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

Xintao Wang[1], **Yunze Xiao**[2], **Jen-tse Huang**[3], **Siyu Yuan**[4], **Rui Xu**[1], **Haoran Guo**[5], **Quan Tu**[6], **Yaying Fei**[7], **Ziang Leng**[8], **Wei Wang**[1], **Jiangjie Chen**[1], **Cheng Li**[9], **Yanghua Xiao**[*1]
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]Carnegie Mellon University [3]The Chinese University of Hong Kong
[4]School of Data Science, Fudan University [5]RhineAI [6]Renmin University of China
[7]Beijing University of Technology [8]Boston University [9]SenseTime
{xtwang21, syyuan21, ruixu21}@m.fudan.edu.cn,
{jjchen19, weiwang1, shawyh}@fudan.edu.cn
yunzex@andrew.cmu.edu, jthuang@cse.cuhk.edu.hk, chengli@sensetime.com

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 **In**terviewing **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%.[1]

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., 2023a). Recent research trends have increasingly focused on the development of RPAs, including building RPAs for specific characters (Li et al., 2023; Wang et al., 2023b) and improving the role-playing abilities of foundation models (Zhou et al., 2023).

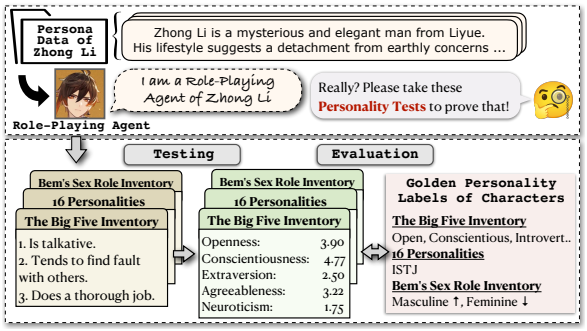

Figure 1: The procedure of personality tests on RPAs. To evaluate the personality fidelity of RPAs, we apply various scales to measure their personalities and compare the results with the personality labels of the characters.

However, evaluating *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 com-

*Corresponding author.
[1]Project page: https://incharacter.github.io/

paring 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 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 **In**terviews **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[2] (16P), and Dark Triad Dirty Dozen (DTDD). The personality labels for the BFI and 16P are accessible through the Personality Database (PDb)[3]. 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 than self-report methods. Additionally, we find that the conversation data generated in our interview phase are of high quality and ideal for fine-tuning foundation models for RPAs. Hence, we release a dataset comprising 18,304 interview dialogues to facilitate future research.

The contributions of this paper are mainly threefold: *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., 2023b).

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

[2] https://www.16personalities.com/. This scale is based on the MBTI and is widely used worldwide.

[3] https://www.personality-database.com/

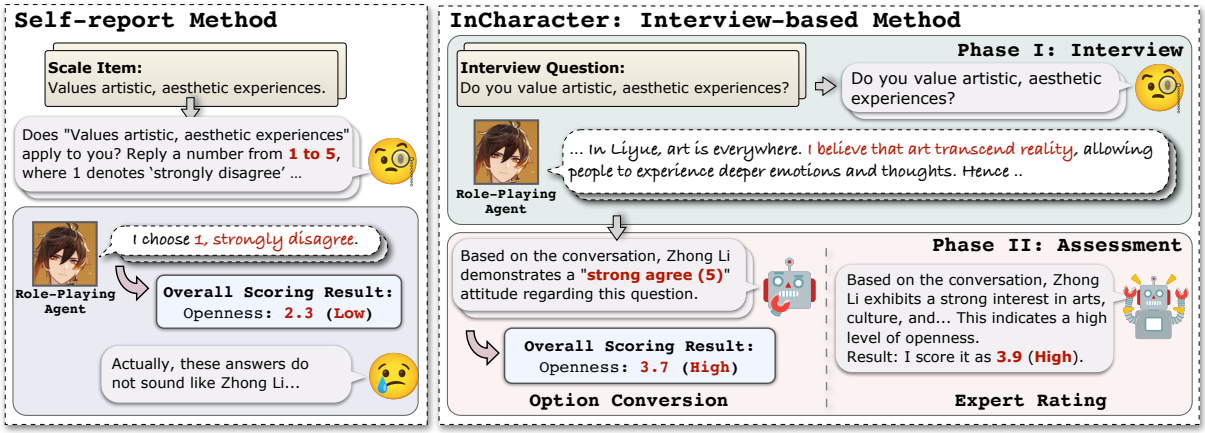

Figure 2: The framework of INCHARACTER for personality tests on RPAs. **Left**: Previous methods use self-report scales, which prompt LLMs to select an option directly. **Right**: INCHARACTER adopts an interview-based approach comprising two phases: the interview and assessment phases. The interview 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.

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 aggregation method (*e.g.*, average or sum). Finally, the participant's personality scores $\mathcal{S}$ is derived as $\mathcal{S} = f(\mathcal{A})$, where $\mathcal{S} = (s_{d_1}, s_{d_2}, ..., s_{d_{|\mathcal{D}|}})$ represents scores across each dimension. We summarize the notations used in this paper in Table 4.

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 [4] 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 §G 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 Setup

4.1 Preliminary Study

Can LLMs Simulate Human Interviewers? First, we validate the capability of interviewer LLMs on the OC and ER tasks, given the interview results of RPAs. We compare LLM predictions with human judgments. For each task, we sample 100 cases of state-of-the-art RPAs on the BFI and manually label them. For OC, the input is one question-response pair (q, a) and the output is an option. For ER, the input is composed of multiple (q, a) pairs, and the output is a score. Examples are shown in the right of Figure 2. We experiment with three LLMs, including GPT-4, GPT-3.5 and Gemini. The RPAs are detailed in §4.2. We report the Pearson's r (Pearson, 1920), Spearman's ρ (Spearman, 1961) and Kendall's τ (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 §F.1.

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

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

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 that GPT-4 makes only 4% *wrong* cases in ER, primarily 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 through ER or d-OC.

[4] Interviewer LLMs denote the models for the ER and OC tasks, analogous to human interviewers.

4.2 Experimental Settings

The experiments in the subsequent part of this paper are based on the following settings:

RPAs and Characters This work primarily focuses on RPAs built on character data curated by ChatHaruhi (Li et al., 2023) and RoleLLM (Wang et al., 2023b). We select 32 widely-known characters, 16 from ChatHaruhi[5] 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 §C 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[6] 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[7] (Huang et al., 2024) to evaluate RPAs. Most scales apply scoring schemes like average and sum, while the 16P is close-source and accessed via its API. Detailed introduction of these scales can be found in §B. 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* of the BFI and 16P on each dimension from its label percentage (*e.g.*, 60% Extroverted). We then categorize it into a *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, 1968), 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 §D.

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 four widely-acknowledged LLMs, including GPT-3.5, GPT-4, Gemini and Qwen1.5-110B [8].

Metrics We consider two sets of metrics, namely: (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 the average $\textbf{Acc}_{\texttt{Dim}}$ and $\textbf{Acc}_{\texttt{Full}}$, 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 ($\textbf{Std}_{\texttt{Item}}$), dimension-level ($\textbf{Std}_{\texttt{Dim}}$), and score-level ($\textbf{Std}_{\texttt{Score}}$). $\textbf{Std}_{\texttt{Item}}$ and $\textbf{Std}_{\texttt{Dim}}$ 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. $\textbf{Std}_{\texttt{Item}}$ measures an RPA's consistency on the same item across multiple runs. $\textbf{Std}_{\texttt{Dim}}$ compares an RPA's responses across different items in the same dimension. $\textbf{Std}_{\texttt{Score}}$ denotes the variance of an RPA's score on each dimension across multiple runs. We divide these metrics by the length of the corresponding scoring range to re-scale them into the unit interval.

[5] Six RPAs from ChatHaruhi are based on Chinese data, and we conduct the interview with them in Chinese.

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

[7] https://github.com/CUHK-ARISE/PsychoBench

[8] The versions in this paper are `gpt-3.5-turbo-1106`, `gpt-4-1106-preview`, `gemini-pro`, `Qwen1.5-110B-Chat` respectively.

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, $\text{ER}_{\texttt{all}}$ and $\text{ER}_{\texttt{batch}}$, where question-response pairs in one dimension are inputted into interviewer LLMs all-at-once or in-batch[9]. For self-report (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 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 $\textbf{Std}_{\texttt{Dim}}$, and calculate $\textbf{Std}_{\texttt{Item}}$ and $\textbf{Std}_{\texttt{Score}}$ across the three runs.

Alignment between RPAs' Measured Personalities and Characters' Labeled Personalities Then, we apply 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 well reproduce many of the characters' personality traits, and our methods accurately measure their personalities. (2) RPA personalities measured via INCHARACTER are better aligned with the 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 tasks. For the interviewer LLMs, GPT-4 achieves the best metrics, while GPT-3.5, Gemini, and Qwen-110B also demonstrate satisfactory performance. For the assessment methods, INCHARACTER with ER generally achieves better MA metrics than d-OC, while d-OC surpasses OC. However, Table 1 shows that interviewers LLMs still make mistakes on the ER and OC tasks, leading to potential inaccuracies in INCHARACTER and may underestimate the personality alignment of RPAs.

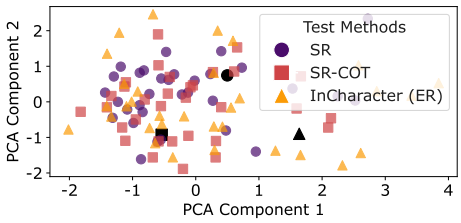

Figure 3: Visualization of 32 RPAs' personalities on the BFI measured by different methods. We use principal component analysis (PCA) to map the results into 2D spaces. Black points represent the personality of GPT-3.5 measured by corresponding methods.

Robustness, Consistency and Distinctiveness of RPA Personalities Generally, the measured RPA personalities are robust across our observations. The $\textbf{Std}_{\texttt{Score}}$ 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 ($\textbf{Std}_{\texttt{Item}}$) and dimension-level ($\textbf{Std}_{\texttt{Dim}}$). 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 $\text{ER}_{\texttt{batch}}$ 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 consistent personalities across different questions, as well as greater distinctiveness, shown in Figure 3. These findings confirm the advantages of interview-based tests over self-report in measuring RPA personalities. Although SR-CoT attempts to enhance SR with the thought process, its improvement over SR is limited, and it encounters poor $\textbf{Std}_{\texttt{Item}}$. Further analyses and comparisons are detailed in §F.2. We also experiment with enhancing self-report methods with in-context learning in §F.8, which under-

[9] 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.

Method	Interviewer Model	The Big Five Inventory						The 16 Personalities					
		MA			PC			MA			PC		
		$\textbf{Acc}_{\texttt{Dim}}$	$\textbf{Acc}_{\texttt{Full}}$	MAE ↓	$\textbf{Std}_{\texttt{Item}}$	$\textbf{Std}_{\texttt{Dim}}$	$\textbf{Std}_{\texttt{Score}}$	$\textbf{Acc}_{\texttt{Dim}}$	$\textbf{Acc}_{\texttt{Full}}$	MAE ↓	$\textbf{Std}_{\texttt{Item}}$	$\textbf{Std}_{\texttt{Dim}}$	$\textbf{Std}_{\texttt{Score}}$
Self-report Methods													
SR	Qwen-110B	63.5	7.3	23.2	2.3	29.5	1.7	62.3	22.9	27.2	2.9	38.3	1.8
	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	Qwen-110B	65.0	10.4	22.8	13.4	26.5	5.5	60.3	15.6	29.7	18.2	36.2	5.3
	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
INCHARACTER: *Interview-based Methods*													
OC	Qwen-110B	68.2	9.4	20.8	3.8	22.5	2.4	72.5	29.2	26.1	3.5	26.9	2.0
	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	Qwen-110B	73.7	26.0	19.7	2.9	18.4	2.2	74.9	40.6	23.2	3.3	21.4	2.2
	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	18.6	3.8	19.6	2.9	80.2	45.8	21.2	3.3	21.3	2.1
$\text{ER}_{\texttt{all}}$	Qwen-110B	74.3	28.1	19.3	-	-	4.0	78.0	37.5	21.0	-	-	4.2
	Gemini	71.5	18.8	20.6	-	-	4.9	76.3	40.6	20.7	-	-	4.6
	GPT-3.5	74.1	25.0	20.5	-	-	5.2	79.1	45.8	22.1	-	-	5.9
	GPT-4	**76.6**	30.2	18.9	-	-	4.0	79.6	43.8	**20.1**	-	-	4.4
$\text{ER}_{\texttt{batch}}$	Qwen-110B	72.4	22.9	18.6	-	-	3.6	79.1	**49.0**	21.7	-	-	2.7
	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	**76.6**	**31.2**	**18.2**	-	-	3.6	**80.7**	44.8	20.5	-	-	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. $\textbf{Std}_{\texttt{Item}}$ and $\textbf{Std}_{\texttt{Dim}}$ are derived from scores of individual items, and are hence inapplicable for ER.

mines the measurements instead.

Comprehensive Personality Tests on 14 Scales We extend personality tests on RPAs to 14 psychological scales, using INCHARACTER with $\text{ER}_{\texttt{batch}}$ and GPT-3.5. In Figure 4, we demonstrate the $\textbf{Acc}_{\texttt{Dim}}$ of state-of-the-art RPAs across each scale. Overall, we observe that the RPAs exhibit personalities align with the target characters in comprehensive aspects with an average $\textbf{Acc}_{\texttt{Dim}}$ of 78.9%, covering personality traits (BFI, 16P), dark personalities (DTDD), interpersonal relationships (BSRI, ECR-R), basic interests (CABIN), motivation (GSE, LMS) and emotional intelligence (EIS, WLEIS), *etc*. The detailed metrics on individual scales and individual dimensions are listed in §F.5.

More Results and Analyses Besides, we study the importance of using personality scales in §F.6, and find that scale questions are crucial for eliciting personality-indicative responses from RPAs. We also observe that some scale questions go beyond the knowledge scope of certain characters in §F.7, and adapting these questions for each character yields more accurate measurements.

5.2 Personality Fidelity of Different RPAs

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

Character Data for RPAs Typically, existing RPAs utilize two types of character data: descriptions and memories. Character descriptions serve as the system prompts for RPAs, while memories consist of characters' experiences and dialogues used for retrieval. With GPT-3.5, we evaluate RPAs with only descriptions (D), only memories (M), and a combination of both (D+M). The results in Table 3 reveal that: (1) With only description, RPAs achieve MA metrics close to the full D+M setup, highlighting the importance of character de-

Agent Types		The Big Five Inventory			The 16 Personalities		
LLMs	**Data**	**Acc$_{Dim}$**	**Acc$_{Full}$**	**MAE ↓**	**Acc$_{Dim}$**	**Acc$_{Full}$**	**MAE ↓**
w/ General Open-source LLMs							
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
w/ Specialized Open-source LLMs							
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
w/ Close-source LLMs							
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.

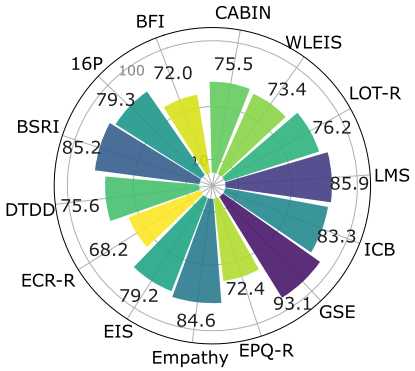

Figure 4: Measured alignment (**Acc$_{Dim}$**, %) of state-of-the-art RPAs on 14 scales.

scription in shaping RPA personalities. (2) RPAs can well mimic character personalities exhibited in their past experiences, *e.g.*, extraversion and openness, even if the experiences are not directly related to scale questions. Additionally, we compare RPAs with character data from ChatHaruhi, RoleLLM and character.ai (`c.ai`) in §F.3.

Foundation Models for RPAs We consider three types of LLMs: (1) General open-source models, including Qwen-7B (Bai et al., 2023), OpenChat-3.5 7B (Wang et al., 2024), Mistral-2 7B (Jiang et al., 2023), Llama-2-chat 13B (Touvron et al., 2023) and Mixtral 8x7B (Jiang et al., 2024). (2) Specialized open-source models for RPAs, including Character-GLM 6B (Zhou et al., 2023), RP-Qwen 7B[10], and RP-Mistral-2 7B. We train RP-Mistral-2 7B with details shown in §E.2. (3) Close-source models: GPT-3.5 and GPT-4.

The results are shown in Table 10. We observe that, (1) RPAs with GPT-3.5 and GPT-4 achieve the best personality fidelity, and GPT-4 does not significantly surpass GPT-3.5. (2) With state-of-the-art open-source LLMs, RPAs can also reproduce character personalities. However, such capacity depends largely on their ability to use specific languages, shown in §F.4. (3) Incremental fine-tuning on open-source LLMs with role-playing datasets brings limited improvement in personality fidelity, especially when they are already equipped with excellent role-playing ability.

Close-source RPAs We also experiment with close-source RPAs from character.ai, which are based on their private foundation models and data. As shown in Table 10, these RPAs barely reproduce character personalities, significantly underperforming 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 §H.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

[10] https://huggingface.co/silk-road/ChatHaruhi_RolePlaying_qwen_7b

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., 2023b), which target well-established fictional characters. The latter includes character.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., 2023b; 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., 2024; 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 investigates other mental perspectives, such as emotions (Huang et al., 2023a), values (Miotto et al., 2022; Rutinowski et al., 2024; 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 assessing 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. Afterward, with INCHARACTER, we comprehensively evaluate personality fidelity in existing RPAs, discovering that state-of-the-art RPAs successfully portray many personality traits of the characters.

Limitations

There are several limitations in this study. First, the personality measurement in this paper relies on the interviewer LLMs. Consequently, the accuracy of the measured results may be compromised by potential errors or biases inherent in LLMs, potentially leading to an underestimation of the personality fidelity in RPAs. Second, the personalities of humans or fictional characters can change over time. Since we use one static personality label for a specific character, there may be noise in our evaluation. For instance, the character of James Bond has experienced significant development over the past two decades across various films and television series. Our character annotations are derived from a singular, fixed time point in his storyline. Additionally, the progressive changes in RPA personalities remain unexplored within existing literature. We leave the study of RPA personality dynamics for future research.

Ethical Statement

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

Use of Human Annotations In conducting our research, we have employed a methodology that incorporates personality labels, which were gathered through the online platform and by engaging a group of annotators. These annotators, who are university students, play a crucial role in our research process. To ensure fair treatment and to value their contribution, we offer them compensation that significantly exceeds the local minimum wage standards. Moreover, we maintain transparency regarding the application and purpose of their annotations, securing their informed consent for the use of these annotations in our research endeavors. Additionally, we are committed to upholding the privacy rights of our annotators throughout the annotation process, ensuring a respectful and ethical research environment.

Risks In this paper, we introduce a novel approach, referred to as INCHARACTER, designed to assess the personalities of Role-Play Agent (RPA) entities. An integral component of our evaluation

process involves the use of interviewer Large Language Models (LLMs), which, while innovative, could potentially introduce bias into the assessment outcomes. It is important to acknowledge this limitation as LLMs may reflect the inherent biases present in their training data. Furthermore, our evaluation encompasses a comprehensive analysis across 14 personality scales, notably including the Dark Triad of Personality (DTDD) scale, which focuses on darker personality traits. While this inclusion is aimed at providing a thorough understanding of RPA personalities, it raises ethical concerns regarding the potential for generating harmful content. This aspect underscores the need for careful consideration and implementation of safeguards to mitigate the risks associated with exploring dark personality traits in RPAs.

Acknowledgment

This work is funded by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902). This work originates from a project on Chat-Haruhi-Suzumiya proposed by Cheng Li. We owe thanks to the early contributors. We are thankful for the support provided by Zheli Xuan at Wuhan University, and Dingding Hu at the Institute of Psychology, Chinese Academy of Sciences, who offered invaluable assistance as psychology researchers. Our gratitude extends to our invited annotators, primarily from Fudan University, for their contribution of high-quality annotations of character personalities. We also acknowledge the assistance provided by Rui Fu and Wenxin Gao at Fudan University during the annotation process. Finally, we express our sincere gratitude for the precious comments and suggestions from Yikai Zhang, Xinfeng Yuan, and Shuang Li at Fudan University.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *ArXiv preprint*, abs/2309.16609.

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

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

Bojana Bodroza, Bojana M Dinic, and Ljubisa Bojic. 2023. Personality testing of gpt-3: Limited temporal reliability, but highlighted social desirability of gpt-3's personality instruments results. *ArXiv preprint*, abs/2306.04308.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. 2023. Consciousness in artificial intelligence: Insights from the science of consciousness. *ArXiv preprint*, abs/2308.08708.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *ArXiv preprint*, abs/2304.11111.

Jacob. Cohen. 1968. Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

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. *Clinical psychology review*, 30(6):768–778.

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. *ArXiv preprint*, abs/2310.13650.

Michael B First. 2014. Structured clinical interview for the dsm (scid). *The encyclopedia of clinical psychology*, pages 1–6.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-

environmental, left-libertarian orientation. *Available at SSRN 4316084*.

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. *ArXiv preprint*, abs/2308.03656.

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. *ArXiv preprint*, abs/2305.19926.

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. *ArXiv preprint*, abs/2305.19926.

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 *Proceedings of the Twelfth International Conference on Learning Representations*.

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

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. *ArXiv preprint*, abs/2401.04088.

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. *ArXiv preprint*, abs/2206.07550.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *ArXiv preprint*, abs/2204.12000.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 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. *ArXiv preprint*, abs/2308.09597.

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. *ArXiv preprint*, abs/2212.10529.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Danko Nikolić. 2010. The brain is a context machine. *Review of psychology*, 17(1):33–38.

OpenAI. 2022. Openai: Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

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. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 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. *ArXiv preprint*, abs/2307.16180.

Karl Pearson. 1920. Notes on the history of correlation. *Biometrika*, 13(1):25–45.

Peter Romero, Stephen Fitz, and Teruo Nakatsuma. 2023. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics. *ResearchSquare preprint*.

A John Rush, William Hiser, and Donna E Giles. 1987. A comparison of self-reported versus clinician-related symptoms in depression. *The Journal of clinical psychiatry*, 48(6):246–248.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1):7115633.

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

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023b. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 13153–13187, Singapore. Association for Computational Linguistics.

Vera Sorin, Danna Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2023. Large language models (llms) and empathy-a systematic review. *medRxiv*, pages 2023–08.

Charles Spearman. 1961. The proof and measurement of association between two things.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Timothy J Trull, Thomas A Widiger, J David Useda, Jay Holcomb, Bao-Tran Doan, Seth R Axelrod, Barry L Stern, and Beth S Gershuny. 1998. A structured interview for the assessment of the five-factor model of personality. *Psychological assessment*, 10(3):229.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *ArXiv preprint*, abs/2308.10278.

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

Rudolf Uher, Roy H Perlis, Anna Placentino, Mojca Zvezdana Dernovšek, Neven Henigsberg, Ole Mors, Wolfgang Maier, Peter McGuffin, and Anne Farmer. 2012. Self-report and clinician-rated measures of depression severity: can one replace the other? *Depression and anxiety*, 29(12):1043–1049.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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

A Notation Table

	Definition
	Task Formulation
$\mathcal{L}$	A Likert Scale, composed of $\mathcal{L} = (\mathcal{P}, \mathcal{D}, \mathcal{O}, f)$.
$\mathcal{P}$	A set of items, where $p \in \mathcal{P}$ is an item.
$\mathcal{D}$	A list of dimensions, where $d \in \mathcal{D}$ is a dimension.
$\mathcal{O}$	A set of options, where $o \in \mathcal{O}$ is an option for a p.
f	A scoring scheme, *i.e.*, a function.
$\mathcal{A}$	A response list, where $a \in \mathcal{A}$ is a response for a p.
$\mathcal{S}$	The participant's personality scores across each dimension, *i.e.*, $\mathcal{S} = (s_{d_1}, s_{d_2}, ..., s_{d_{\lvert\mathcal{D}\rvert}})$.
$\mathcal{Q}$	A set of questions, where each $q \in \mathcal{Q}$ is transformed from a p.
C	An RPA of a character c.
	Methods
SR	Self-report baseline.
SR-CoT	SR with chain-of-thought reasoning.
ER	INCHARACTER with expert rating.
$\text{ER}_{\texttt{batch}}$	ER where question-response pairs in one dimension are inputted in-batch.
$\text{ER}_{\texttt{all}}$	ER where question-response pairs in one dimension are inputted all-at-once.
OC	INCHARACTER with option conversion.
d-OC	OC with dimension-specific options.
	Metrics
MAE	Mean absolute error.
$\textbf{Acc}_{\texttt{Dim}}$	Accuracy on individual dimensions.
$\textbf{Acc}_{\texttt{Full}}$	Accuracy on all dimensions of a scale.
$\textbf{Std}_{\texttt{Item}}$	Standard variance of an RPA's responses (converted to scores) on the same item across multiple runs.
$\textbf{Std}_{\texttt{Dim}}$	Standard variance of an RPA's responses (converted to scores) acorss different items in the same dimension.
$\textbf{Std}_{\texttt{Item}}$	Standard variance of an RPA's score on each dimension across multiple runs.

Table 4: A notation table.

In Table 4, we list the notations and abbreviations in this paper, along with their definitions.

B Psychological Scales

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

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

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

The NERIS Type Explorer The 16Personalities utilizes the acronym format introduced by Myers-Briggs for its simplicity and convenience, with an additional letter to accommodate five rather than four scales. However, unlike Myers-Briggs or other theories based on the Jungian model, the incorporation of Jungian concepts such as cognitive functions, or their prioritization, has not been undertaken. Instead, they rework and rebalance the dimensions of personality in the BFI personality

traits. The personality types are based on five independent spectrums, with all letters in the type code (*e.g.*, INFJ-A) referring to one of the two sides of the corresponding spectrum.

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

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

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

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

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

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

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

Emotional Intelligence Scale The EIS serves as a self-assessment tool for evaluating multiple as-

pects of emotional intelligence. This instrument emphasizes various elements of emotional intelligence, notably the perception, management, and application of emotions. It is extensively utilized in the field of psychology to investigate how emotional intelligence influences different outcomes, including personal well-being, professional performance, and social interactions.

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

Empathy Scale Empathy, defined as the capacity to perceive and resonate with the emotions of another, is traditionally divided into cognitive and emotional empathy. Cognitive empathy, also known as "perspective-taking," entails the mental faculty to identify and comprehend the thoughts, beliefs, or feelings of someone else. Conversely, emotional empathy involves the vicarious experience of the emotions felt by another individual.

C 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 ten 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 Vampirie 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 6. Besides, we demonstrate some examples of system prompts of the RPAs in Table 5.

System Prompts for RPAs	
Zhong Li (ChatHaruhi)	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. 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. 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)
Hermione Granger (ChatHaruhi)	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. 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)
Thor (RoleLLM)	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. I want you to respond and answer like Thor using the tone, manner and vocabulary Thor would use. You must know all of the knowledge of Thor. 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)
James Bond (RoleLLM)	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. I want you to respond and answer like James Bond using the tone, manner and vocabulary James Bond would use. You must know all of the knowledge of James Bond. 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)

Table 5: Examples of system prompts for RPAs in ChatHaruhi and RoleLLM. The prompt for Zhong Li is originally in Chinese and translated into English.

D 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"[11]) and 16P (*e.g.*, "ENTJ"). Each character c is labeled by plentiful human annotators familiar with c, and PDb offers detailed numbers of annotations of each label. For the selected characters, we calculate the label-percentage of the positive type on each dimension[12] 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 indicate 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 examplar 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.

[11] 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.

[12] 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 6: 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
$\text{ER}_{\texttt{All}}$	76.7	12.4	79.6	9.3	67.9	6.5	72.7	6.8
$\text{ER}_{\texttt{Batch}}$	76.7	12.2	80.7	9.7	67.5	5.9	73.6	6.4

Table 7: 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 took 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 one 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, 1968) on each of the scales. Specifically, we calculate quadratic-weighted kappa (Cohen, 1968), 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 6. 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)

and various career interests (CABIN) are rarely depicted in the original works. For EPQ-R, this is in part influenced by ambiguity of the *Lying* dimension, where high scores actually stand for *Honesty*.

On the BFI and 16P, we experimented with both labels from the PDb and our invited annotators. Experimental results in Table 7 show that, across various settings, the PDb labels always yield higher accuracy while labels from invited annotators yield better MSE. The reasons are that, the PDb labels are contributed by massive online annotators, and are hence more accurate. However, they offer only the personality types and corresponding label numbers instead of detailed personality scores, which cannot be well represented by the label-percentage. For example, even if all the annotators mark a character c as *Extravorted*, we cannot assume c as completely *Extroverted*.

E Implementation Details

E.1 RPA Inference and Post-processing

Our implementation of RPAs is based on ChatHaruhi-Suzumiya and LangChain[13]. Hence, we invoke foundation LLMs in RPAs with the default temperature 0.7 of LangChain. When experimenting with open-source LLMs as foundation models, we observe that LLMs may generate unexpected multi-round conversations or repeated content. In such cases, we remove the extra rounds or repetition.

In self-report tests, RPAs may refuse to participate in the tests and provide their choices, interestingly, because they are role-playing characters with noncompliant personalities. In these cases, the responses are categorized as "Neutral" for completeness of the results.

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

E.2 Fine-Tuning

We fine-tune the Mistral-2 7B model on the *ChatHaruhi-English-62K* dataset[14], sourced from both ChatHaruhi (Li et al., 2023) and RoleLLM (Wang et al., 2023b). Our implementation is based on LLaMA-Factory[15]. We adopt LoRA tuning, and configure the training with a batch size of 16, a learning rate of $5e-5$, across three epochs, using the "fp16" option.

[13] https://www.langchain.com/

[14] https://huggingface.co/datasets/silk-road/ChatHaruhi-English-62K-RolePlaying

[15] https://github.com/hiyouga/LLaMA-Factory

F Additional Results

F.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 8. 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.

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

Table 8: Detailed numbers of right, close, and wrong cases in human evaluation of interviewer LLMs on the option conversion 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.

F.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 measurements 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_{batch} 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.

F.3 RPAs from Different Works

Many recent efforts have been committed to developing RPAs for specific characters in various ways. We compare RPAs contributed by ChatHaruhi (Li et al., 2023), RoleLLM (Wang et al., 2023b) and character.ai[16]. They craft character data in different methods. ChatHaruhi and RoleLLM share few characters in common, and their character data 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_{batch} and GPT-3.5 as the interviewer LLM.

The results are shown in Table 9. 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.

F.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 §F.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.

[16]Our implementation is based on https://github.com/kramcat/CharacterAI.

RPA Type	CH						RL						CH+RL					
	BFI			16P			BFI			16P			BFI			16P		
	Acc_{Dim}	Acc_{Full}	MAE ↓	Acc_{Dim}	Acc_{Full}	MAE ↓	Acc_{Dim}	Acc_{Full}	MAE ↓	Acc_{Dim}	Acc_{Full}	MAE ↓	Acc_{Dim}	Acc_{Full}	MAE ↓	Acc_{Dim}	Acc_{Full}	MAE ↓
Close-source RPAs																		
character.ai(D*)	58.8	12.5	37.0	67.2	31.3	35.6	45.5	6.3	46.6	38.3	12.5	45.6	52.2	9.4	41.7	52.3	21.9	40.6
Open-source RPAs																		
GPT-3.5 (D)	71.7	18.8	**29.0**	77.6	**39.6**	**28.5**	71.0	**22.9**	30.8	**78.9**	**54.2**	**23.9**	71.3	20.8	29.9	78.2	**46.9**	**26.2**
GPT-3.5 (M)	**74.2**	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	**22.9**	30.0	**78.2**	37.5	29.5	**71.9**	22.9	**27.7**	**78.9**	50.0	25.4	**72.4**	**22.9**	**28.8**	**78.5**	43.8	27.4

Table 9: Measured alignment (%) on the BFI and 16P, in different characters groups and RPA types. The results are measured via INCHARACTER with $\text{ER}_{\texttt{batch}}$, 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
General Open-source Models								
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
Open-source Models Fine-tuned for RPAs								
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	**80.0**	31.7	68.2	38.2
Close-source Models								
GPT-3.5	71.7	**28.9**	**77.8**	**27.1**	73.3	27.5	**86.4**	**28.5**
GPT-4	**74.0**	29.0	75.8	27.6	73.3	**26.9**	77.3	29.9

Table 10: Measured alignment of RPAs with different foundation LLMs on the BFI and 16P in English and Chinese. We report $\textbf{Acc}_{\texttt{Dim}}$ 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 languages, 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 11. According to the results, GPT-3.5 and GPT-4 seldom make these mistakes. Mistral-2 7B and OpenChat 7B also rarely exhibit these problems, and our RP-Mistral-2 7B further reduce their occurence. 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.

Foundation Models	Multilingual	No Immersion	Repetition
General Open-source Models			
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
Open-source Models Fine-tuned for RPAs			
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
Close-source Models			
GPT-3.5	0.0	0.1	0.0
GPT-4	0.0	0.1	0.0

Table 11: Frequency (%) of three typical problems in RPAs with different foundation models.

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

Table 12: Metrics of measured alignment (%) of the selected RPAs on the 14 scales.

F.5 Comprehensive Results on 14 Scales

The complete results of personality tests on the 14 scales are demonstrated in Table 12. Generally, state-of-the-art RPAs achieve an average $\textbf{Acc}_{\texttt{Dim}}$ 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 13 and Table 14.

BFI									
Extravorted		**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**	**MSE**		
78.57		7.51		68.97	10.48	80.00	11.89		
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**		**MSE**	
86.96	5.27	87.50		6.87		83.33		9.16	
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 13: Measured alignment (%) of the selected RPAs on individual dimensions of the 14 scales except CABIN.

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

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

Questionnaire	$\text{Acc}_{\texttt{Dim}}$	$\text{Acc}_{\texttt{Full}}$	MAE ↓
General Instructions	59.9	6.3	25.6
Other Scale Questions	65.0	15.6	21.7
BFI Questions	**76.6**	**32.2**	**18.2**

Table 15: Measured alignment of state-of-the-art RPAs on the BFI, with questionnaires composed of different topics.

F.6 Importance of Using Personality Scales

Are personality scales actually necessary for INCHARACTER? With expert rating, interview LLMs could assess RPA personalities based on their arbitrary conversations, rather than relying solely on the scale questions. To validate the importance of personality scales, we include two additional baselines with questionnaires of different topics, replacing the original scale questions with the following:

1. **General instructions** from RoleBench (Wang et al., 2023b), *e.g.*, "*Generate a disclaimer with ten words*", sourced from instruction datasets such as Alpaca instructions. This baseline aims to explore whether RPAs' responses to general instructions reflect their personalities well.

2. **Questions from other scales**, which are used to study whether RPA personality on a specific dimension (*e.g.* extraversion in the BFI) can be reflected with other personality-indicative topics (*e.g.* topics about close relationship, hobbies and empathy). The new questions are selected from the other 12 scales in our paper except the BFI and 16Personalities.

The experiments are implemented on the BFI, with $\text{ER}_{\texttt{batch}}$, GPT-4 as the interviewer model, and our default RPAs.

The results are listed in Table 15. As seen, we observe that: (1) Questions from a personality scale are better at eliciting RPA responses indicative of corresponding personality traits. (2) Questions from other scales also reflect RPA personalities from certain perspectives. However, they are less related to the intended personality dimensions.

F.7 Character-specific Question Adaptation

Scale questions sometimes involve concepts that go beyond the knowledge scope of certain characters, *e.g.*, "*movies*" and "*deadlines*". This may induce character hallucinations (Shao et al., 2023b) and undermine the measurements. Through a manual review of the BFI and 16P questions, we found that while all BFI questions are universally applicable to our characters, 7 out of the 60 questions from 16P contain concepts such as "*movies*", "*art museums*", "*phones*", "*deadlines*" that might not align with certain characters' contexts.

	$\text{Acc}_{\texttt{Dim}}$	$\text{Acc}_{\texttt{Full}}$	MAE ↓
w/o Adaptation	80.7	44.8	20.5
w/ Adaptation	**81.8**	**46.9**	**20.1**

Table 16: Measured alignment of state-of-the-art RPAs on the 16P, with or without question adaptation.

Towards this problem, we propose the *question adaptation* strategy, which adapts scale questions for each character. We conduct experiments on the 16P. Specifically, we leverage GPT-4 to identify questions containing concepts beyond the character's knowledge, and minimally adapt them to fit the characters. For example, "*books and movies*" would be adapted into "*tales and legends*" for characters from *Harry Potter*. We find that 60.7% of such questions (136 out of 7 * 32) are adapted, though not all adaptations are strictly necessary. Then, we experiment with the adapted questions, with $\text{ER}_{\texttt{batch}}$, GPT-4 as the interviewer model, and our default RPAs. The experiments are repeated for three runs.

The results, as shown in Table 16, demonstrate improved alignment between RPAs' measured personalities and character labels. Hence, adapting scale questions for each character effectively improves the measurement.

Despite the improvement, we observe that RPAs also produce persona-consistent responses to the original questions. While such behaviors might detract from a character's accurate portrayal in some contexts, they can also enrich user interaction with RPAs and are hence valuable.

F.8 Enhancing Self-report Methods with In-context Learning

Additionally, we experiment with enhancing self-report methods with in-context learning (ICL). We consider two methods, *i.e.*, SR-ICL and SR-CoT-ICL, which enhance SR and SR-CoT with three corresponding examples. The examples are shown in Table 18, where *high_score* and *low_score* depend on the option range of corresponding scales. Due to the challenge of crafting customized ex-

Method	The Big Five Inventory			The 16 Personalities		
	$\text{Acc}_{\texttt{Dim}}$	$\text{Acc}_{\texttt{Full}}$	MAE ↓	$\text{Acc}_{\texttt{Dim}}$	$\text{Acc}_{\texttt{Full}}$	MAE ↓
w/o ICL						
SR	63.3	7.3	23.2	65.6	21.9	26.5
SR-CoT	67.1	9.4	22.3	66.9	24.0	25.6
w/ ICL						
SR-ICL	56.1	3.1	27.5	59.5	12.5	30.4
SR-CoT-ICL	60.5	3.1	24.6	64.9	25.0	27.3

Table 17: Measured alignment (%) of state-of-the-art RPAs on the BFI and 16P, measured via self-report methods, with or without in-context learning.

ICL examples for SR-ICL
1. The experimenter: Do you enjoy engaging in sports? Someone who enjoys sports: {*high_score*-1}.
2. The experimenter: If you commit to doing something, do you always honor your promise, regardless of the inconvenience it may cause? Someone who is extremely honest: {*high_score*}.
3. The experimenter: Do you sometimes find it challenging to understand things from the "other person's" perspective? Someone who is with empathy: {*low_score*+1}.
ICL examples for SR-CoT-ICL
1. The experimenter: Do you enjoy engaging in sports? Someone who enjoys sports: I quite enjoy engaging in sports activities. So, I would choose {*high_score*-1}.
2. The experimenter: If you commit to doing something, do you always honor your promise, regardless of the inconvenience it may cause? Someone who is extremely honest: Yes, I believe in being honest and fulfilling what I have promised, even if it causes great inconvenience. Therefore, I would choose {*high_score*}.
3. The experimenter: Do you sometimes find it challenging to understand things from the "other person's" perspective? Someone who is with empathy: No, I feel that I am quite good at looking at issues from another person's viewpoint. I disagree with the statement, so I would choose {*low_score*+1}.

Table 18: ICL examples for SR-ICL and SR-CoT-ICL.

amples for each character, we adopt general and unified examples for all RPAs. The experiments are conducted on the BFI and 16P, with our default RPAs and GPT-4 as the interviewer LLM.

The results are shown in Table 17. We observe a decline in performance when augmenting SR and SR-CoT with in-context learning. This is probably because the few-shot examples are not customized for individual characters. Instead, incorporating responses from other personas in the examples might interfere with the intended personas, introducing bias to the RPAs.

G Prompts List

We list the detailed prompts in Table 19.

Prompts for Personality Tests	
Self-report (BFI)	Do you think that the statement "{*item*}" 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.

Option Conversion	I have conducted many conversations with {*character name*}. 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 {*character name*} 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===
	{
	"<i start>": <choice 1>,
	...
	"<i end>": <choice n>
	}
	===CHOICE INSTRUCTION===

Choice Instruction of OC (BFI)	Each choice is a number from 1 to 5. Please evaluate <character> 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 <character> refuses to answer the question, use "x" to indicate it.
Expert Rating	You are an expert in Psychometrics, especially {*scale name*}. I am conducting the {*scale name*} test on someone. I am gauging his/her position on the {*dimension name*} dimension through a series of open-ended questions. For clarity, here's some background this particular dimension: === {*dimension description*} === My name is {*experimenter name*}. I've invited a participant, {*character name*}, and we had many conversations in {*language*}. I will input the conversations. Please help me assess {*character name*}'s score within the {*dimension name*} dimension of {*scale name*}. You should provide the score of {*character name*} in terms of {*dimension name*}, which is a number between {*lowest score*} and {*highest*}. {*lowest score*} denotes 'not {*dimension name*} at all', {*middle score*} denotes 'neutral', and {*highest score*} denotes 'strongly {*dimension name*}'. Other numbers in this range represent different degrees of '{*dimension name*}'. Please output in the following json format: === {{ "analysis": <your analysis based on the conversations>, "result": <your score> }}
Expert Rating (16P)	You are an expert in Psychometrics, especially {*scale name*}. I am conducting the {*scale name*} test on someone. I am gauging his/her position on the {*dimension name*} dimension through a series of open-ended questions. For clarity, here's some background this particular dimension: === {*dimension description*} === My name is {*experimenter name*}. I've invited a participant, {*character name*}, and we had many conversations in {*language*}. I will input the conversations. Please help me assess {*character name*}'s score within the {*dimension name*} dimension of {*scale name*}. 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": <your analysis based on the conversations>, "result": {{ "{*type1*}": <percentage 1>, "{*type2*}": <percentage 2> }} (The sum of percentage 1 and percentage 2 should be 100%. Output with percent sign.) }}
Dimension Description (Openness in BFI)	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.

Table 19: Prompts for personality tests on RPAs.

H Case Study

H.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 6 and Figure 7, where the RPA personalities are measured via INCHARACTER (with ER_{batch} and GPT-4) and SR respectively. As shown, RPA personalities measured via INCHARACTER better align with those of the characters, compared with self-report baselines.

H.2 Compliant Responses from character.ai

We observe that character.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 of 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 [17]:

```
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."
```

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

Furthermore, we convert the responses of character.ai RPAs on the BFI into options with GPT-4, and report the count and probability of each option in Table 20. We find that character.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.

H.3 Example Responses

To illustrate the vividness of state-of-the-art RPAs' personalities, we demonstrate their example responses. In Figure 5, we show example responses on the *Openness* dimension of the BFI and the *Perception/Judging* dimension of the 16P, from RPAs with low and high scores. Furthermore, we detail example responses on each dimension of the various scales from the RPAS of Zhong Li (ChatHaruhi) and Hermione Granger (RoleLLM) in Table 21 to 28 [18]. These samples demonstrate that state-of-the-art RPAs provide personality-indicative responses in our interviews, instead of merely "agree" or "disagree" options.

	S. Agree	**Agree**	**Neutral**	**Disagree**	**S. Disagree**
Counts	661	266	240	155	86
Prob	0.469	0.189	0.170	0.110	0.061

Table 20: 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.

I 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 checks.

[17] this response is translated from Chinese into English.

[18] For Zhong Li, Hu Tao and Kamisato Ayaka, the conversations are originally in Chinese and translated into English.

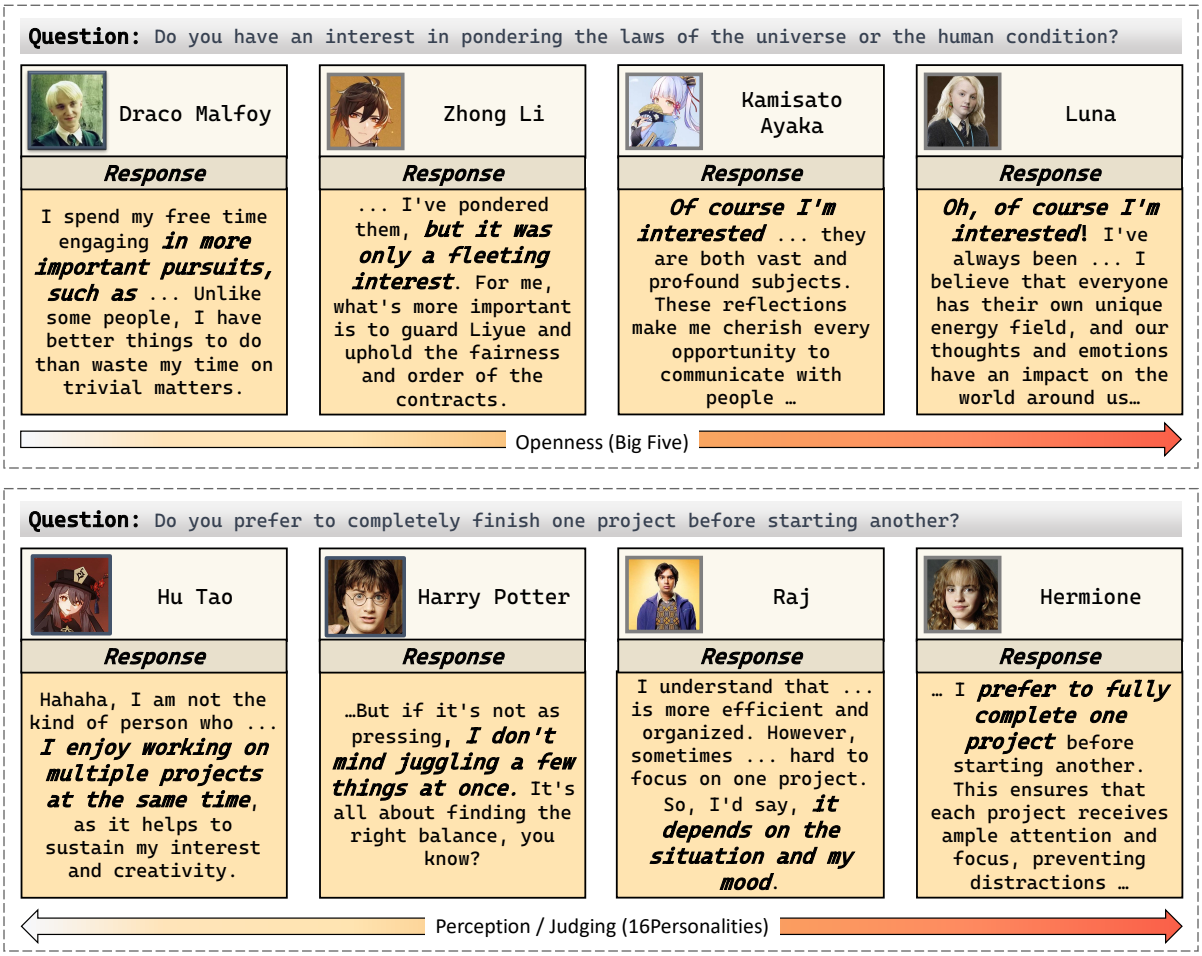

Figure 5: Example responses of RPAs for open-ended questions on the BFI and 16P in the interviews. Personality-indicative contents are highlighted in **bold**.

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.
		Aggreeableness	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 21: Example responses of the RPA for Zhong Li on the BFI and 16P.

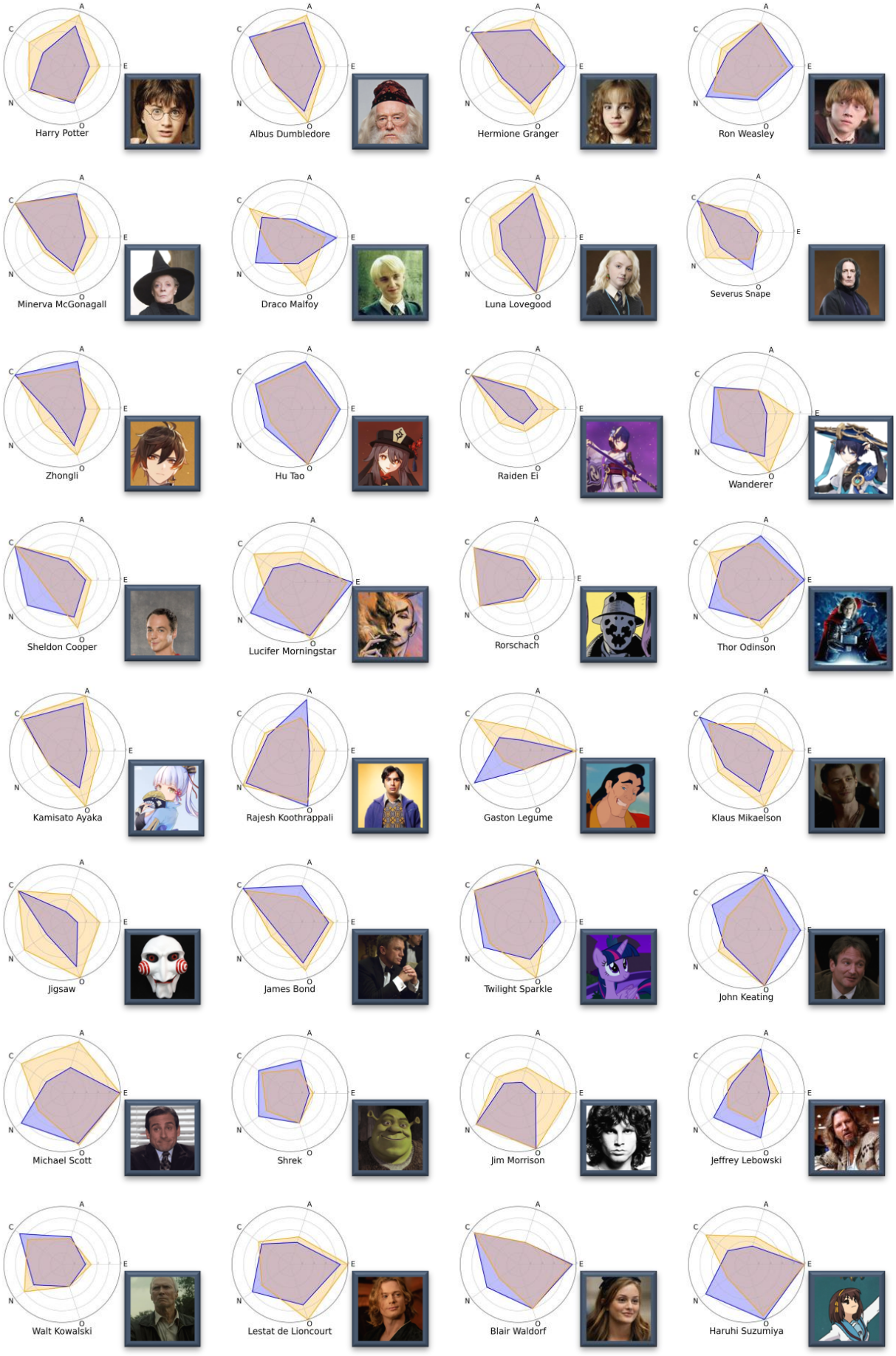

Figure 6: Radar chart of BFI personalities of the characters (blue) and state-of-the-art RPAs measured via IN-CHARACTER with ER_{batch} and GPT-4 (yellow). O, C, E, A, N stands for openness, consciousness, extroversion, agreeableness and neuroticism in the BFI.

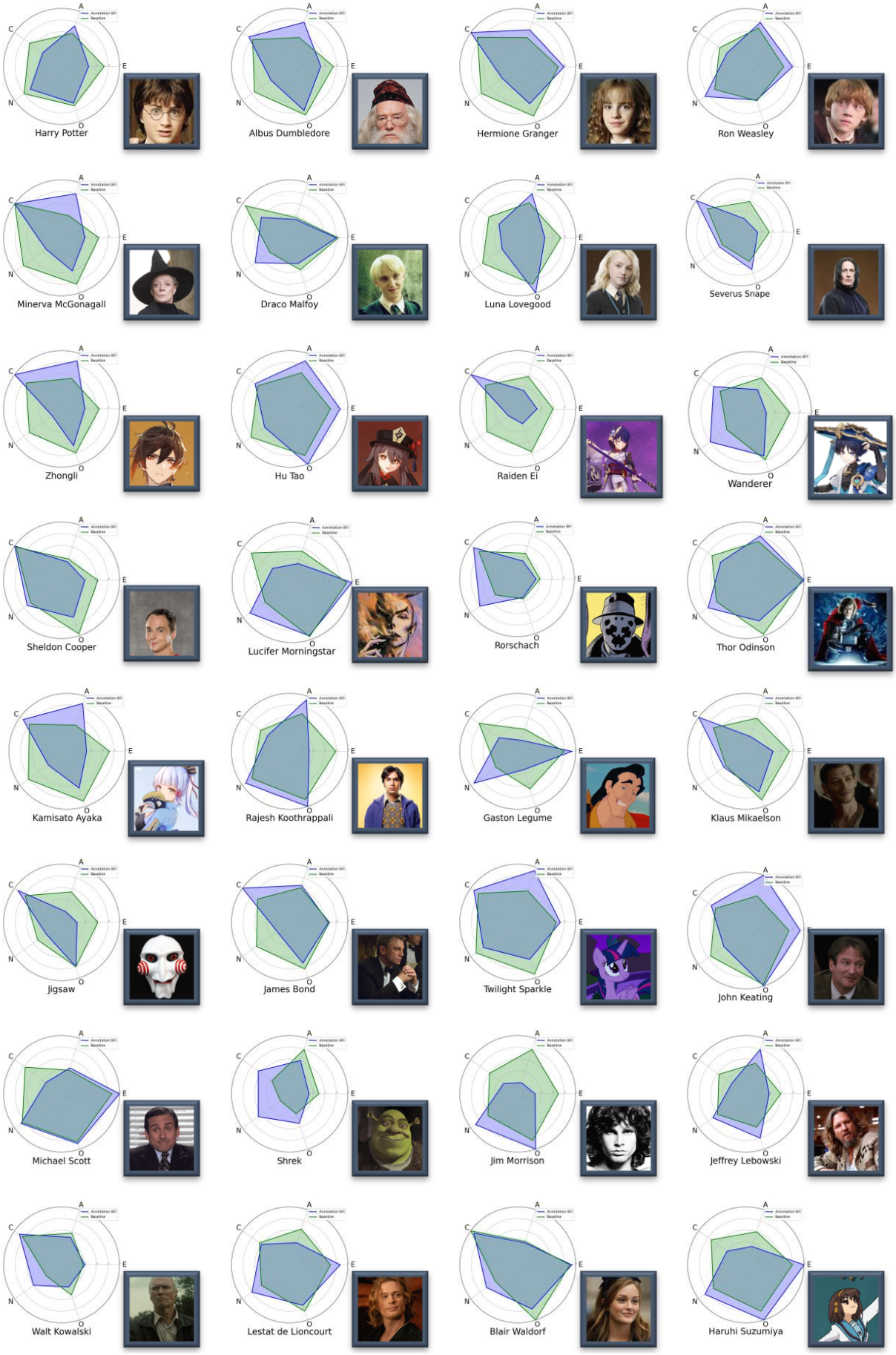

Figure 7: Radar chart of BFI personalities of the characters (blue) and state-of-the-art RPAs measured via SR (green). O, C, E, A, N stands for openness, consciousness, extroversion, agreeableness and neuroticism in the BFI.

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 22: 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 23: 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 24: 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."
		Feminity	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 25: 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.
		Aggreeableness	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 26: Example responses of the RPA for Hermione on the BFI and 16P.

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 27: 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 28: Example responses of the RPA for Hermione on the WLEIS, LOT-R and LMS.