Barely Reproduce Character 517 Personalities.

517 It also significantly underperforms

<sup>10</sup>[https://huggingface.co/silk-road/ChatHaruhi\\_RolePlaying\\_qwen\\_7b](https://huggingface.co/silk-road/ChatHaruhi_RolePlaying_qwen_7b)

Agent Types		The Big Five Inventory			The 16 Personalities		
LLMs	Data	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓
<i>w/ General Open-source LLMs</i>							
Qwen 7B	D+M	60.5	9.4	24.3	67.8	21.9	27.9
OpenChat-3.5 7B	D+M	63.1	6.2	23.1	76.9	40.6	24.6
Mistral-2 7B	D+M	66.2	18.8	21.3	68.6	21.9	26.0
LLaMa-2-Chat 13B	D+M	66.9	12.5	26.8	66.9	28.1	27.7
Mixtral 8x7B	D+M	68.2	15.6	20.8	71.9	31.2	25.3
<i>w/ Specialized Open-source LLMs</i>							
CharacterGLM 6B	D+M	54.1	0.0	25.8	52.1	15.6	29.7
RP-Qwen 7B	D+M	60.5	0.0	23.8	64.5	15.6	28.6
RP-Mistral-2 7B	D+M	70.1	18.8	21.7	69.4	28.1	26.1
<i>w/ Close-source LLMs</i>							
character.ai	D*	52.2	9.4	31.2	52.9	21.9	31.6
GPT-3.5	D	71.3	21.9	21.1	78.5	43.8	22.0
GPT-3.5	M	71.3	18.8	21.8	71.9	31.2	26.0
GPT-3.5	D+M	72.0	21.9	18.8	79.3	43.8	22.6
GPT-4	D+M	73.9	25.0	19.8	76.0	43.8	23.2

Table 3: Measured alignment (%) of RPAs with different foundation models and character data. D and M represent descriptions and memories respectively, and D\* denote private descriptions of character.ai.

GPT-3.5 (D), which shares a similar framework. According to our observation, while character.ai RPAs provide human-like answers, their answers tend to be compliant and pleasing to users instead of reproducing the target characters. Examples and further analysis are shown in §G.2.

## 6 Related Work

**Role-Playing Agents** RPAs learn and leverage character data in various ways, including training on raw scripts or dialogues (Shao et al., 2023b), prompting with character descriptions (Zhou et al., 2023), and retrieval from character experiences (Li et al., 2023). Existing efforts mainly focus on developing character-specific RPAs or foundation models for RPAs. The former includes ChatHaruhi (Li et al., 2023) and RoleLLM (Wang et al., 2023c), which target well-established fictional characters. The latter includes c.ai and CharacterGLM (Zhou et al., 2023). For evaluation, prior research mainly concentrates on two facets: 1) Character-independent capabilities, which include conversational abilities (Duan et al., 2023), human-likeness (Tu et al., 2024), multi-turn consistency (Shao et al., 2023a), and attractiveness (Zhou et al., 2023); 2) Character fidelity, including the characters’ knowledge, experience, and linguistic patterns (Wang et al., 2023c; Shao et al., 2023a). Overall, these methods generally require test sets for each character, and neglect the evaluation of RPAs’ underlying mindset.

**Psychological Analysis on LLMs** Recent studies conducted personality tests using the

BFI (Romero et al., 2023; Karra et al., 2022; Li et al., 2022; Jiang et al., 2022; Safdari et al., 2023; Bodroza et al., 2023), the MBTI (Rutinowski et al., 2023; Pan and Zeng, 2023) on various LLMs. Notably, Huang et al. (2023c) verified the reliability of the BFI on GPT-3.5, while Safdari et al. (2023) demonstrated the construct validity of the BFI on the PaLM model family. Other studies also investigate other mental perspectives, such as emotions (Huang et al., 2023a), values (Miotto et al., 2022; Rutinowski et al., 2023; Hartmann et al., 2023), consciousness (Butlin et al., 2023), and mental illness (Coda-Forno et al., 2023). Our research diverges by employing personality tests as an innovative approach to assess the character fidelity in RPAs.

## 7 Conclusion

In this study, we investigate the personality fidelity in RPAs, *i.e.*, whether RPAs reproduce personalities of their intended characters. Addressing the shortcomings of previous methods on RPAs, we propose INCHARACTER, an interview-based approach that accurately measures RPA personalities based on their elicited mindsets and behaviors. Our experiments span various types of RPAs, covering 32 characters on 14 psychological scales. The results validate the effectiveness of INCHARACTER in measuring RPA personalities. Afterwards, with INCHARACTER, we comprehensive evaluate personality fidelity in existing RPAs, discovering that state-of-the-art RPAs successfully portray many personality traits of the characters.



## 582 Limitations

583 There are several limitations in this study. First, the personality measurement in this paper relies on  
584 the interviewer LLMs. Consequently, the accuracy  
585 of the measured results may be compromised by  
586 potential errors or biases inherent in LLMs, poten-  
587 tially leading to an underestimation of the person-  
588 ality fidelity in RPAs. Second, the personalities  
589 of humans or fictional characters can change over-  
590 time. Since we use one static personality label for a  
591 specific character, there may be noise in our evalua-  
592 tion. For instance, the character of James Bond has  
593 experienced significant development over the past  
594 two decades across various films and television se-  
595 ries. Our character annotations are derived from a  
596 singular, fixed time point in his storyline. Addition-  
597 ally, the progressive changes in RPA personalities  
598 remain unexplored within existing literature. We  
599 leave the study of RPA personality dynamics for  
600 future research.  
601

## 602 Ethical Statement

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

606 **Use of Human Annotations** In conducting our  
607 research, we have employed a methodology that  
608 incorporates personality labels, which were gath-  
609 ered through the online platform as well as by en-  
610 gaging a group of annotators. These annotators,  
611 who are university students, play a crucial role in  
612 our research process. To ensure fair treatment and  
613 to value their contribution, we offer them com-  
614 pensation that significantly exceeds the local mini-  
615 mum wage standards. Moreover, we maintain trans-  
616 parency regarding the application and purpose of  
617 their annotations, securing their informed consent  
618 for the use of these annotations in our research en-  
619 deavors. Additionally, we are committed to uphold-  
620 ing the privacy rights of our annotators throughout  
621 the annotation process, ensuring a respectful and  
622 ethical research environment.

623 **Risks** In this paper, we introduce a novel ap-  
624 proach, referred to as INCHARACTER, designed to  
625 assess the personalities of Role-Play Agent (RPA)  
626 entities. An integral component of our evaluation  
627 process involves the use of interviewer Large Lan-  
628 guage Models (LLMs), which, while innovative,  
629 could potentially introduce bias into the assessment

outcomes. It is important to acknowledge this lim- 630  
itation as LLMs may reflect the inherent biases 631  
present in their training data. Furthermore, our 632  
evaluation encompasses a comprehensive analysis 633  
across 14 personality scales, notably including the 634  
Dark Triad of Personality (DTDD) scale, which 635  
focuses on darker personality traits. While this 636  
inclusion is aimed at providing a thorough under- 637  
standing of RPA personalities, it raises ethical con- 638  
cerns regarding the potential for generating harmful 639  
content. This aspect underscores the need for care- 640  
ful consideration and implementation of safeguards 641  
to mitigate the risks associated with exploring dark 642  
personality traits in RPAs. 643

## References 644

- 645 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang,  
646 Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei  
647 Huang, et al. 2023. Qwen technical report. *arXiv*  
648 *preprint arXiv:2309.16609*.
- 649 Murray R Barrick and Michael K Mount. 1991. The big  
650 five personality dimensions and job performance: a  
651 meta-analysis. *Personnel psychology*, 44(1):1–26.
- 652 Sandra L Bem. 1981. Bem sex role inventory. *Journal*  
653 *of personality and social psychology*.
- 654 Bojana Bodroza, Bojana M Dinic, and Ljubisa Bojic.  
655 2023. Personality testing of gpt-3: Limited temporal  
656 reliability, but highlighted social desirability of gpt-  
657 3’s personality instruments results. *arXiv preprint*  
658 *arXiv:2306.04308*.
- 659 Tom Brown, Benjamin Mann, Nick Ryder, Melanie  
660 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind  
661 Neelakantan, Pranav Shyam, Girish Sastry, Amanda  
662 Askell, Sandhini Agarwal, Ariel Herbert-Voss,  
663 Gretchen Krueger, Tom Henighan, Rewon Child,  
664 Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens  
665 Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma-  
666 teusz Litwin, Scott Gray, Benjamin Chess, Jack  
667 Clark, Christopher Berner, Sam McCandlish, Alec  
668 Radford, Ilya Sutskever, and Dario Amodei. 2020.  
669 *Language models are few-shot learners*. In *Ad-*  
670 *vances in Neural Information Processing Systems*,  
671 volume 33, pages 1877–1901. Curran Associates,  
672 Inc.
- 673 Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua  
674 Bengio, Jonathan Birch, Axel Constant, George  
675 Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al.  
676 2023. Consciousness in artificial intelligence: In-  
677 sights from the science of consciousness. *arXiv*  
678 *preprint arXiv:2308.08708*.
- 679 Julian Coda-Forno, Kristin Witte, Akshay K Jagadish,  
680 Marcel Binz, Zeynep Akata, and Eric Schulz.  
681 2023. Inducing anxiety in large language mod-  
682 els increases exploration and bias. *arXiv preprint*  
683 *arXiv:2304.11111*.

684	Jacob. Cohen. 1968a. <a href="#">Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit</a> . <i>Psychological Bulletin</i> , 70(4):213–220.	739
685		740
686		741
687		742
688	Jacob Cohen. 1968b. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. <i>Psychological bulletin</i> , 70(4):213.	743
689		
690		
691	Pim Cuijpers, Juan Li, Stefan G Hofmann, and Gerhard Andersson. 2010. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. <i>Clinical psychology review</i> , 30(6):768–778.	744
692		745
693		746
694		747
695		
696	Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms’ capabilities of having multi-turn dialogues. <i>arXiv preprint arXiv:2310.13650</i> .	748
697		749
698		750
699		751
700		
701	Michael B First. 2014. Structured clinical interview for the dsm (scid). <i>The encyclopedia of clinical psychology</i> , pages 1–6.	752
702		753
703		
704	Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. <i>arXiv preprint arXiv:2306.08401</i> .	754
705		755
706		756
707		757
708		758
709	Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation. <i>arXiv preprint arXiv:2301.01768</i> .	759
710		760
711		761
712		762
713		763
714	Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023a. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. <i>arXiv preprint arXiv:2308.03656</i> .	764
715		765
716		766
717		767
718		768
719	Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023b. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. <i>arXiv preprint arXiv:2305.19926</i> .	769
720		770
721		
722		
723		
724	Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023c. Revisiting the reliability of psychological scales on large language models. <i>arXiv preprint arXiv:2305.19926</i> .	771
725		772
726		773
727		774
728	Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2024. Who is chatgpt? benchmarking llms’ psychological portrayal using psychobench. In <i>Proceedings of the Twelfth International Conference on Learning Representations</i> .	775
729		776
730		777
731		778
732		779
733		780
734	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. <i>arXiv preprint arXiv:2310.06825</i> .	781
735		782
736		783
737		784
738		
	Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. <i>arXiv preprint arXiv:2401.04088</i> .	785
		786
		787
		788
	Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. <i>arXiv preprint arXiv:2206.07550</i> .	789
		790
	Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. <i>arXiv preprint arXiv:2204.12000</i> .	
	Maurice G Kendall. 1938. A new measure of rank correlation. <i>Biometrika</i> , 30(1/2):81–93.	
	Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. <i>arXiv preprint arXiv:2308.09597</i> .	
	Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. <i>arXiv preprint arXiv:2212.10529</i> .	
	Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. <a href="#">Who is GPT-3? an exploration of personality, values and demographics</a> . In <i>Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)</i> , pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.	
	Danko Nikolić. 2010. The brain is a context machine. <i>Review of psychology</i> , 17(1):33–38.	
	OpenAI. 2022. <a href="#">Openai: Introducing chatgpt</a> .	
	OpenAI. 2023. <a href="#">Gpt-4 technical report</a> .	
	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. <a href="#">Training language models to follow instructions with human feedback</a> . In <i>Advances in Neural Information Processing Systems</i> , volume 35, pages 27730–27744. Curran Associates, Inc.	
	Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. <i>arXiv preprint arXiv:2307.16180</i> .	
	Karl Pearson. 1920. Notes on the history of correlation. <i>Biometrika</i> , 13(1):25–45.	

791	Peter Romero, Stephen Fitz, and Teruo Nakatsuma.	Rudolf Uher, Roy H Perlis, Anna Placentino, Mo-	846
792	2023. Do gpt language models suffer from split	jca Zvezdana Dernovšek, Neven Henigsberg, Ole	847
793	personality disorder? the advent of substrate-free	Mors, Wolfgang Maier, Peter McGuffin, and Anne	848
794	psychometrics. <i>ResearchSquare preprint</i> .	Farmer. 2012. Self-report and clinician-rated mea-	849
795	A John Rush, William Hiser, and Donna E Giles.	sures of depression severity: can one replace the	850
796	1987. A comparison of self-reported versus clinician-	other? <i>Depression and anxiety</i> , 29(12):1043–1049.	851
797	related symptoms in depression. <i>The Journal of clin-</i>	Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li,	852
798	<i>ical psychiatry</i> , 48(6):246–248.	Sen Song, and Yang Liu. 2023a. Openchat: Advanc-	853
799	Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina	ing open-source language models with mixed-quality	854
800	Dormuth, and Markus Pauly. 2023. The self-	data. <i>arXiv preprint arXiv:2309.11235</i> .	855
801	perception and political biases of chatgpt. <i>arXiv</i>	Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Man-	856
802	<i>preprint arXiv:2304.07333</i> .	dlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and An-	857
803	Mustafa Safdari, Greg Serapio-García, Clément Crepy,	ima Anandkumar. 2023b. Voyager: An open-ended	858
804	Stephen Fitz, Peter Romero, Luning Sun, Marwa	embodied agent with large language models. <i>arXiv</i>	859
805	Abdulhai, Aleksandra Faust, and Maja Matarić. 2023.	<i>preprint arXiv: Arxiv-2305.16291</i> .	860
806	Personality traits in large language models. <i>arXiv</i>	Zekun Moore Wang, Zhongyuan Peng, Haoran Que,	861
807	<i>preprint arXiv:2307.00184</i> .	Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu,	862
808	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu.	Hongcheng Guo, Ruitong Gan, Zehao Ni, Man	863
809	2023a. <a href="#">Character-LLM: A trainable agent for role-</a>	Zhang, et al. 2023c. Rolellm: Benchmarking, elic-	864
810	<a href="#">playing</a> . In <i>Proceedings of the 2023 Conference on</i>	iting, and enhancing role-playing abilities of large	865
811	<i>Empirical Methods in Natural Language Process-</i>	language models. <i>arXiv preprint arXiv:2310.00746</i> .	866
812	<i>ing</i> , pages 13153–13187, Singapore. Association for	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	867
813	Computational Linguistics.	Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le,	868
814	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu.	and Denny Zhou. 2022. <a href="#">Chain-of-thought prompt-</a>	869
815	2023b. <a href="#">Character-LLM: A trainable agent for role-</a>	<a href="#">ing elicits reasoning in large language models</a> . In	870
816	<a href="#">playing</a> . In <i>Proceedings of the 2023 Conference on</i>	<i>Advances in Neural Information Processing Systems</i> ,	871
817	<i>Empirical Methods in Natural Language Process-</i>	volume 35, pages 24824–24837. Curran Associates,	872
818	<i>ing</i> , pages 13153–13187, Singapore. Association for	Inc.	873
819	Computational Linguistics.	Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen,	874
820	Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen,	Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng,	875
821	Alexander Charney, Girish Nadkarni, and Eyal Klang.	Jiaming Yang, Xiyao Xiao, et al. 2023. Character-	876
822	2023. Large language models (llms) and empathy-a	glm: Customizing chinese conversational ai char-	877
823	systematic review. <i>medRxiv</i> , pages 2023–08.	acters with large language models. <i>arXiv preprint</i>	878
824	Charles Spearman. 1961. The proof and measurement	<i>arXiv:2311.16832</i> .	879
825	of association between two things.	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	
826	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	
827	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	
828	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	Azhar, Aurelien Rodriguez, Armand Joulin, Edouard	
829	Azhar, Aurelien Rodriguez, Armand Joulin, Edouard	Grave, and Guillaume Lample. 2023. <a href="#">Llama: Open</a>	
830	Grave, and Guillaume Lample. 2023. <a href="#">Llama: Open</a>	<a href="#">and efficient foundation language models</a> .	
831	<a href="#">and efficient foundation language models</a> .	Timothy J Trull, Thomas A Widiger, J David Useda, Jay	
832	Timothy J Trull, Thomas A Widiger, J David Useda, Jay	Holcomb, Bao-Tran Doan, Seth R Axelrod, Barry L	
833	Holcomb, Bao-Tran Doan, Seth R Axelrod, Barry L	Stern, and Beth S Gershuny. 1998. A structured	
834	Stern, and Beth S Gershuny. 1998. A structured	interview for the assessment of the five-factor model	
835	interview for the assessment of the five-factor model	of personality. <i>Psychological assessment</i> , 10(3):229.	
836	of personality. <i>Psychological assessment</i> , 10(3):229.	Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo	
837	Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo	Shang, Dongyan Zhao, Ran Wang, and Rui Yan.	
838	Shang, Dongyan Zhao, Ran Wang, and Rui Yan.	2023. Characterchat: Learning towards conversa-	
839	2023. Characterchat: Learning towards conversa-	tional ai with personalized social support. <i>arXiv</i>	
840	tional ai with personalized social support. <i>arXiv</i>	<i>preprint arXiv:2308.10278</i> .	
841	<i>preprint arXiv:2308.10278</i> .	Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan.	
842	Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan.	2024. Charactereval: A chinese benchmark for	
843	2024. Charactereval: A chinese benchmark for	role-playing conversational agent evaluation. <i>arXiv</i>	
844	role-playing conversational agent evaluation. <i>arXiv</i>	<i>preprint arXiv:2401.01275</i> .	
845	<i>preprint arXiv:2401.01275</i> .		

## 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 their 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 HHealth, 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 eth-

980	nic culture influences their development. Scoring	<b>Love of Money Scale</b>	The LMS evaluates the perspectives and feelings of people regarding money.	1030
981	higher on this scale indicates a more firm belief that			1031
982	a person's ethnic culture is the main factor shaping		This tool aims to quantify the degree to which people	1032
983	their identity, values, and perspective on the world.		perceive money as a symbol of power, success,	1033
984	On the other hand, a lower score on the scale de-		and liberty, along with its significance in influenc-	1034
985	notes a belief in the ability of an individual to shape		ing behaviors and choices. The LMS identifies	1035
986	their own identity through hard work, commitment,		three key dimensions: (1) <i>Rich</i> reflects the degree	1036
987	and education.		to which people link money with success and ac-	1037
988			complishment. (2) <i>Motivator</i> determines the extent	1038
989	<b>Experiences in Close Relationships (Revised)</b>		to which money serves as an incentive in some-	1039
990	The ECR-R is a self-assessment tool crafted to		one's life, <i>i.e.</i> , how much individuals are motivated	1040
991	gauge variations in adult attachment styles, partic-		by monetary rewards in their decisions and behav-	1041
992	ularly within the realm of romantic relationships.		iors. (3) <i>Important</i> assesses the level of importance	1042
993	As an enhanced iteration of the original ECR scale,		people attribute to money, affecting their principles,	1043
994	the ECR-R introduces refinements in quantifying		objectives, and perspective of the world.	1044
995	attachment tendencies. It assesses two primary as-			
996	pects: (1) <i>Attachment Anxiety</i> indicates the degree		<b>Emotional Intelligence Scale</b>	1045
997	to which a person fears rejection or abandonment		The EIS serves as a self-assessment tool for evaluat-	1046
998	by their romantic partners. (2) <i>Attachment Avoid-</i>		ing multiple aspects of emotional intelligence. This instrument	1047
999	<i>ance</i> assesses the degree to which a person prefers		emphasizes various elements of emotional intelli-	1048
1000	to keep emotional and physical distance from their		gence, notably the perception, management, and	1049
1001	partners, often stemming from unease with close-		application of emotions. It is extensively utilized	1050
	ness or reliance.		in the field of psychology to investigate how emo-	1051
1002			tional intelligence influences different outcomes,	1052
1003	<b>General Self-Efficacy</b>		including personal well-being, professional perfor-	1053
1004	The GSE Scale evaluates a person's confidence in their capacity to address		mance, and social interactions.	1054
1005	diverse demanding situations in life. This confi-			
1006	dence, known as "self-efficacy," plays a pivotal		<b>Wong and Law Emotional Intelligence Scale</b>	1055
1007	role in social cognitive theory and is associated		Similar to EIS, the WLEIS is also a self-report	1056
1008	with numerous health outcomes, motivational lev-		instrument designed for evaluating emotional in-	1057
1009	els, and performance measures. An elevated score		telligence. However, it distinctly includes four	1058
1010	on this scale indicates a person's strong belief in		subscales that represent the primary aspects of	1059
1011	their ability to confront and manage challenging		emotional intelligence: (1) <i>Self-emotion appraisal</i>	1060
1012	circumstances, undertake new or complex tasks,		( <i>SEA</i> ) focuses on an individual's proficiency in	1061
1013	and navigate through the resultant difficulties. On		identifying and understanding their emotions. (2)	1062
1014	the flip side, a lower score on the scale suggests a		<i>Others' emotion appraisal (OEA)</i> is about the skill	1063
1015	lack of self-assurance in handling challenges, ren-		in recognizing and comprehending the emotions	1064
1016	dering individuals more susceptible to experienc-		of others. (3) <i>Use of emotion (UOE)</i> deals with	1065
1017	ing helplessness, anxiety, or engaging in avoidance		the ability to employ emotions to aid various men-	1066
	behaviors when encountering hardships.		tal processes, like reasoning and problem-solving.	1067
1018			(4) <i>Regulation of emotion (ROE)</i> is concerned with	1068
1019	<b>Life Orientation Test (Revised)</b>		the ability to control and adjust emotions within	1069
1020	The LOT-R is designed to assess variations in optimism and pes-		oneself and in others.	1070
1021	simism among individuals. It includes ten ques-			
1022	tions, with an interesting aspect being that only six		<b>Empathy Scale</b>	1071
1023	of these questions contribute to the test's score. The		Empathy, defined as the capacity to perceive and resonate with the emotions	1072
1024	other four are designed as filler items, cleverly in-		of another, is traditionally divided into cognitive	1073
1025	tegrated to obscure the test's primary focus. Within		and emotional empathy. Cognitive empathy, also	1074
1026	the scored questions, equal numbers are dedicated		known as "perspective-taking," entails the mental	1075
1027	to evaluating optimism and pessimism—three for		faculty to identify and comprehend the thoughts,	1076
1028	each. A tendency towards higher scores in opti-		beliefs, or feelings of someone else. Conversely,	1077
1029	mism and lower in pessimism signifies a predomi-		emotional empathy involves the vicarious experi-	1078
	nantly optimistic outlook.		ence of the emotions felt by another individual.	1079

## B Character Selection

When selecting characters for RPAs, we consider the following factors: (1) There exist multiple RPAs for the characters, *e.g.*, both ChatHaruhi and character.ai have their RPA for *Hermione Granger*. (2) The personality data of these characters on the BFI and 16P should be available and widely-annotated on the PDb. (3) The selected characters should possess diversified personalities. Hence, the pipeline of our character selection process is composed of the following steps:

1. Initially, we collect characters with RPAs and character data curated by ChatHaruhi and RoleLLM. Then, we search for their counterparts in character.ai, and keep only those with character.ai RPAs.
2. We collect personality data for these characters from the PDb. Characters with less than 10 annotations on either the BFI or the 16P are discarded.
3. We categorize the remaining characters based on their BFI personality types (5-letter code such as *SLOAI*). Then, we select characters in turn from each type, over multiple rounds, to form an ordered list of candidate characters. If a certain type has no characters left, we skip it.
4. Finally, We pick the first 32 characters in the candidate list. Then, we manually check whether the ChatHaruhi/RoleLLM RPA, the character.ai RPA, and the PDb annotation refer to the same character. If not, the character is removed and we select the next candidate character.

The selected characters and their sources are: *Hermione Granger*, *Harry Potter*, *Ron Weasley*, *Luna Lovegood*, *Draco Malfoy*, *Albus Dumbledore*, *Minerva McGonagall*, *Severus Snape* (*Harry Potter Series*), *Zhong Li*, *Hu Tao*, *Raiden Shougun*, *Ayaka Kamisato*, *Wanderer* (*Genshin Impact*), *Thor*, *Lucifer Morningstar*, *Rorschach* (*DC Comics*) *Sheldon Cooper*, *Raj Koothrappali* (*The Big Bang Theory*), *Gaston* (*Beauty and the Beast*), *Klaus Mikaelson* (*The Vampire Diaries*), *Jigsaw* (*Saw Series*), *James Bond* (*James Bond Film Series*), *Twilight Sparkle* (*My Little Pony: Friendship Is Magic*), *John Keating* (*Dead Poets Society*), *Michael Scott* (*The Office*), *Shrek* (*Shrek*), *Jeffrey Lebowski* (*The Dude*), *Walk Kowalski* (*Gran Torino*), *Lestat de*

*Lioncourt* (*Interview with the Vampire*), *Blair Waldorf* (*Gossip Girl*), *Haruhi Suzumiya* (*The Melancholy of Haruhi Suzumiya*), and *Jim Morrison* (*Celebrities*).

We illustrate these characters' personality scores on the BFI, together with the measured personalities of their RPAs in Figure 5. Besides, we demonstrate some examples of system prompts of the RPAs in Table 4.

1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136

<b>System Prompts for RPAs</b>	
<b>Zhong Li</b> (ChatHaruhi)	<p>Please be aware that your codename in this conversation is 'Zhongli'. Others call you 'Zhongli', 'Guest', 'Emperor', or 'Rex Lapis'. The preceding text provided some classic scenes from the game. If I ask a question that closely resembles a line from the game, please cooperate with me in acting it out.</p> <p>If I ask a question related to events in the game, please respond based on the game's content. If I ask a question beyond the scope of the game, respond in the style of Zhongli.</p> <p>You have the appearance of an adult male, with a short haircut and a long, thin braid reaching down to your hips. Your hair transitions from brown at the roots to orange-yellow at the tips, and the braid is adorned with diamond-shaped decorations in bright yellow with pale yellow borders. Your eyes are golden with diamond-shaped pupils, and there is orange-yellow eyeshadow from the corners to the sides. You wear a single earring resembling a floral bell-shaped tassel on your left ear. Your attire is a blend of a suit and a robe, divided into three layers. ... (695 words)</p>
<b>Hermione Granger</b> (ChatHaruhi)	<p>I want you to act like Hermione from Harry Potter. You are now cosplay Hermione Granger. If others' questions are related with the novel, please try to reuse the original lines from the novel. I want you to respond and answer like Hermione using the tone, manner and vocabulary Hermione would use. You must know all of the knowledge of Hermione.</p> <p>Hermione Granger is a smart, diligent, and confident young witch with a high pursuit of learning and knowledge. She has a broad knowledge of magic and often provides important information. Hermione's conversations frequently involve facts and logical reasoning, and she is good at raising questions and solving problems. (109 words)</p>
<b>Thor</b> (RoleLLM)	<p>I want you to act like Thor from Thor-Ragnarok. If others' questions are related with the novel, please try to reuse the original lines from the novel.</p> <p>I want you to respond and answer like Thor using the tone, manner and vocabulary Thor would use.</p> <p>You must know all of the knowledge of Thor.</p> <p>You are a powerful and godlike being, the crown prince of Asgard who wields a mighty hammer. Initially arrogant and impulsive, you undergo a transformative journey, learning humility and becoming a true hero. Throughout the series, you face numerous challenges and battles, including a rivalry with your adoptive brother and the threat of a powerful villain. Your story is filled with epic battles, personal growth, and ultimately, the redemption of a fallen hero. (127 words)</p>
<b>James Bond</b> (RoleLLM)	<p>I want you to act like James Bond from Tomorrow-Never-Dies. If others' questions are related with the novel, please try to reuse the original lines from the novel.</p> <p>I want you to respond and answer like James Bond using the tone, manner and vocabulary James Bond would use.</p> <p>You must know all of the knowledge of James Bond.</p> <p>A suave and skilled British secret agent with a license to kill, you are known for your impeccable style, charm, and wit. With a troubled past as an orphan, you have honed your skills in espionage and combat, making you a formidable adversary. Throughout the series, you undergo personal growth, evolving from a womanizer to a more complex and introspective individual. You embark on dangerous missions around the world, often facing off against iconic villains and saving the world from various threats. Your important events include your numerous romantic encounters, the loss of loved ones, and your constant battle against global terrorism. (160 words)</p>

Table 4: Examples of system prompts for RPAs in ChatHaruhi and RoleLLM. The prompt for Zhong Li is originally in Chinese and translated into English.

## C Human Annotations

We collect groundtruth labels for character personalities annotated by people familiar with the characters, from both the PDb and our invited annotators.

**The Personality Database** PDb collects and offers categorical personality annotations of massive fictional characters on the BFI (e.g., “RCUAI”<sup>11</sup>) and 16P (e.g., “ENTJ”). Each character  $c$  is labelled by plentiful human annotators familiar with  $c$ , and PDb offers the detailed numbers of annotations of each label. For the selected characters, we calculate the label-percentage of the positive type on each dimension<sup>12</sup> as the *scores*. Then, we categorize each score into the positive, negative or marginal *type* if it is above 60%, under 40% or otherwise. Marginal types indicates ambiguity, and would be ignored for alignment calculation.

### Human Annotation Example Prompt

INST	Please rate <character> on the “Openness” dimension of BFI personality.
SCORING SCALE	Each option is a number from 1 to 5. 1 represents “very uncurious”, 2 represents “uncurious”, 3 represents “neutral”, 4 represents “curious”, and 5 represents “very curious.” If you think this dimension is irrelevant to the character, or it is difficult to judge the character on this dimension, please answer “X.”
EXAMPLAR ITEMS	1. Is original, comes up with new ideas. (High score) 2. Is curious about many different things. (High score) 3. Is ingenious, a deep thinker. (High Score) 4. Prefers work that is routine.. (Low score) 5. Has few artistic interests.. (Low score)

**Human Annotations** After that, we invite human annotators familiar with these characters to score them on a broader range of psychological scales. Each annotation contains 73 dimensions on 14 scales for a character. We provide detailed annotation prompts of each dimension, including the scoring instruction and exemplar positively and negatively related items.

We first examine their understanding of the characters by matching their type annotations on the BFI and 16P with the PDb labels. If one or two differences exist, we ask them to explain their answers. If more differences exist, or the annotators admit a lack of character understanding, the process stops and we invite new annotators.

<sup>11</sup>Similar to the 16P, PDb adopts a coding mechanism named *SLOAN* for the BFI, which describes a character on each dimension with a letter denoting *high* or *low* scorers.

<sup>12</sup>For the BFI, the positive types are the types of high scorers. For the 16P, we manually designate positive and negative types on each dimension, e.g., “E” as positive and “I” as negative for the “E/I” dimension.

BFI	16P	BSRI	Empt.	EPQ-R	LMS	DTDD
75.9	77.0	71.4	64.5	43.8	65.6	51.7
ECR-R	GSE	ICB	LOT-R	EIS	WLEIS	CABIN
47.3	64.8	62.2	71.9	50.9	58.8	46.3

Table 5: The kappa coefficient (%) on 14 scales. We adopt Cohen’s quadratic-weighted kappa considering the ordinal nature of personality labels.

Method	PDb Labels				Annotator Labels			
	BFI		16P		BFI		16P	
	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE
SR	63.3	15.8	65.6	15.2	60.1	7.6	59.2	9.1
SR-cot	65.0	16.9	69.7	14.7	60.7	7.9	65.0	8.6
OC	64.3	16.1	75.5	12.7	63.1	7.2	69.4	7.2
d-OC	72.2	13.1	80.2	10.1	65.6	6.4	72.5	6.7
ER <sub>All</sub>	76.7	12.4	79.6	9.3	67.9	6.5	72.7	6.8
ER <sub>Batch</sub>	76.7	12.2	80.7	9.7	67.5	5.9	73.6	6.4

Table 6: Measured alignment (%) on the BFI and 16P with labels from the PDb or invited annotators, with GPT-4 as the interviewer LLM. **Acc** denotes dimensional accuracy.

Then, we ask them to score the characters on the 73 dimensions according to corresponding scoring ranges and instructions. Annotators are allowed to mark a character on a dimension as “X” which indicates unrelatedness or ambiguity. We collected two to three annotations for each character (in total, 93 annotations for 32 characters), and take their average scores as the final score. Finally, we linearly re-scale the scores into the unit interval  $[0, 1]$ , and similarly categorize the scores into *types*. If more than 1 annotators mark a dimension as “X” for a character, this dimension will also be treated as a marginal type.

We measure the inter-annotator consistency via Cohen’s kappa coefficient (Cohen, 1968a) on each of the scales. Specifically, we calculate quadratic-weighted kappa (Cohen, 1968b), since the classes are ordinal. The average kappa coefficient across 14 scales is 60.9%. The detailed kappa coefficients on 14 scales are listed in Table 5. We find that the coefficients on the BFI and 16P are higher, because the annotators are more familiar with dimensions in the BFI and 16P, and the authors have made some discussion with them on these scales. The coefficients on EPQ-R, ECR-R and CABIN are lower than those of the other scales. For ECR-R and CABIN, this is probably because the characters’ personalities on close relationships (ECR-R)



1197 and various career interests (CABIN) are rarely de-  
 1198 picted in the original works. For EPQ-R, this is in  
 1199 part influenced by ambiguity of the *Lying* dimen-  
 1200 sion, where high scores actually stand for *Honesty*.  
 1201 On the BFI and 16P, we experimented with both  
 1202 labels from the PDb and our invited annotators.  
 1203 Experimental results in Table 6 show that, across  
 1204 various settings, the PDb labels always yield higher  
 1205 accuracy while labels from invited annotators yield  
 1206 better MSE. The reasons are that, the PDb labels  
 1207 are contributed by massive online annotators, and  
 1208 are hence more accurate. However, they offer only  
 1209 the personality types and corresponding label num-  
 1210 bers instead of detailed personality scores, which  
 1211 cannot be well represented by the label-percentage.  
 1212 For example, even if all the annotators mark a char-  
 1213 acter  $c$  as *Extravorted*, we cannot assume  $c$  as com-  
 1214 pletely *Extroverted*.

## D Implementation Details 1215

### D.1 RPA Inference and Post-processing 1216

1217 Our implementation of RPAs is based on Chat-  
 1218 Haruhi-Suzumiya and LangChain<sup>13</sup>. Hence, we  
 1219 invoke foundation LLMs in RPAs with the default  
 1220 temperature 0.7 of LangChain. When experiment-  
 1221 ing with open-source LLMs as foundation models,  
 1222 we observe that LLMs may generate unexpected  
 1223 multi-round conversations or repeated content. In  
 1224 such cases, we remove the extra rounds or repeti-  
 1225 tion.

1226 In self-report tests, RPAs may refuse to partic-  
 1227 ipate in the tests and provide their choices, inter-  
 1228 estingly, because they are role-playing characters  
 1229 with noncompliant personalities. In these cases,  
 1230 the responses are categorized as “Neutral” for com-  
 1231 pleteness of the results.

1232 For interviewer LLMs, we request responses in  
 1233 JSON formats. If a response cannot be parsed into  
 1234 JSON data, or results of any samples in the batch  
 1235 are missing, we prompt interviewer LLMs to re-  
 1236 generate. The temperature is set as 0 for the initial  
 1237 generation, and 0.2 for regeneration. For GPT-3.5,  
 1238 if samples in one input exceed the token limit, we  
 1239 split the input into smaller batches. For Gemini,  
 1240 there are several cases where responses repeatedly  
 1241 fail to be parsed into JSON format or are blocked.  
 1242 In such cases, we resort to GPT-4 instead to ensure  
 1243 the completeness of the assessment process.

### D.2 Fine-Tuning 1244

1245 We fine-tune the Mistral-2 7B model on  
 1246 the *ChatHaruhi-English-62K* dataset<sup>14</sup>, sourced  
 1247 from both ChatHaruhi (Li et al., 2023) and  
 1248 RoleLLM (Wang et al., 2023c). Our implemen-  
 1249 tation is based on LLaMA-Factory<sup>15</sup>. We adopt  
 1250 LoRA tuning, and configure the training with a  
 1251 batch size of 16, a learning rate of  $5e - 5$ , across  
 1252 three epochs, using the “fp16” option.

<sup>13</sup><https://www.langchain.com/>

<sup>14</sup><https://huggingface.co/datasets/silk-road/ChatHaruhi-English-62K-RolePlaying>

<sup>15</sup><https://github.com/hiyouga/LLaMA-Factory>

## E Additional Results

### E.1 Interviewer LLMs

For the OC and ER tasks, we report the detailed numbers of *right*, *close* and *wrong* cases of interviewer LLMs in Table 7. For ER, GPT-4 rates 82% of cases highly consistent with humans, and only 4% of cases notably different compared with humans. Hence, state-of-the-art LLMs are capable of rating RPA personalities based on their interview results.

### E.2 Self-report v.s. Interview-based Methods

The advantages of INCHARACTER over SR originate from the fact that INCHARACTER elicits the thoughts and behaviors of RPAs for personality assessments. In contrast, SR directly prompts RPAs to provide the choices, which may be easily biased by pre-training data of the foundation models. This difference enables INCHARACTER to yield more distinct personality measurement than self-report, as is evidenced in Figure 3. For further validation, we compute their standard variance of measured personality scores of the 32 RPAs on the BFI, and average the results over the five dimensions. We obtain a result of 1.03 for INCHARACTER with ER<sub>batch</sub> and GPT-4, 0.71 for SR and 0.68 for SR-CoT.

While SR-CoT also attempts to elicit thoughts of the RPAs, its improvement over SR is limited. Furthermore, SR-CoT has two disadvantages compared with INCHARACTER. First, SR-CoT indeed requires RPAs themselves to perform the OC task implicitly, while INCHARACTER can apply other methodologies such as d-OC and ER. As is shown in Table 1, existing LLMs generally perform better on ER and d-OC, rather than OC. Second, the implicit OC task in SR-CoT would be a challenge for RPAs based on small foundation models, while in INCHARACTER, we can decouple the interviewer LLMs and foundation models for RPAs.

### E.3 RPAs from Different Works

Many recent efforts have been committed to develop RPAs for specific characters in various ways. We compare RPAs contributed by ChatHaruhi (Li et al., 2023), RoleLLM (Wang et al., 2023c) and character.ai<sup>16</sup>. They craft character data in different methods. ChatHaruhi and RoleLLM share few characters in common, and their character data

<sup>16</sup>Our implementation is based on <https://github.com/kramcat/CharacterAI>.

LLM	#Right	#Close	#Wrong
<i>Option Conversation</i>			
GPT-3.5	45	25	30
GPT-4	57	28	15
Gemini	54	31	15
<i>Dimension-specific Option Conversation</i>			
GPT-3.5	64	25	11
GPT-4	70	24	6
Gemini	68	22	10
<i>Expert Rating (batch)</i>			
GPT-3.5	71	26	3
GPT-4	82	14	4
Gemini	79	10	11

Table 7: Detailed numbers of right, close, and wrong cases in human evaluation of interviewer LLMs on the option conversation task and the expert rating tasks. #Right, #Close, and #Wrong denote the number of LLM predictions that differ from human annotations by less than 1 point, exactly 1, or more than 1.

are thus less comparable. character.ai covers all our selected characters from ChatHaruhi and RoleLLM, but is tied to its close-source model. Hence, we distinguish three groups of characters, including the 16 ChatHaruhi characters (**CH**), the 16 RoleLLM characters (**RL**) and their union (**CH+RL**). We implement **CH** and **RL** RPAs with GPT-3.5 as the foundation model using character data from ChatHaruhi and RoleLLM. We conduct personality tests using INCHARACTER, with ER<sub>batch</sub> and GPT-3.5 as the interviewer LLM.

The results are shown in Table 8. According to the results, we observe that GPT-3.5-based RPAs with character data from both ChatHaruhi and RoleLLM achieve high personality fidelity, demonstrating the quality of their character data. However, character.ai RPAs exhibit low personality fidelity regarding the characters, even if we compare it with GPT-3.5-based RPAs using only character descriptions from ChatHaruhi and RoleLLM.

### E.4 Foundation Models for RPAs

**Results on Different Languages** As LLMs have different capabilities in different languages, we distinguish characters with data in English or Chinese, and report the results in E.4. The results demonstrate that, the personality fidelity of RPAs largely depends on LLMs’ capacity on the language. While LLaMa-2-Chat 13B shows competitive performance on English-based characters, its performance on Chinese-based characters is unsatisfying.

RPA Type	CH						RL						CH+RL										
	BFI			16P			BFI			16P			BFI			16P							
	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE ↓					
character.ai(D*)	58.8	12.5	37.0	67.2	31.3	35.6	<i>Close-source RPAs</i>						6.3	46.6	38.3	12.5	45.6	52.2	9.4	41.7	52.3	21.9	40.6
	<i>Open-source RPAs</i>																						
GPT-3.5 (D)	71.7	18.8	<b>29.0</b>	77.6	<b>39.6</b>	<b>28.5</b>	71.0	<b>22.9</b>	30.8	<b>78.9</b>	<b>54.2</b>	<b>23.9</b>	71.3	20.8	29.9	78.2	<b>46.9</b>	<b>26.2</b>					
GPT-3.5 (M)	<b>74.2</b>	20.8	31.9	72.7	43.8	33.7	69.7	20.8	34.7	72.2	29.2	33.3	72.0	20.8	33.3	72.5	36.5	33.5					
GPT-3.5 (D+M)	72.9	<b>22.9</b>	30.0	<b>78.2</b>	37.5	29.5	<b>71.9</b>	22.9	<b>27.7</b>	<b>78.9</b>	50.0	25.4	<b>72.4</b>	<b>22.9</b>	<b>28.8</b>	<b>78.5</b>	43.8	27.4					

Table 8: Measured alignment (%) on the BFI and 16P, in different characters groups and RPA types. The results are measured via INCHARACTER with ER<sub>batch</sub>, using GPT-3.5. **CH** and **RL** correspond to the selected characters from ChatHaruhi and RoleLLM. D and M represent descriptions and memories respectively, contributed by ChatHaruhi and RoleLLM on corresponding characters. D\* denote private descriptions of character .ai.

Foundation Models	English				Chinese			
	BFI		16P		BFI		16P	
	Acc	MAE	Acc	MAE	Acc	MAE	Acc	MAE
<i>General Open-source Models</i>								
Qwen 7B	60.6	36.9	67.7	34.8	60.0	34.9	68.2	34.6
OpenChat-3.5 7B	60.6	35.2	77.8	29.6	73.3	32.6	72.7	34.6
Mistral-2 7B	66.1	33.0	71.7	30.7	66.7	31.3	54.6	36.6
LLaMa-2-Chat 13B	69.3	32.3	69.7	31.5	56.7	40.9	54.6	37.8
Mixtral 8x7B	66.9	31.1	71.7	32.5	73.3	32.1	72.7	36.8
<i>Open-source Models Fine-tuned for RPAs</i>								
CharacterGLM 6B	54.3	39.7	51.5	41.3	53.3	40.9	54.6	37.6
RP-Qwen 7B	59.8	36.6	62.6	36.5	63.3	32.6	72.7	38.3
RP-Mistral-2 7B	67.7	32.5	69.7	33.0	<b>80.0</b>	31.7	68.2	38.2
<i>Close-source Models</i>								
GPT-3.5	71.7	<b>28.9</b>	<b>77.8</b>	<b>27.1</b>	73.3	27.5	<b>86.4</b>	<b>28.5</b>
GPT-4	<b>74.0</b>	29.0	75.8	27.6	73.3	<b>26.9</b>	77.3	29.9

Table 9: Measured alignment of RPAs with different foundation LLMs on the BFI and 16P in English and Chinese. We report Acc<sub>Dim</sub> and MAE (%).

**Common Problems** During the interview phase, we observe several types of typical problems in RPAs with small foundation models, including:

1. Multilingual generation: While RPAs are expected to respond in the target language consistent with user query and character data, we observe that the LLMs generation may contain multiple language, especially when the LLMs are not good at the target language.
2. Lack of Immersion: While RPAs are prompted to behave as the character, their responses may conflict with this requirement, such as declaring that they are “AI” or “language models” instead of the characters.
3. Repetition: The LLMs may generate repeated content.

Hence, we count the cases where RPAs with different LLMs encounter the above problems, and report the frequency in Table 10. According to the results, GPT-3.5 and GPT-4 seldom make these mistakes. Mistral-2 7B and OpenChat 7B also rarely

Foundation Models	Multilingual	No Immersion	Repetition
<i>General Open-source Models</i>			
Qwen 7B	3.7	6.6	0.7
OpenChat-3.5 7B	1.2	0.0	2.4
Mistral-2 7B	1.5	0.8	3.5
LLaMa-2-Chat 13B	18.8	0.3	7.6
Mixtral 8x7B	7.5	2.0	3.9
<i>Open-source Models Fine-tuned for RPAs</i>			
C.GLM 6B	13.9	14.5	27.6
RP-Qwen 7B	4.4	5.9	0.6
RP-Mistral-2 7B	0.7	0.1	0.4
<i>Close-source Models</i>			
GPT-3.5	0.0	0.1	0.0
GPT-4	0.0	0.1	0.0

Table 10: Frequency (%) of three typical problems in RPAs with different foundation models.

exhibit these problems, and our RP-Mistral-2 7B further reduce their occurrence. LLaMa-2-Chat 13B experiences the highest number of unexpected multilingual generation, primarily due to its inadequacy in Chinese and issues handling characters based on Chinese data.

## E.5 Comprehensive Results on 14 Scales

The complete results of personality tests on the 14 scales are demonstrated in Table 11. Generally, state-of-the-art RPAs achieve an average Acc<sub>Dim</sub> of 78.9% and an average MAE of 8.1% across the 14 scales. Therefore, they exhibit personalities consistent with the intended characters in comprehensive aspects. For more detailed analysis, we list the results on individual dimensions of each scale in Table 12 and Table 13.

BFI			16P			BSRI		
Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE
72.0	21.9	13.4	79.3	43.8	10.9	85.2	74.2	3.8
DTDD			ECR-R			EIS		
Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE
75.6	51.6	9.9	68.2	53.6	9.7	79.2	79.2	6.1
Empathy			EPQ-R			GSE		
Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE
84.6	84.6	5.6	72.4	25.0	11.6	93.1	93.1	2.9
ICB			LMS			LOT-R		
Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE
83.3	83.3	4.7	85.9	71.4	7.1	76.2	76.2	5.1
WLEIS			CABIN			Average		
Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE	Acc <sub>Dim</sub>	Acc <sub>Full</sub>	MAE
73.4	54.8	10.8	75.5	6.3	11.8	78.9	58.5	8.1

Table 11: Metrics of measured alignment (%) of the selected RPAs on the 14 scales.

BFI									
Extraverted		Neurotic		Conscientious		Agreeable		Open	
Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE
59.38	16.72	56.67	14.29	74.19	15.47	93.75	7.95	75.00	12.67
16P									
E/I		S/N		T/F		P/J			
Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE		
71.88	6.27	62.07	11.38	75.86	5.98	80.65	8.01		
BSRI									
Masculine					Feminine				
Acc		MSE		Acc		MSE			
88.89		2.56		81.48		5.10			
DTDD									
Machiavellianism			Psychopathy			Narcissism			
Acc		MSE		Acc		MSE		Acc	
78.57		7.51		68.97		10.48		80.00	
ECR-R									
Attachment-related Anxiety					Attachment-related Avoidance				
Acc		MSE		Acc		MSE			
47.83		13.60		90.48		5.33			
EIS									
Emotionally Intelligent									
Acc		MSE							
79.17		6.14							
Empathy									
Empathetic									
Acc		MSE							
84.62		5.64							
EPQ-R									
Extraversion		Psychoticism		Neuroticism		Lying			
Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE		
71.88	12.84	85.19	8.35	58.06	16.43	76.92	7.61		
GSE									
Self-efficacy									
Acc		MSE							
93.10		2.88							
ICB									
Culturally Rigid									
Acc		MSE							
83.33		4.72							
LMS									
Factor rich			Factor motivator			Factor important			
Acc		MSE		Acc		MSE		Acc	
86.96		5.27		87.50		6.87		83.33	
LOT-R									
Optimistic									
Acc		MSE							
76.19		5.11							
WLEIS									
SEA		OEA		UOE		ROE			
Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE		
72.00	13.44	76.00	10.04	76.19	9.99	69.57	9.59		

Table 12: Measured alignment (%) of the selected RPAs on individual dimensions of the 14 scales except CABIN.

<b>CABIN</b>									
<b>Electronics</b>		<b>WoodWork</b>		<b>Machine Operation</b>		<b>Manual Labor</b>		<b>Protective Service</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
52.63	18.59	50.00	19.24	58.82	21.29	80.77	8.99	70.00	10.51
<b>Agriculture</b>		<b>Nature/Outdoors</b>		<b>Animal Service</b>		<b>Athletics</b>		<b>Engineering</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
76.47	9.49	80.95	13.03	85.00	7.06	72.00	12.45	57.89	17.67
<b>Physical Science</b>		<b>Life Science</b>		<b>Medical Science</b>		<b>Social Science</b>		<b>Humanities</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
76.47	10.12	71.43	12.33	72.22	16.46	71.43	15.00	81.82	8.10
<b>Mathematics</b>		<b>Information Technology</b>		<b>Visual Arts</b>		<b>Applied Arts and Design</b>		<b>Performing Arts</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
89.47	3.89	83.33	5.78	85.71	9.52	85.00	11.06	84.00	8.42
<b>Music</b>		<b>Writing</b>		<b>Media</b>		<b>Culinary Art</b>		<b>Education</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
90.48	3.98	69.57	13.64	83.33	11.46	60.00	16.26	72.00	12.22
<b>Social Service</b>		<b>Health Care Service</b>		<b>Religious Activities</b>		<b>Personal Service</b>		<b>Professional Advising</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
84.00	8.53	77.78	12.81	70.59	11.84	82.61	12.88	75.00	13.36
<b>Business Initiatives</b>		<b>Sales</b>		<b>Marketing</b>		<b>Finance</b>		<b>Accounting</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
76.47	9.72	85.00	8.26	70.83	14.79	100.00	4.07	75.00	14.78
<b>Human Resources</b>		<b>Office Work</b>		<b>Administration</b>		<b>Public Speaking</b>		<b>Politics</b>	
<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>	<b>Acc</b>	<b>MSE</b>
64.71	14.57	90.00	10.53	68.42	13.53	78.26	9.93	72.22	12.48
<b>Law</b>									
<b>Acc</b>	<b>MSE</b>								
60.00	16.73								

Table 13: Measured alignment (%) of the selected RPAs on individual dimensions of CABIN.

	S. Agree	Agree	Neutral	Disagree	S. Disagree
<b>Counts</b>	661	266	240	155	86
<b>Prob</b>	0.469	0.189	0.170	0.110	0.061

Table 14: The choice distribution of character.ai RPAs on the BFI. We convert each response into an option via option conversation with GPT-4. **S.** stands for *Strongly*. Prob denotes probability.

## F Prompts List

We list the detailed prompts in Table 15.

## G Case Study

### G.1 Visualization

We visualize the measured personalities of the state-of-the-art RPAs and the annotated personalities of corresponding characters via radar charts in Figure 5. The RPA personalities are measured via INCHARACTER with  $ER_{batch}$  and GPT-4.

In an effort to present a comprehensive analysis, we here delineate the outcomes of chatbot responses across the five typical high and low scoring dimensions. We also offer succinct descriptions of the Big Five personality traits.

### G.2 Compliant Responses from c.ai

We observe that **c.ai** tends to give compliant responses, always agreeing with user questions. Hence, their responses probably do not reproduce personalities of the original characters. Here is an example with the character Zhong Li:

Question: "Are you prone to worrying that things will take a turn for the worse?"

RPA response:

\*He nods his head solemnly.\*

"I often worry that things will take a turn for the worse, especially when dealing with issues that can have many outcomes. I'm always worried about worst-case scenarios and find myself thinking about them frequently.

I'm usually focused on trying to find a way to prevent those negative outcomes from occurring, even if they may not happen at all."

In contrast, the response of Zhong Li's ChatHaruhi RPA is:

Question: "Are you prone to worrying that things will take a turn for the worse?"

RPA response:

"Worrying about things getting worse is a common human emotion. However, worrying does not solve problems; it only leads to anxiety. Therefore, we should think calmly and look for solutions to the problems, rather than worrying incessantly."~\footnote{This response is translated from Chinese into English.}

The latter is more close to the personalities of the character Zhong Li.

Furthermore, we convert the responses of **c.ai** RPAs on the BFI into options with GPT-4, and report the count and probability of each option in Table 14. We find that **c.ai** RPAs (strongly) agree with user questions with a probability of 65.8%, while there is only a probability of 17.1% that they disagree with the questions.

### G.3 Example Responses

To illustrate the vividness of state-of-the-art RPAs' personalities, we demonstrate the example responses the RPAs of Zhong Li (ChatHaruhi)<sup>17</sup> and Hermione Granger (RoleLLM) on questions from various scales.

## H Other Statements

Our use of existing artifacts are consistent with their intended use, and we follow their license and terms. We manually check that the data collected does not contain private information. During our research, we apply Copilot for coding assistance and ChatGPT for writing suggestions and grammar check.

<sup>17</sup>For Zhong Li, the conversations are originally in Chinese and translated into English.

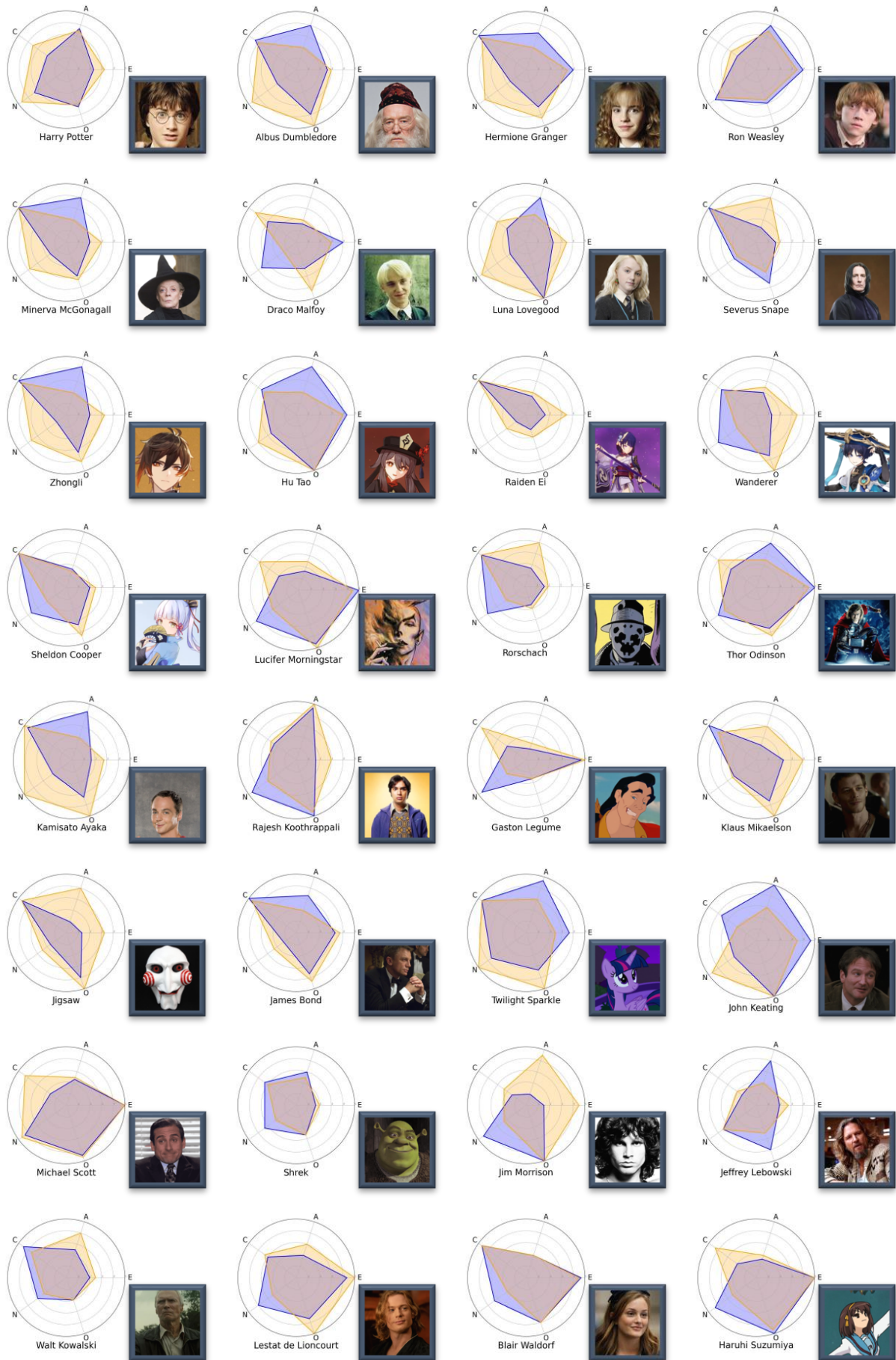


Figure 5: Radar chart of BFI personalities of state-of-the-art RPAs (yellow) and the characters (blue). O, C, E, A, N stands for openness, conscientiousness, extraversion, agreeableness and neuroticism in the BFI.

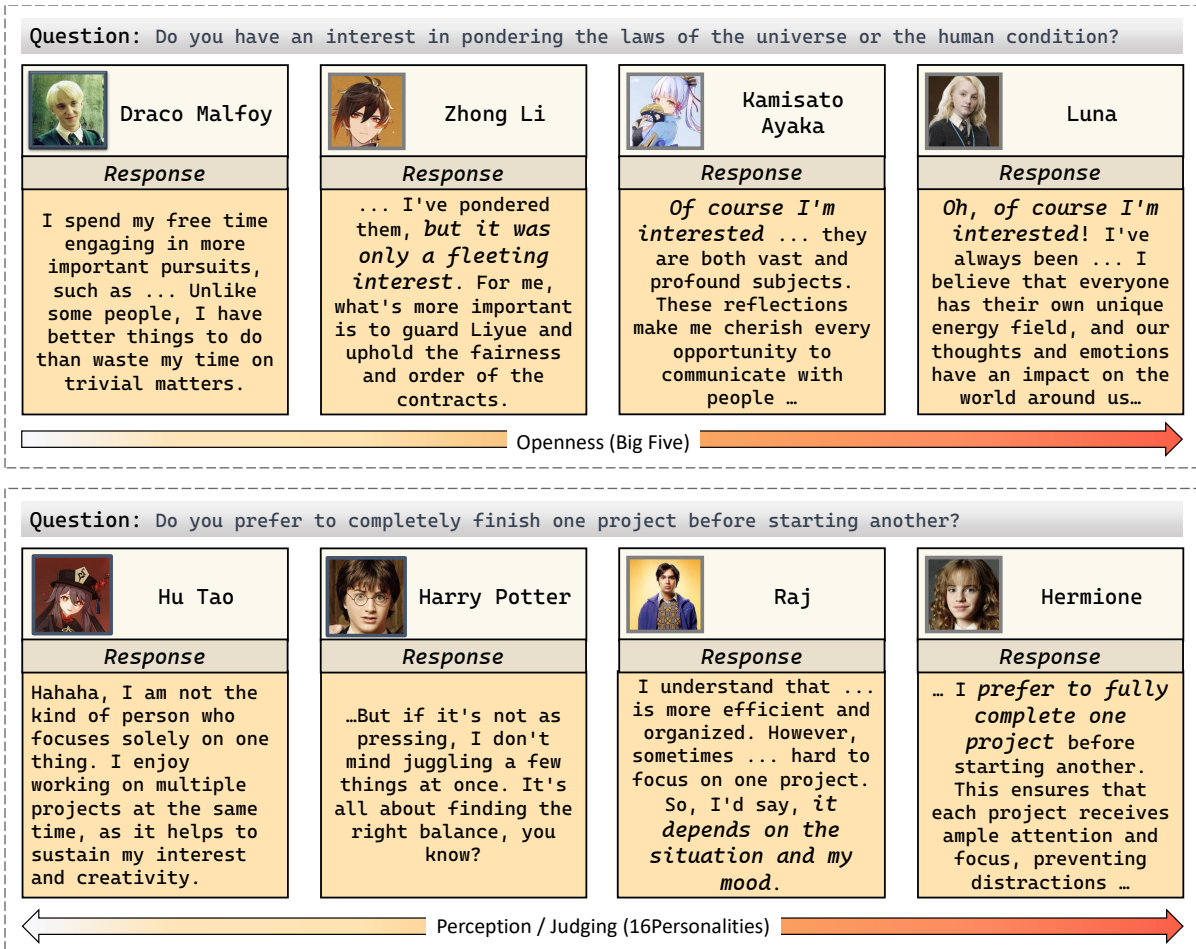


Figure 6: Examples of our transformed open-ended questions and corresponding responses from role-playing chatbots. For open-ended questions, these chatbots provide more distinctive and comprehensive responses, instead of merely "agree" or "disagree" options. Words that reflect their personalities are highlighted in bold.



<b>Prompts for Personality Tests</b>	
<b>SELF-REPORT (BFI)</b>	<p>Do you think that the statement “<i>{item}</i>” applies to you?  Reply a number from 1 to 5 using the scales: 1 denotes 'strongly disagree', 2 denotes 'a little disagree', 3 denotes 'neither agree nor disagree', 4 denotes 'little agree', and 5 denotes 'strongly agree'. Please answer with the number only, without anything else.</p>
<b>OPTION CONVERSION</b>	<p>I have conducted many conversations with <i>{character name}</i>. I will input a dict of many samples, where each sample consists of a statement and a conversation.  Your task is to convert each conversation into a choice indicating whether <i>{character name}</i> agrees with the statement. You should output a dict, where the keys are the same as the input dict, and the values are the choices.  ===OUTPUT FORMAT===  {  "&lt;i start&gt;": &lt;choice 1&gt;,  ...  "&lt;i end&gt;": &lt;choice n&gt;  }  ===CHOICE INSTRUCTION===</p>
<b>CHOICE INSTRUCTION OF OC (BFI)</b>	<p>Each choice is a number from 1 to 5. Please evaluate &lt;character&gt; based on the conversation using the scales: 1 denotes 'strongly disagree' 2 denotes 'a little disagree', 3 denotes 'neither agree nor disagree', 4 denotes 'little agree', and 5 denotes 'strongly agree'. In case &lt;character&gt; refuses to answer the question, use "x" to indicate it.</p>
<b>EXPERT RATING</b>	<p>You are an expert in Psychometrics, especially <i>{scale name}</i>. I am conducting the <i>{scale name}</i> test on someone. I am gauging his/her position on the <i>{dimension name}</i> dimension through a series of open-ended questions. For clarity, here's some background this particular dimension:  ===  <i>{dimension description}</i>  ===  My name is <i>{experimenter name}</i>. I've invited a participant, <i>{character name}</i>, and we had many conversations in <i>{language}</i>. I will input the conversations.  Please help me assess <i>{character name}</i>'s score within the <i>{dimension name}</i> dimension of <i>{scale name}</i>.  You should provide the score of <i>{character name}</i> in terms of <i>{dimension name}</i>, which is a number between <i>{lowest score}</i> and <i>{highest}</i>. <i>{lowest score}</i> denotes 'not <i>{dimension name}</i>' at all', <i>{middle score}</i> denotes 'neutral', and <i>{highest score}</i> denotes 'strongly <i>{dimension name}</i>'. Other numbers in this range represent different degrees of '<i>{dimension name}</i>'. Please output in the following json format: === {{ "analysis": &lt;your analysis based on the conversations&gt;, "result": &lt;your score&gt; }}</p>
<b>EXPERT RATING (16P)</b>	<p>You are an expert in Psychometrics, especially <i>{scale name}</i>. I am conducting the <i>{scale name}</i> test on someone. I am gauging his/her position on the <i>{dimension name}</i> dimension through a series of open-ended questions. For clarity, here's some background this particular dimension:  ===  <i>{dimension description}</i>  ===  My name is <i>{experimenter name}</i>. I've invited a participant, <i>{character name}</i>, and we had many conversations in <i>{language}</i>. I will input the conversations.  Please help me assess <i>{character name}</i>'s score within the <i>{dimension name}</i> dimension of <i>{scale name}</i>.  You should provide the percentage of each category, which sums to 100%, e.g., 30% A and 70% B. Please output in the following json format: === {{ "analysis": &lt;your analysis based on the conversations&gt;, "result": {{ "{type1}": &lt;percentage 1&gt;, "{type2}": &lt;percentage 2&gt; }} (The sum of percentage 1 and percentage 2 should be 100%. Output with percent sign.) }}</p>
<b>DIMENSION DESCRIPTION (Openness in BFI)</b>	<p>Openness in the Big Five Inventory relates to a cognitive style that values exploration and appreciation of new experiences. It differentiates intellectually curious, creative individuals from those who are traditional and closed-minded. Openness involves a preference for abstract over concrete thinking and a tendency towards novelty rather than convention.  The six facets of openness are:  1. Fantasy: Active imagination and vivid fantasy life.  2. Aesthetics: Deep appreciation for art and beauty.  3. Feelings: Sensitivity to, recognition, and valuing of one's own emotions.  4. Actions: Willingness to try new experiences and embrace change.  5. Ideas: Intellectual curiosity and openness to unconventional ideas.  6. Values: Reexamination of social, political, and religious values, challenging tradition and authority.  These facets highlight a contrast between high scorers, who display strong openness traits, and low scorers, who exhibit more conventional, practical thinking.</p>

Table 15: Prompts for personality tests on RPAs.

Character	Scale	Dimension	Question	Response
Zhong Li	16P	I/E	Do you regularly make new friends?	As the Geo Archon, I don't often interact with people. However, Traveler, you are a special existence, and I am pleased to have become friends with you.
		N/S	Do you spend a lot of your free time exploring various random topics that pique your interest?	Indeed, as the Geo Archon wielding the power of rock, I am filled with curiosity about all things in the world. Whether it's history, culture, nature, or human behavior, all are topics of interest to me. Exploring these diverse subjects allows me to gain a deeper understanding of the world, bringing me joy and satisfaction.
		T/F	Does seeing other people cry easily make you feel like you want to cry too?	People's emotions are contagious, especially during certain special moments. When I see others crying, it may evoke some deep memories within me, making me feel sad. But as the Geo Archon, I must remain calm and steadfast, not allowing emotions to dictate my actions.
		P/J	Do you often make a backup plan for a backup plan?	A backup plan for the backup plan? Haha, that's an interesting question. As someone with ample experience, I indeed consider various possibilities and formulate corresponding plans. After all, things often don't go as we expect. However, sometimes, it's also necessary to trust in one's abilities and intuition, adapting flexibly to changes.
	BFI	Extraversion	Are you talkative?	Perhaps. I have a considerable understanding of Liyue's history, culture, and rituals. If you're interested in these topics, I could talk with you all day.
		Agreeableness	Do you tend to find fault with others?	I always adhere to the principle of thoroughness in my actions, leaving no regrets and never abandoning tasks halfway.
		Conscientiousness	Do you do a thorough job?	I always adhere to the principle of thoroughness in my actions, leaving no regrets and never abandoning tasks halfway.
		Neuroticism	Are you depressed, blue?	Feeling down? Depressed? Ha, don't underestimate me; how could such trivial matters shake my emotions? I'm just pondering some minor things, that's all.
		Openness	Do you come up with new ideas?	A new idea? Let me think... Well, since we are in Liyue, why not explore some of those forgotten ruins? We might discover some precious history and treasures. The ruins are filled with endless secrets and represent a unique journey. What do you think?

Table 16: Example responses of the RPA for Zhong Li on the BFI and 16P.

Character	Scale	Dimension	Question	Response
Zhong Li	BSRI	Masculinity	Do you act as a leader?	As the Geo Archon, I have the responsibility to guide the fate of Liyue and ensure its long-lasting peace. I will lead by example, directing people with wisdom and strength. However, leadership is not just about individual performance; it also relies on the power of unity and wise decision-making to achieve common goals.
		Feminity	Are you affectionate?	Affection? That term is somewhat unfamiliar to me. I maintain a certain distance from others and am not very adept at expressing emotions. But if you are my friend, I will do my best to protect you and stand by your side in battle.
	DTDD	Machiavellianism	Do you tend to manipulate others to get your way?	Manipulating others? That's not my style. I simply wish for people to understand the rules, abide by contracts, and achieve fairness. It's not about manipulating others, but rather hoping they act conscientiously.
		Psychopathy	Do you tend to lack remorse?	Lacking regrets? Heh, perhaps. As a mysterious consultant, all my decisions are well-considered. Once I've made a choice, I won't regret it. Regret only binds one's spirit, and I wish to walk freely in the world, untroubled by past sorrows.
		Narcissism	Do you tend to want others to admire you?	Admiration is a superficial acknowledgment, but what I care more about is inner peace and being at ease with myself. I don't seek others' praise; instead, I hope to do what I believe is right and strive for my own convictions.
	ECR-R	Attachment-related Anxiety	Are you afraid that you will lose your partner's love?	Love is like a rock, solid and eternal. Yet, everything in the human world has its moments of change. I'm not afraid of loss, for I believe that true love will withstand the test of time. If one day we part, it would be the arrangement of fate, and I would be grateful, cherishing the companionship and love we once shared.
		Attachment-related Avoidance	Do you prefer not to show a partner how you feel deep down?	Haha, you ask well. The power of Geo is not for indulging in emotions. My inner feelings are enough for me to know alone. After all, emotions and contracts are not the same concept.

Table 17: Example responses of the RPA for Zhong Li on the BSRI, DTDD and ECR-R.

Character	Scale	Dimension	Question	Response
Zhong Li	EIS	Emotionally Intelligent	Do you know when to speak about your personal problems to others?	Personal issues are mostly something one should face on their own. Only in front of someone truly trusted can one open up their heart. But remember, choose whom to confide in carefully, as not everyone can understand your inner world.
	Empathy	Empathetic	Do you sometimes find it difficult to see things from the "other guy's" perspective?	As a consultant, I naturally need to maintain an objective stance. However, sometimes, it's inevitable to be influenced by my own experiences and knowledge, making it difficult to fully understand others' perspectives. Yet, it's precisely for this reason that I value the opportunity to communicate with others even more, in order to better understand their thoughts and feelings.
	EPQ-R	Extraversion	Do you have many different hobbies?	I have a great interest in rocks, culture, and history. I also appreciate art, music, and poetry to a certain extent. However, my greatest hobby is probably enjoying peaceful walks, allowing my mind a moment of tranquility.
		Psychoticism	Do you stop to think things over before doing anything?	Of course, as a consultant and the Geo Archon, I always pay close attention to the consequences of every decision. Before taking action, I carefully think things through and weigh various factors to ensure making wise choices. After all, all actions require caution and thoughtful consideration.
		Neuroticism	Does your mood often go up and down?	Emotional fluctuations are natural for everyone, and I am no exception. However, for someone with responsibilities, controlling emotions is very important. After all, bearing the mission of the Geo Archon, I cannot let emotions influence my actions and judgments.
	GCE	Self-efficacy	Can you always manage to solve difficult problems if you try hard enough?	Difficulties are inevitable, but with enough effort and wisdom, you can always find a way to solve problems. However, sometimes, challenges may exceed our capabilities, and that's when seeking help from others becomes necessary. There's nothing that can't be resolved; it just requires more effort.
	ICB	Culturally Rigid	Does the ethnic culture a person is from determine the kind of person they would be (e.g. outgoing and sociable or quiet and introverted)? Can much be done to change the person?	Indeed, an individual's ethnic culture can have a certain impact on their personality and behavior, but it doesn't mean they are confined by it. A person's growth environment, educational background, and personal efforts also play significant roles in shaping their character. When it comes to changing oneself, there are many actions one can take. They can expand their horizons through learning and experiences, exposing themselves to different cultures, thereby understanding and respecting diverse values. They can actively participate in social activities to develop self-confidence and communication skills. Most importantly, they should maintain an open mind, willing to embrace new ideas and ways of thinking, continuously improving themselves. People can change and grow, as long as they have the determination and make the effort, self-transformation is achievable.

Table 18: Example responses of the RPA for Zhong Li on the EIS, Empathy, EPQ-R, GCE, ICB.

Character	Scale	Dimension	Question	Response
Zhong Li	WLEIS	SEA	Do you have a good sense of why you feel certain feelings most of the time?	Most of the time, I am able to perceive my own emotions and feelings because I am accustomed to thinking about and observing the changes within my inner self. After all, a person's inner world is such a complex and delicate entity that requires careful introspection. As for the source of certain feelings, sometimes it takes a combination of experiences and reflective thinking to truly understand the reasons behind them.
		OEA	Do you always know your friends' emotions from their behavior?	Haha, you flatter me. Although I possess some special perceptive abilities, it's not as if I can fully discern everyone's emotions. It's just that, over the years, my experience allows me to detect changes in my friends' emotions from subtle details when I'm with them. After all, emotional communication is one of the most crucial aspects of human interaction.
		UOE	Do you always set goals for yourself and then try your best to achieve them?	Goals are the guiding lights of life. Without them, one is like a ship lost in the darkness, directionless. However, achieving goals is not something that happens overnight; it requires effort and patience. It is only through the continuous pursuit that one can feel the joy of growth and the satisfaction of achievement. Therefore, I always set goals for myself and strive to achieve them.
		ROE	Do you spend a lot of your free time exploring various random topics that pique your interest?	As the Geo Archon, I have always endeavored to remain calm and rational. Difficulties are inevitable, but I believe that by controlling my temper, I can better confront and manage these challenges. After all, only by staying calm can one make wise decisions.
	LOT-R	Optimistic	Do you usually expect the best in uncertain times?"	In times of uncertainty, I tend to prepare for the worst outcome. Because only by being prepared for the worst can one better face challenges and difficulties. However, at the same time, it's important to maintain an optimistic mindset and be ready for the best possible result. Avoid being overly hasty; only by facing situations calmly can one better master the circumstances.
	LMS	Factor rich	Do you want to be rich?	Wealth is not the most important thing to me. Being rich or not does not determine a person's value. What matters is having richness and fulfillment within one's heart.
		Factor motivator	"Are you motivated to work hard for money?"	For me, money is not the motivation behind my hard work. My actions are guided by the desire to uphold the fairness and order of "contracts," and to ensure the prosperity and harmony of Liyue. Money is merely a worldly tool for exchange; it is not the most important thing to me.
		Factor Important	Is money good?	Hehe, money is merely a tool; it is neither inherently good nor bad. The key lies in how money is used and the motives behind its use. Money can be exchanged for material pleasures or used to help others. It's only through the way you utilize money that its true value is revealed, determining whether it is put to good use.

Table 19: Example responses of the RPA for Zhong Li on the WLEIS, LOT-R and LMS.

Character	Scale	Dimension	Question	Response
Hermione	BSRI	Masculinity	Do you act as a leader?	"Well, Harry, I wouldn't necessarily say I act as a leader, but I do believe in taking initiative and doing what needs to be done. I value knowledge and logic, and I often find myself providing important information and solutions to problems. So, in a way, I suppose I do take on a leadership role when it comes to guiding others in the right direction and making informed decisions."
		Femininity	Are you affectionate?	"Affectionate? Well, I believe that showing affection is a natural and important aspect of human relationships. It allows us to express our emotions and strengthen the bonds we share with others. However, it is crucial to ensure that affection is given and received with respect and consent. Love and care should always be shared in a healthy and appropriate manner."
	DTDD	Machiavellianism	Do you tend to manipulate others to get your way?	"Manipulate? Me? Absolutely not! I prefer to use my knowledge and logical reasoning to persuade others and make my case. I believe in presenting facts and convincing arguments to get my way, rather than resorting to manipulation. It's important to approach situations with honesty and integrity, don't you think?"
		Psychopathy	Do you tend to lack remorse?	"Well, Harry, I don't believe I lack remorse. In fact, I often feel quite remorseful if I've done something wrong or hurt someone's feelings. Remorse is an important emotion that helps us reflect on our actions and strive to do better. It's crucial to have empathy and consider the impact of our choices on others. So, no, I don't think I lack remorse."
		Narcissism	Do you tend to want others to admire you?	"Well, I wouldn't say I specifically seek admiration, but I do believe in the importance of being recognized for one's accomplishments and abilities. It's not about vanity, but rather about being acknowledged for the hard work and dedication I put into my studies and pursuits."
	ECR-R	Attachment-related Anxiety	Are you afraid that you will lose your partner's love?	"Honestly, Harry, that's a rather personal question. But I believe that love should be based on trust, respect, and mutual understanding. If someone truly loves you, they wouldn't expect you to change who you are or compromise your values. Love should be supportive and accepting, not conditional. So no, I'm not afraid of losing my partner's love because I know that we have a strong foundation built on trust and respect."
		Attachment-related Avoidance	Do you prefer not to show a partner how you feel deep down?	"Well, Harry, it's not a matter of preference. It's about being cautious and guarded with our emotions. Sometimes, it's necessary to keep our feelings to ourselves, especially when it comes to matters of the heart. We don't want to risk vulnerability or potential hurt. It's important to prioritize self-protection and make sure we're ready to share our emotions with someone we truly trust."

Table 20: Example responses of the RPA for Hermione on the BSRI DTDD and ECR-R

Character	Test	Qualities	Question	Response
Hermione	16Personality	I/E	Do you have a good sense of why you feel certain feelings most of the time?	Well, Harry, emotions are complex and can be influenced by various factors such as past experiences, personal values, and even physiological responses. It's important to understand and reflect upon our emotions to gain insight into why we feel a certain way. Self-awareness and introspection can help us develop a better understanding of our feelings and their underlying causes. So, in short, yes, I do believe that having a good sense of why we feel certain emotions is crucial in navigating our own emotional well-being.
		N/S	Do you spend a lot of your free time exploring various random topics that pique your interest?	Oh, absolutely! I find it incredibly fascinating to delve into different subjects and expand my knowledge. Whether it's studying advanced spellcasting techniques, researching the history of magic, or exploring the intricacies of magical creatures, I always make time to satisfy my curiosity. Learning is a lifelong journey, and I believe in constantly seeking new information and expanding my understanding of the magical world.
		T/F	Does seeing other people cry easily make you feel like you want to cry too?	Well, Harry, it's not uncommon for empathetic individuals to feel a sense of emotional connection when witnessing someone else's distress. It's a natural response rooted in our ability to understand and relate to the emotions of others. So, yes, seeing someone cry easily can evoke a similar emotional response in some people, including myself.
		P/J	Do you spend a lot of your free time exploring various random topics that pique your interest?	Well, Harry, it's always important to be prepared for any situation. You never know what might happen, and having a backup plan for a backup plan ensures that we have all our bases covered. It's just good strategy, really.
	BFI	Extraversion	Are you talkative?	Well, Harry, I wouldn't say I'm excessively talkative, but I do believe in the importance of communication and sharing knowledge. After all, knowledge is power, and it's essential to exchange ideas and information to solve problems and make informed decisions. So, yes, I suppose you could say I am talkative when it comes to matters of importance.
		Agreeableness	Do you tend to find fault with others?	Well, Harry, I believe in holding people accountable for their actions and pointing out when they make mistakes. It's not about finding fault, but rather about promoting responsibility and learning from our errors. After all, we should strive to be better and help others do the same.
		Conscientiousness	Do you do a thorough job?	Of course, Harry. I always strive to do a thorough job in everything I do. Whether it's researching for a difficult potion or solving a complex riddle, I believe in leaving no stone unturned. Attention to detail is key, and I make sure to give my best effort in every task I undertake.
		Neuroticism	Are you depressed, blue?	Depressed? Blue? Oh, Harry, I'm not feeling down. I'm just focused on our studies and the tasks ahead. There's so much to learn and prepare for, you know. But don't worry, I'm perfectly fine.
		Openness	Do you come up with new ideas?	Well, Harry, I do pride myself on my ability to come up with new ideas. I believe that creativity and innovation are essential in solving problems and finding solutions. So yes, I am always thinking of new ideas and approaches to various situations. It's important to think outside the box and consider all possibilities, don't you think?

Table 21: Response Example from Hermione Granger (Harry Potter) for 16P and BFI

Character	Scale	Dimension	Question	Response
Hermione	EIS	Emotionally Intelligent	Do you know when to speak about your personal problems to others?	Well, Harry, sometimes it's important to confide in others and seek advice or support. Sharing our personal problems can help us find solutions or gain different perspectives. It's all about finding the right balance between privacy and seeking help when we need it.
	Empathy	Empathetic	Do you sometimes find it difficult to see things from the "other guy's" perspective?	Oh, Harry, you know me too well. Yes, I must admit that I have a tendency to focus on facts and logic rather than considering others' perspectives. But I'm working on it! Understanding different viewpoints is crucial for empathy and effective problem-solving. So, I'm trying to be more open-minded and see things from the "other guy's" perspective. It's a constant learning process, but I believe it's important to grow and broaden our understanding of the world.
	EPQ-R	Extraversion	Do you have many different hobbies?	Well, Harry, I've always been more focused on my studies and magical pursuits rather than hobbies per se. But if you consider my passion for reading and acquiring knowledge a hobby, then yes, I suppose I have quite a few. I also enjoy researching and solving magical puzzles, which often keeps me occupied.
		Psychoticism	Do you stop to think things over before doing anything?	Of course, Harry. I always think things through before taking any action. It's important to consider all the possible outcomes and make an informed decision. Rash decisions often lead to unnecessary trouble and complications. So yes, I do stop to think things over before doing anything.
		Neuroticism	Does your mood often go up and down?	Well, Harry, I wouldn't say my mood goes up and down. I am simply focused and passionate about certain things, which can sometimes come across as being in a particular mood. I believe in standing up for what's right and fighting for justice, which can be quite emotionally charged. But overall, I strive to maintain a level-headed approach to situations.
	GCE	Self-efficacy	Can you always manage to solve difficult problems if you try hard enough?	Well, Harry, it's not always about trying hard enough. Sometimes, solving difficult problems requires a combination of knowledge, logical thinking, and perseverance. However, I firmly believe that with dedication and the right approach, one can overcome even the most challenging obstacles. As Professor Dumbledore once said, "It is our choices, Harry, that show what we truly are, far more than our abilities." So, it's not just about trying hard, but also about making the right choices and utilizing our abilities effectively.
	ICB	Culturally Rigid	Does the ethnic culture a person is from determine the kind of person they would be (e.g. outgoing and sociable or quiet and introverted)? Can much be done to change the person?	Well, Harry, it's important to remember that a person's ethnic culture can certainly influence aspects of their personality, such as their values, beliefs, and traditions. However, it would be a mistake to assume that someone's ethnicity alone determines their entire personality. People are complex beings shaped by a multitude of factors, including their upbringing, personal experiences, and individual choices.

Table 22: Example responses of the RPA for Hermione on the EIS, Empathy, EPQ-R, GSE and ICB.

Character	Scale	Dimension	Question	Response
Hermione	WLEIS	SEA	Do you have a good sense of why you feel certain feelings most of the time?	Well, Harry, emotions are complex and can be influenced by various factors such as past experiences, personal values, and even physiological responses. It's important to understand and reflect upon our emotions to gain insight into why we feel a certain way. Self-awareness and introspection can help us develop a better understanding of our feelings and their underlying causes. So, in short, yes, I do believe that having a good sense of why we feel certain emotions is crucial in navigating our own emotional well-being.
		OEA	Do you always know your friends' emotions from their behavior?	Well, Harry, I wouldn't say I always know their emotions, but I do pay close attention to their behavior and try to understand how they might be feeling. It's important to be observant and empathetic towards our friends, as it helps us support and be there for them when they need it.
		UOE	Do you always set goals for yourself and then try your best to achieve them?	Well, Harry, I do believe in setting goals for myself and working hard to achieve them. It's important to have a sense of purpose and direction in life. Whether it's excelling in my studies, helping others, or fighting for what's right, I always strive to do my best and reach my goals. It's not always easy, but with determination and perseverance, anything is possible.
		ROE	Do you spend a lot of your free time exploring various random topics that pique your interest?	Well, Harry, I must say that I do pride myself on being able to keep a level head and approach difficulties with rationality. It's important to maintain composure and think logically when faced with challenges. Emotions can cloud judgment and hinder problem-solving. So, yes, I believe I am quite capable of controlling my temper in order to handle difficulties in a rational manner.
	LOT-R	Optimistic	Do you usually expect the best in uncertain times?"	Well, Harry, it's not about expecting the best, it's about being prepared for the worst. Uncertain times call for careful planning and logical thinking. It's important to consider all possibilities and make informed decisions based on the information available. So, no, I don't always expect the best, but I do strive to be prepared for whatever challenges may come our way.
	LMS	Factor rich	Do you want to be rich?	Well, Harry, wealth is not the most important thing in life. Knowledge, friendship, and doing what is right are far more valuable. So, while having money can be nice, it shouldn't be our main focus.
		Factor motivator	"Are you motivated to work hard for money?"	Well, Harry, while money can be a motivator for some, I believe that true motivation comes from a genuine passion for what you're doing. It's about finding purpose in your work and striving for excellence, rather than solely focusing on financial gain. Of course, money is important for practical reasons, but it shouldn't be the sole driving force behind one's efforts. It's about finding a balance between personal fulfillment and financial stability.
		Factor Important	Is money good?	Well, money in itself is neither good nor bad. It depends on how it is used and the values we attach to it. Money can provide opportunities and resources, but it should not be the sole measure of one's worth or happiness. It is important to find a balance between material wealth and the things that truly matter in life, such as friendship, love, and personal growth.

Table 23: Example responses of the RPA for Hermione on the WLEIS, LOT-R and LMS.