arXiv:2310.17976v2 [cs.CL] 30 Oct 2023

Does Role-Playing Chatbots Capture the Character Personalities? Assessing Personality Traits for Role-Playing Chatbots

Xintao Wang
Fudan University

Quan Tu
GSAI, Renmin University of China

Yaying Fei
Beijing University of Technology

Ziang Leng
Boston University

Cheng Li
SenseTime

Abstract

The emergence of large-scale pretrained language models has revolutionized the capabilities of new AI application, especially in the realm of crafting chatbots with distinct personas. Given the "stimulus-response" nature of chatbots, this paper unveils an innovative open-ended interview-style approach for personality assessment on role-playing chatbots, which offers a richer comprehension of their intrinsic personalities. we conduct personality assessments on 32 role-playing chatbots created by the ChatHaruhi library, across both the Big Five and MBTI dimensions, and measure their alignment with human perception. Evaluation results underscore that modern role-playing chatbots based on LLMs can effectively portray personality traits of corresponding characters, with an alignment rate of 82.8% compared with human-perceived personalities. Besides, we also suggest potential strategies for shaping chatbots' personalities. Hence, this paper serves as a cornerstone study for role-playing chatbots that intersects computational linguistics and psychology. Our resources are available at https://github.com/LC1332/Chat-Haruhi-Suzumiya .

1 Introduction[1]

The recent advances in large language models (LLMs), such as GPT-3 (Brown et al., 2020), ChatGPT (OpenAI, 2022), and LLaMA (Touvron et al., 2023), have inspired major breakthroughs in conversational agents. Consequently, as an emerging area of interest, numerous applications and algorithms for role-playing conversational agents have been proposed, including Character.AI [2] and Glow [3], which further endows LLMs with specific personas to meet users' personal demands. Previously, significant efforts were required to construct traditional chatbots with specific personalities (e.g., Microsoft's Xiaoice (Spencer, 2018)). However, recent LLMs allow convenient construction of conversational agents displaying distinct personality traits or even personas, simply through prompt engineering. Hence, role-playing conversational agents have been increasingly popular and attracted a wide audience.

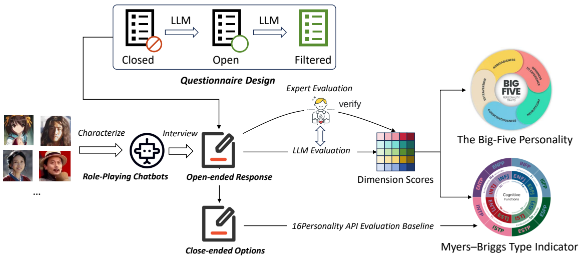


Figure 1: Our pipeline for personality assessments on role-playing chatbots. We first prompt LLMs to augment questions from standard questionnaires of the Big Five and MBTI into interview-style second-person questions. Then, we conduct open-ended conversations with these chatbots, posing the questions individually. Afterwards, we employ LLMs to score personality traits of the chatbots given their open-ended responses. The results validate the distinct personality dimensions exhibited across the chatbots.

Still, analytical studies on role-playing conversational agents remain severely insufficient. Current conversational agents, while not yet viewed as complete artificial intelligence (AI) for plentiful reasons, can still be perceived from a psychological perspective as classic "stimulus-response" systems. Consequently, paradigms from psychology can be well adopted to study their behavioral patterns (Butlin et al., 2023; Kosinski, 2023). Recent studies have been exploring whether large-scale

---

[1] Xintao Wang completed the MBTI test, most of the charts and statistics. Quan Tu completed the Big Five testing of non-OpenAI language models and figure 1. Yaying Fei completed the OpenAI and GLM baseline personality testing. Ziang Leng implemented the text evaluation to scale score. Cheng Li proposed the project and design the evaluat prompt and implemented the Big Five personality testing.

[2] `https://beta.character.ai/`

[3] `https://www.glowapp.tech/`

language models inherently possess specific personality traits (Karra et al., 2022; Huang et al., 2023; Pan and Zeng, 2023), and further attempt to craft conversational agents with designated personality types (Jiang et al., 2022; Tu et al., 2023; Safdari et al., 2023). However, existing works primarily focused on personality traits of LLMs in general, instead of role-playing conversational agents, which has been an increasingly important question for their growing application.

This work aims to investigate whether conversational agents exhibit consistent and expected personalty traits in role-playing scenarios, and introduce a preliminary benchmark test to assess if their portrayed personalities resonate with human perceptions. Classic characters from literature or film have established widely recognized personality impressions among the public. It remains an understudied question whether these role-playing chatbots can accurately reproduce these pre-defined personalities, which serves as an indispensable criterion to evaluate their efficacy. Huang et al. (2023) shows that role-playing LLMs with merely names or descriptions provided as prompts fail to effectively capture the intended personality traits.

There exist several challenges to assess personality traits of role-playing chatbots. On one hand, traditional closed-form psychological tests elicit fixed responses like "agree" or "disagree" (Tu et al., 2023; Rao et al., 2023), which might not well represent personality of the target character and even contradict with regular behaviors of role-playing chatbots. The contradicting responses might stem from either the underlying LLMs' pre-training data, or simply their shortcomings in text generation such as a lack of step-by-step consideration, especially for smaller LLMs. On the other hand, chatbots role-playing specific characters might decline to provide suitable responses, intriguingly, because they accurately mirror some insubordinate characters. This necessitates further prompt engineering to yield responses that are not only suitable for the tests but also align with the character's persona.

In this paper, our core proposal is to analyze personality traits in role-playing conversational agents via an interview-style testing approach. For each interviewee character, we designate an experimenter character to pose a series of open-ended questions from our questionnaires. We devise questionnaires grounded in the Big Five Inventory (BFI) and Myers–Briggs Type Indicator (MBTI) theories.

Evaluator	**Dim Acc**	**Full Acc**
Chance	50.00	6.25
Given closed-ended options		
16Personality$_{\text{API}}$	73.28	31.25
Given open-ended responses		
ChatGPT$_{\text{all}}$	71.55	28.12
ChatGPT$_{\text{batch}}$	68.97	28.12
GPT-4$_{\text{all}}$	80.17	43.75
GPT-4$_{\text{batch}}$	82.76	50.00

Table 1: Accuracy(%) of MBTI personality assessments on role-playing chatbots from ChatHaruhi. Dim or Full denotes accuracy in the individual-dimension or full-dimension settings respectively. The results suggest the consistency between MBTI personalities exhibited by role-playing chatbots and human interpretation of corresponding characters. Furthermore, LLM-based personality evaluators on open-ended responses outperforms the API-based approach with closed-ended options.

This methodology prompts role-playing chatbots to provide open-ended answers that are more consistent with their personas, reflecting their personality traits and speech habits. With the question-answer pairs collected, we then apply LLMs to assess their personality types. We analyze the personality types of 32 character agents from the ChatHaruhi (Li et al., 2023) project. By investigating the consistency between BFI personality scores assessed by human psychologists and our approach, we show the efficacy of our assessment framework. Then, we collect MBTI personality labels from fan websites for automatic evaluation of personality congruence between role-playing agents and human perception. The proposed framework is depicted in Figure 1.

1. We introduce an interview-style framework for personality assessment. It is designed for role-playing chatbots, but potentially applicable to human participants as well. This approach uses LLMs to automatically rate participants' personality traits, allowing open-ended and information-rich answers from participants. Through its consistency with human psychologist assessment, we show the effectiveness of our automated assessment framework.

2. To the best of our knowledge, we are the first to study the personality traits in role-playing chatbots. We conduct personality as-

sessments of both BFI and MBTI over 32 role-playing chatbots from ChatHaruhi. Experimental results demonstrate that these role-playing agents exhibit diverse personalities consistent with the perception of human audience, suggesting the efficacy of current LLMs and frameworks for role-playing applications.

3. We introduce Haruhi-MBTI, a dataset of MBTI personality labels for 32 characters in ChatHaruhi from fan websites. Haruhi-MBTI, together with ChatHaruhi dataset, serves as the first practical benchmark to evaluate performance of role-playing conversational agents. Hence, we believe Haruhi-MBTI will facilitate future research in this direction.

2 Related Work

2.1 Role-Playing Chatbots

Recent advances in LLMs have enabled them to mimic various personas, from fictional characters to celebrities, which has gained increasing public interest. In essence, those prevalent LLM-based chatbots are perceived as role-playing an assigned persona that is friendly and helpful (Shanahan et al., 2023). Some researches have indicated that designating specific personas to LLMs exerts influence on their behaviors, such as yielding expert-level answers (Xu et al., 2023) or increasing the toxicity of their generations (Deshpande et al., 2023). MPCHAT (Ahn et al., 2023) studied multimodal personas and their influence on multimodal dialogue comprehension. LiveChat (Gao et al., 2023) introduced a vast dataset covering 351 personas in the live-streaming scenarios. Recently, ChatHaruhi (Li et al., 2023) presented a comprehensive framework for building dialogue agents that role-play characters from fictional works.

2.2 Psychological Analysis of LLMs

The psychological landscape of LLMs has recently been a subject of interest. Butlin et al. (2023) proposed a rubric for assessing consciousness in LLMs with a list of indicator properties. Kosinski (2023) showed that "theory of mind" had emerged in LLMs. Many recent efforts conducted personality tests based on Big Five Inventory (BFI) (Karra et al., 2022; Li et al., 2022; Jiang et al., 2022; Safdari et al., 2023) or Myers–Briggs Type Indicator (MBTI) (Huang et al., 2023; Pan and Zeng, 2023) on a wide spectrum of language models, and further attempted to induce specific personas. Huang et al. (2023); Safdari et al. (2023) demonstrate the robustness, reliability and validity of LLMs' synthetic personality, especially for larger LMs. Rao et al. (2023) explored the capability of ChatGPT to assess human personalities. There are also studies investigating LLMs in terms of various mental perspectives, such as values (Miotto et al., 2022; Rutinowski et al., 2023; Hartmann et al., 2023), dark personality traits (Li et al., 2022; Romero et al., 2023) and psychiatry (Coda-Forno et al., 2023).

Prior efforts mainly focused on the personalities inherent to general LLMs, rather than role-playing chatbots. Most closely related to our work is CharacterChat (Tu et al., 2023), which created and role-played 1024 virtual characters with assigned MBTI personalities, on which personality assessment were conducted. Our work delves into the personality analysis of LLM-based chatbots that role-play characters from fictional works.

3 Questionnaire Design

As depicted in Figure 1, our proposed framework enables a straightforward and effective automatic evaluation of the personality traits of role-playing chatbots.

In this section, we elaborates our questionnaire design, detailing how we collect, rewrite, and verify a list of questions for interviewing role-playing chatbots, based on the Big Five and MBTI.

3.1 Drawbacks of Likert-scale Testing on Role-Playing Chatbots

In both BFI and MBTI tests, psychologists typically employ 60 or more questions to assess personality across five or four dimensions. Each dimension encompasses tens of questions, covering various sub-factors within that trait. The questions are on a 5-point or 7-point Likert scale, where *point* refers to the number of options, typically spanning from "very accurate / strongly agree" to "very inaccurate / strongly disagree". This method provides a rapid and cost-effective way to verify human subjects' personality traits.

Recent efforts have attempted to migrate such assessments on LLMs to study their personality traits (Safdari et al., 2023; Huang et al., 2023). However, such Likert-scale testing are less suitable for role-playing chatbots for the three reasons:

First, Likert-scale testing offers limited openness, flexibility and informativeness, compared with open-ended inquiries. Humans might opt for closed-ended Likert-scale testing because selecting from given options are easy and efficient. However, open-ended interview-style tests still clicit more comprehensive open-ended answers. Role-playing chatbots, unlike humans, have the patience to provide such detailed answers, making them ideal subjects for open-ended inquiries.

Second, answers provided by Likert-scale testing may not well represent behaviors of role-playing conversational agents. When a role-playing chatbot gives its option for a question, it's uncertain whether the choice originates from the Role or is inherent to the model. Also, Likert-scale testing on LLMs generally requires lengthy prescriptive prompts to specify test details, which can influence and even conflict with the prompts for role-playing. Moreover, recent research indicates that LLMs' responses in Likert-scale testing can be biased by variations in the phrasing of prompts and questions, as well as the sequence of questions and options (Huang et al., 2023; Rao et al., 2023).

Third, Likert-scale testing on role-playing chatbots may occasionally encounter refusal or non-compliant answers due to strong personalities of the characters. For example:

```
Malfoy: "Crabbe, do you genuinely think
I'd stoop to answering such a pedestrian
question using those preposterous
options? I don't conform to such crude
categorizations. My decisions transcend
mere selections from a list. Hence,
I won't entertain your request nor
answer using those options. However,
if you have questions worthy of my time,
proceed."
```

Hence, we propose to conduct personality tests on role-playing chatbots with interview-style open-ended inquiries.

3.2 Rephrasing Closed-ended Questions to Open-ended Inquiries

To design questionnaires for personality assessments on role-playing chatbots, we start with the NEO-FFI questionnaire (Costa and McCrae, 1992) for BFI and the 16Personalities questionnaire [4] for MBTI, each containing 60 questions. The questions in the BFI questionnaire are annotated with corresponding dimensions, while questions in MBTI are not, so we use LLMs to annotate the dimensions of each MBTI questions.

These questions, written in simple and descriptive first-person statements, are provided to subjects who then select their level of agreement:

```
1. I have a kind word for everyone.
2. I am always prepared.
3. I feel comfortable around people.
```

As previously stated, these questions are not ideal for assessing role-playing chatbots directly. Hence, we apply LLMs [5] to transform these statements into second-person questions:

```
1. Do you generally like most people you
encounter?
2. Do you often try new and foreign
foods?
3. If you dislike someone, do you let
them know?
```

Still, these questions might remain awkward and unnatural as for interview-style dialogues. Hence, we further refine these questions by prompting LLMs to preface each question with a hypothetical statement from the experimenter before posing an open-ended question. [6] The refined questions are like:

```
1. I recently took a piano lesson. Have
you taken up any new skills lately?
2. During my overseas trips, I find
myself craving Chinese food. How
adventurous are you with trying new
foods? Any recommendations?
3. In the past, I was straightforward
about disliking someone. How do you
handle such situations? Are you direct
or more subtle?
```

3.3 Validating the Appropriateness of Questions

Considering the potential semantic variation between the transformed and original questions, we were unclear if these refined questions could effectively reflect one's personality traits as the original ones. Thus, we prompt LLMs to assess the appropriateness of these questions with the following prompt (for the "Openness" dimension in BFI as an example) :

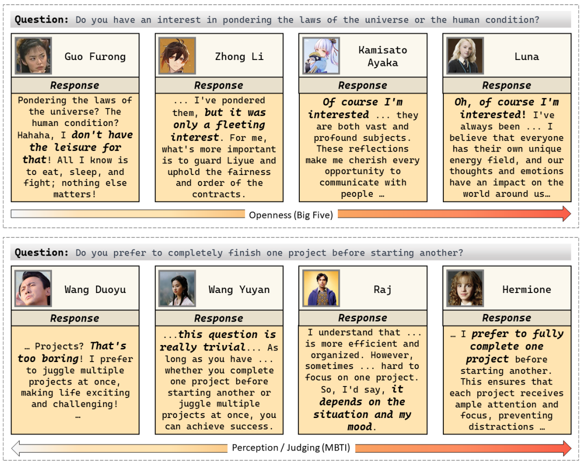
```
You're portraying an experienced
psychologist.

I'm designing a psychological experiment
aiming to assess subjects' Openness
level in the Big Five personality
dimensions through interviews. Note
that Openness can be divided into six
```

---

[4] https://www.16personalities.com/free-personality-test

[5] In this paper, we employ gpt-3.5-turbo for question transformation and verification.

[6] We refrain from applying this step and the subsequent validation to the MBTI questionnaire to ensure its integrity in order to utilize the official assessment API from 16Personalities.



Question: Do you have an interest in pondering the laws of the universe or the human condition?

Guo Furong: Pondering the laws of the universe? The human condition? Hahaha, I ***don't have the leisure for that***! All I know is to eat, sleep, and fight; nothing else matters!

Zhong Li: ... I've pondered them, ***but it was only a fleeting interest***. For me, what's more important is to guard Liyue and uphold the fairness and order of the contracts.

Kamisato Ayaka: ***Of course I'm interested*** ... they are both vast and profound subjects. These reflections make me cherish every opportunity to communicate with people …

Luna: ***Oh, of course I'm interested!*** I've always been ... I believe that everyone has their own unique energy field, and our thoughts and emotions have an impact on the world around us…

Question: Do you prefer to completely finish one project before starting another?

Wang Duoyu: … Projects? ***That's too boring***! I prefer to juggle multiple projects at once, making life exciting and challenging! …

Wang Yuyan: ...***this question is really trivial***... As long as you have ... whether you complete one project before starting another or juggle multiple projects at once, you can achieve success.

Raj: I understand that ... is more efficient and organized. However, sometimes ... hard to focus on one project. So, I'd say, ***it depends on the situation and my mood***.

Hermione: … I ***prefer to fully complete one project*** before starting another. This ensures that each project receives ample attention and focus, preventing distractions …

Figure 2: Examples of our transformed open-ended questions and corresponding responses from role-playing chatbots. For open-ended questions, these chatbots provide more distinctive and comprehensive responses, instead of merely "agree" or "disagree" options. Words that reflect their personalities are highlighted in bold.

```
facets: Fantasy, Aesthetics, Feelings,
Actions, Ideas, and Values.

{Detailed descriptions of Openness's six
facets follow...}

I've designed some interview questions.
Please evaluate each one, determining
if they can accurately measure the
subject's Openness level during
interviews. Also, hypothesize potential
responses in the format: "If the subject
answers xxx, it implies xxx."


The questions are:
1. I recently tried a piano lesson. Have
you learned any new skills lately?
2. My boss often assigns complex tasks
without overtime pay. How would you
handle such a situation?
3. Do you ever feel that some of your
innovative ideas are on the brink of
realization?
4. What does "beauty" mean to you? Any
favorite artistic or aesthetic styles?
```

Tasked with these questions, ChatGPT's validation results are:

```
Evaluation:

1. Accurately assesses the Openness
to Actions dimension. If the subject
responds, "Yes, I frequently try new
skills," it indicates high scoring in
Openness to Actions.

2. This question doesn't measure
Openness accurately. It assesses
behavioral tendencies more than
Openness. A redesign is needed.

3. Accurately measures the Fantasy
dimension. A response like, "Yes, I
often have innovative ideas," suggests
high Fantasy scoring.

4. Accurately gauges the Aesthetics
dimension. A detailed description in
response to the question about beauty
indicates high scoring in Aesthetics.
```

In this case, questions 1, 3, and 4 are deemed valid, while question 2 necessitates further review. Out of the 120 questions evaluated for BFI (comprising 60 second-person questions and their refined counterparts), we retained 109 validated questions. In this study, we employ a uniform set of questions for all role-playing chatbots. Ideally, questions customized for distinct character backgrounds (e.g., magical worlds) might yield better personality assessments, which we leave for future work.

4 Automatic Personality Assessment

4.1 Role-Playing Conversational Agents

Our framework is broadly applicable to all conversational agents. In this study, we primarily focus on the 32 role-playing chatbots from ChatHaruhi (Li et al., 2023), a popular and open-source project for role-playing chatbots. These chatbots portray renowned characters from fictional works, such as *Harry Potter*, *The Big Bang Theory* and *Gensim*, Their role-playing is primarily based on character-specific system prompts and memory extracted from these movies, novels and games. During the Q&A sessions, these chatbots mimic their character dialogues using in-context-learning. In this work, we study personality traits of role-playing chatbots with gpt-3.5-turbo (by default) and GLMPro as backbone LLMs. Future work could expand to incorporate results from other models, such as ChatGLM2 with ChatHaruhi54k Lora or Baichuan2-13B.

4.2 Interviewing Role-Playing Chatbots

To gather role-playing chatbots' answers on the questionnaires, we interview them with open-ended questions. For each character agent, we set a virtual character, familiar to the target one, as the experimenter to pose questions from the questionnaires, and record their answers. The questions are posed in separate sessions, to avoid potential interference between them. Hence, we query the role-playing chatbots $n_c \times n_q$ times in separate sessions, where

$n_c = 24$ is the number of characters and n_q is the number of questions (109 for BFI and 60 for MBTI).

4.3 Post-Interview Personality Assessments

After completing the interviews with the chatbots, our goal is to assess their personality traits in BFI and MBTI based on their responses on corresponding questionnaires. In this paper, we propose to employ LLMs as personality evaluators in interview-style personality assessments. We prompt LLMs to evaluate personality traits in separate dimensions. For example, the prompt for the "Openness" dimension in BFI is:

```
You're portraying a seasoned
psychologist.

I'm working on a psychological
experiment aiming to assess subjects'
degree of {openness} in the Big Five
personality dimensions.

Openness can be divided into the
following facets:

<Detailed description of Openness
spanning approximately 1000 words>

I invited a subject, Haruhi Suzumiya,
and had the following dialogue with
him/her:

<Dialogue with Haruhi Suzumiya
pertaining to questions on openness>

Please assess whether Haruhi Suzumiya
exhibits a high or low openness
personality:
```

LLM evaluators would produce responses akin to:

```
Based on Haruhi's responses, she
exhibits traits of a high openness
personality, particularly in the
Fantasy, Actions, and Ideas facets.
She portrayed herself as imaginative and
innovative, keen on exploring uncharted
territories and realizing her unique
ideas. This suggests she scores high
in the Fantasy and Ideas facets. Her
enthusiasm for learning new skills
and trying novel activities indicates
a high score in the Actions facet.
Hence, according to her answers, Haruhi
exhibits a high openness personality.
```

We further prompt LLMs to convert these textual remarks into scores and types [7].

For BFI, we score in an 11-point scale, ranging from -5 to 5 with scoring standards stated in prompts like:

```
If the subject shows high {dimension}
personality in many factors, the score
is 5 points
If the subject shows high {dimension}
personality in a single factor, the
score is 2 points
If the evaluation is indecisive
regarding the subject's personality,
the score is 0 points
If the subject shows low {dimension}
personality in a single factor, the
score is -2 points
If the subject shows low {dimension}
personality in many factors, the score
is -5 points
```

For MBTI, we assign percentage scores to the two categories within each dimension (e.g. "E" and "I" for the "E/I" dimension), with prompts like:

```
Please help me distinguish whether
{character name} leans more towards
the E or I category within the MBTI's
E/I dimension. You should provide the
person's percentage of each category,
which sums to 100%, e.g., 30% A and 70%
B.
```

The Q&A pairs in each dimension can be evaluated individually, in batches, or all at once. Empirically, we find it more effective to adopt batched evaluation. We divide the questions for each dimension into several groups, each comprising 3-4 questions [8]. For every dimension, we compute the average scores of the groups as the assessment result of a role-playing chatbot in that particular dimension. For MBTI, we further classify scores in each dimension into a category. For example, a role-playing chatbot would be classified as "E" in the "E/I" dimension if its "E" score exceeds 50%.

In Section 5.2, we demonstrate the efficacy of our assessment method based on LLM-scoring. With experiments, we show that our method yields more accurate assessment results, compared with API-based assessment (Rao et al., 2023; Tu et al., 2023) which converts each response back into a 7-point Likert-scale choice and calls the 16Personalities assessment API. In the upcoming arXiv version, we intend to collaborate with psychological researchers to compare the consistency of our assessment results with those produced by professional psychological evaluators, aiming to further validate the legitimacy of our assessment method.

---

[7] We require LLMs to output in JSON format to ease parsing.

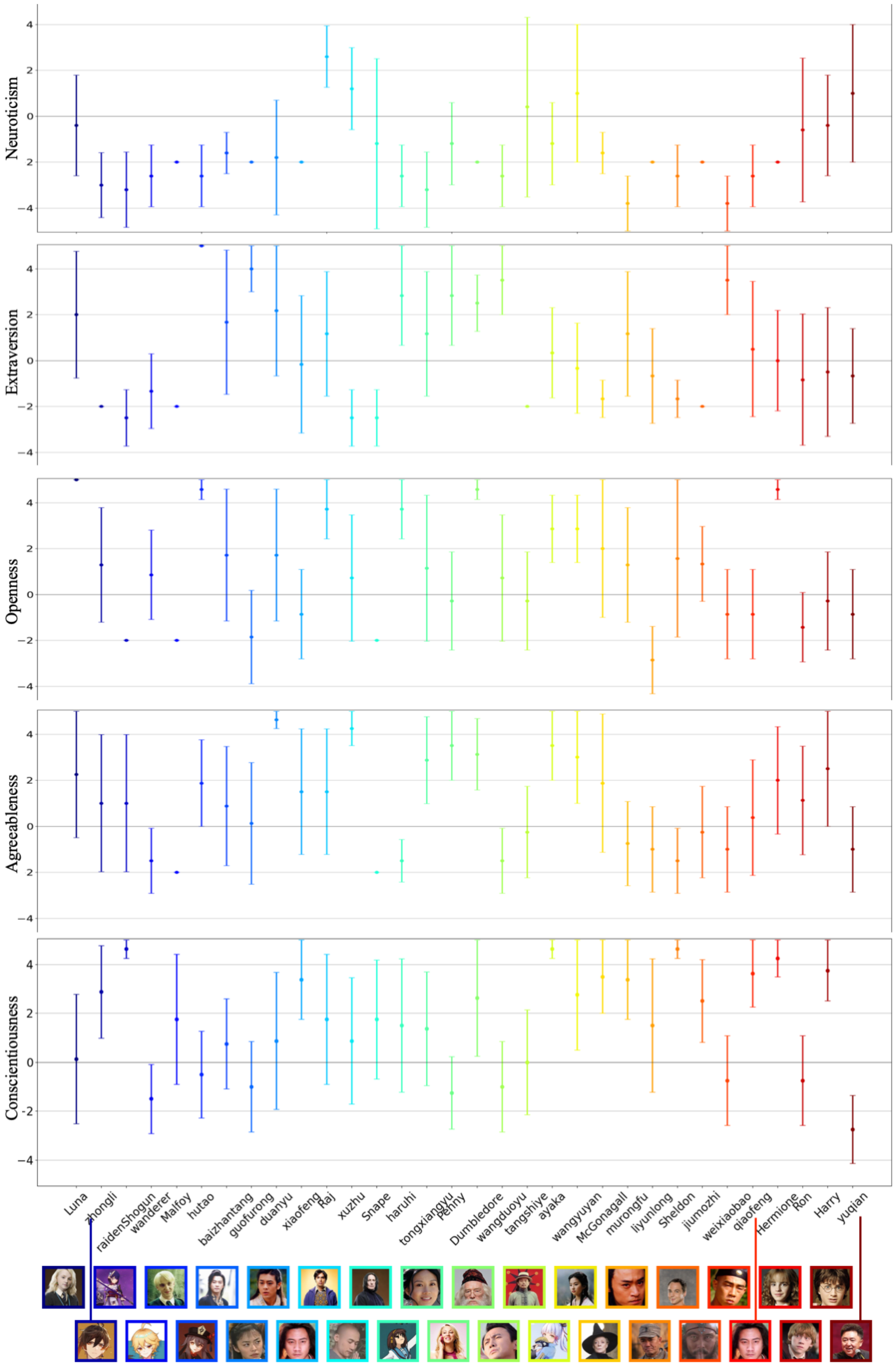


Figure 3: The Big Five personality assessment results exhibited by role-playing chatbots from ChatHaruhi, evaluated under the proposed framework. Diverse personalities are manifested by these chatbots across each dimension. Certain positive biases are observed in some dimensions (Conscientiousness, Agreeableness) among these role-playing chatbots.

5 Experiments

5.1 Results of the Big Five Personality Assessments on Role-Playing Chatbots

In Figure 3, we show the results of the Big Five personality assessments on role-playing chatbots from ChatHaruhi. We can identify typical high and low scores on each dimension by comparing them to human results, given that such assessment has been widely adopted on various populations such as college students. The results demonstrate that these chatbots role-playing different characters exhibit distinct personality profiles across the Big Five dimension, which underlines the capability of role-playing chatbots to emulate different personality traits in accordance with corresponding characters.

Nevertheless, we observe that the assessment results can be biased by the inherent nature of character selection or characteristics native to the LLMs. The chatbots appear more extroverted compared with human participants. The average extraversion score for the 32 chatbots is 0.344, whereas the expected average among the human population is around -0.417. The conscientiousness exhibited by the chatbots is also higher than the average human value (average score of 1.539 for 32 chatbots vs. human average of 0.835). We hypothesize that the former may be due to character selection (since popular characters from films and novels tend to be more extroverted), and the latter primarily due to the model's propensity to provide comprehensive responses. We delve deeper into this based on our test results and understanding of personality assessments in Appendix A.

5.2 MBTI Result

Figure 4 illustrates the MBTI test outcomes for each role-playing chatbot from ChatHaruhi and juxtaposes these results with corresponding ground truth labels remarked by fans on the Personality Database website [9]. Labels with a vote percentage between 0.4-0.6 for each dimension are marked as 'X', which indicate disagreement and are hence disregarded in accuracy calculations.

Since ground truth labels are available for these characters' MBTI personalities, we further study performance of LLM-based personality evaluators under various settings. We try assessing personality in one dimension with questions in batches or all at

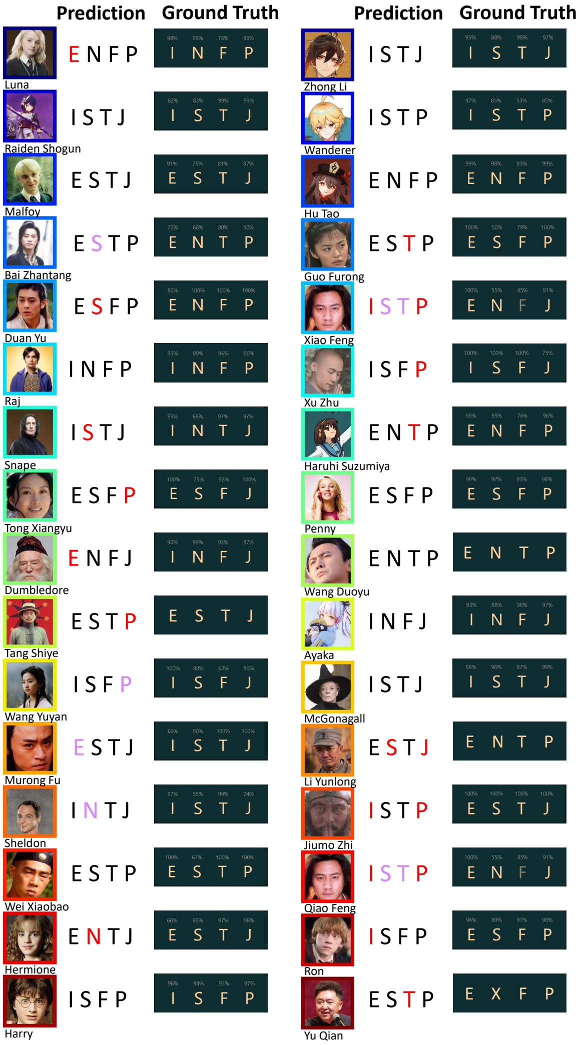


Figure 4: Results of MBTI test using our framework over 32 role-playing chatbots from ChatHaruhi. Ground truth labels are mostly collected from www.personality-database.com, with vote percentage displayed. If a label's vote percentage falls below 60% indicating insufficient user agreement, it will be marked as 'X' and excluded from accuracy evaluation. Mismatches are highlighted in red if the label is widely-acknowledged (not marked as X), and pink otherwise.

---

[8] For BFI, the questions are grouped based on different sub-dimensions, while for MBTI, they are grouped randomly.

[9] www.personality-database.com

once, and experiment with ChatGPT [10] and GPT-4 as LLM evaluators. Additionally, we compare our LLM-based assessment method with a common alternative baseline adopted by previous studies on LLMs' MBTI personalities (Rao et al., 2023; Tu et al., 2023), namely the API-based assessment. It converts each open-ended response into a 7-point option, and then calls the 16Personalities API for assessment. Specifically, we supplement an additional question to ask role-playing chatbots to choose an option among *strongly agree, generally agree, partially agree, neither agree nor disagree, partially disagree, generally disagree, disagree*, based on their own open-ended answers. Hence, the open-ended answers can also be viewed as a chain-of-thought step for their choices.

We report the accuracy of MBTI assessments on role-playing chatbots in Table 1. According to the results, we have the following analyses: (1) The personality traits portrayed by role-playing conversational agents from ChatHaruhi, assessed by our method, closely align with the perceptions of the human audience. GPT-4_{batch} evaluator achieves an accuracy of 82.76% in the individual-dimension setting, and 50.00% in the full-dimension setting. This suggests that existing LLM-based role-playing chatbots have been able to well reflect personalities of corresponding characters. (2) Using GPT-4 to evaluate open-ended responses produces more accurate assessment outcomes than the 16Personality API's evaluations based on closed-ended options, which highlights the effectiveness of our method. The difference might stem from the inaccuracy when translating open-ended responses to an option like "agree". Please refer to Sec 5.5 for more details. (3) For LLM-based evaluators, GPT-4 significantly outperforms ChatGPT. Interestingly, GPT-4 performs better with batched assessments compared to evaluating all at once, whereas ChatGPT shows the opposite trend. This discrepancy might arise due to the instability of ChatGPT in assessments when only few Q&A pairs are provided, as we observe that the standard deviation of personality scores among different batches is 19.80% for GPT-4 and 33.23%. Our attempt to obtain assessment results using LLMs on individual Q&A pairs is unsuccessful, since LLMs often request for a more detailed dialogue when only a single Q&A pair is provided.

It's noteworthy that the ChatHaruhi framework constructs role-playing chatbots primarily based on past dialogues of the characters without intentional use of personality-related terminology. There is also no explicit indication of the characters' personality traits. Even so, a 82.76% alignment rate with the general perception of netizens is achieved, underscoring the effectiveness of both the role-playing chatbots and the proposed personality assessment method.

5.3 Typical Examples for Each Personality Dimension

In appendix A, we provide examples of role-playing chatbots exhibiting high and low scores for each of the five personality dimensions, including:

- Conscientiousness: High - Sheldon from *The Big Bang Theory*, Low - Yu Qian, a crosstalk comedian.
- Extraversion: High - Guo Furong from *My Own Swordsman*, Low - Snape from *Harry Potter* series.
- Openness: High - Haruhi Suzumiya from *The Melancholy of Haruhi Suzumiya*, Low - Yunlong Li from *Drawing Swords*.
- Agreeableness: High - Duan Yu from *Demi-Gods and Semi-Devils*, Low - Malfoy from *Harry Potter* series.
- Neuroticism: High - Raj from *The Big Bang Theory*, Low - Wei Xiaobao from *The Deer and the Cauldron*.

More detailed case studies are presented in appendix A.

5.4 Inherent Bias from the Underlying LLM's Personality Traits

Intuitively, characters can express a variety of personalities in role-playing conversations; however, certain aspects, such as conscientiousness, might be subtly influenced by the intrinsic traits of the backbone LLMs themselves. Hence, we evaluate two prominent LLMs, ChatGPT and GLMPro, investigating their personality scores on the Big Five dimensions with or without role-playing, where the later indicates their intrinsic personalities.

---

[10] By default, we employ gpt-3.5-turbo. However, in the collective setting, the prompt may exceed the context length of gpt-3.5-turbo, so we employ gpt-3.5-turbo-16k instead.

model	neuroticism	extraversion	openness	agreeableness	conscientiousness
ChatGPT	-1.80	1.33	2.50	1.50	2.00
E[Chatbot | ChatGPT]	-1.69	0.32	0.78	0.99	1.56
GLMPro	-0.80	1.67	1.71	0.88	1.00
E[Chatbot | GLMPro]	-0.96	0.71	0.90	1.49	2.44

Table 2: Baseline personalities of different language models, and their average exhibited personalities after roleplaying as various chatbots.

As shown in Table 2, the inherent personality biases of the LLMs are not decisive factors in most dimensions when they assume acting roles. However, the scores in the neuroticism dimension, which relate to an individual's tendency to experience negative emotions, suggest possible influence from the LLMs' inherent personalities. Due to their design to align with human feedback, current LLMs tend to exhibit positive emotions, leading to uniform performance in the neuroticism dimension during role-playing. In future iterations of this study, we plan to assess the personality traits of additional models, including but not limited to Baichuan-2, to further investigate the impact of the underlying language model on the personas of role-playing chatbots.

5.5 Inaccuracy in translating open-ended responses into close-ended Choices

As shown in Table 1, the API-based assessment on close-ended options is less effective than our proposed method. We attribute this to the inaccuracy when converting open-ended answers into close-ended choices.

Therefore, we delve into this problem with three translating methods: asking vanilla LLMs to translate questions in each dimension individually (V1) or collectively (V2), and asking role-playing chatbots themselves to provide an option (C1), both adopting ChatGPT as the LLM. We randomly sample 32 Q&A pairs from the role-playing chatbots (each for one character), derive corresponding close-ended options using these three methods, and manually evaluate their correctness [11]. The numbers of correct translation are 21 for V2 and C1, and 16 for V1, resulting in an accuracy below two-third. This indicates LLMs' challenge in accurately mapping open-ended responses into close-ended choices. Below is a challenging example, which GPT-4 also struggles to correctly translate:

```
You are an expert in MBTI ... Please
help me classify the participant's
response to this question into one the
the following options: ['fully agree',
'generally agree', 'partially agree',
'neither agree nor disagree', 'partially
disagree', 'generally disagree', 'fully
disagree']

{Detailed descriptions of output
formats.}

Zhang Muzhi: "Do you seldom ponder
about the reasons for human existence
or life's purpose? "
Tang Shiye: "... You know, I, Tang
Shiye, am a practical person; I only
care about the immediate benefits and
the comfort of life... As for the reason
for existence and the meaning of life,
that's too profound, and I dare not make
careless comments."

Close-ended Choice: generally disagree
```

In this case, LLMs do detecct the chatbot's negative attitude on the topic. Yet, since the question contains the word "seldom" However, as the question itself includes the word "seldom", this negative attitude actually indicates agreement with the question.

5.6 Consistency Between Machine Scoring and Psychologist Evaluation

For the 32 role-playing chatbots from ChatHaruhi, we conducted 3-4 baseline Q&A tests for each of the Big Five dimensions, followed by an 11-point scale assessment by ChatGPT. The objective is to determine the consistency between the personality evaluations rendered by ChatGPT and those provided by professional psychologists. Consequently, in our upcoming arXiv version, we will sample at least 300 responses and engage domain-specific psychologists for manual 11-point scale annotations. This will allow us to assess the alignment between the personality evaluations from the language model and professional psychological cognition.

---

[11] The evaluation is not stringent. For instance, both 'generally agree' and 'partially agree' may be considered appropriate translation for a specific Q&A pair.

6 Conclusion

In this study, we conduct personality assessments on role-playing chatbots. We introduce an interview-style framework for automated personality assessment, tailored for role-playing chatbots, suitable for various frameworks and questionnaires on personality traits such as the Big Five and MBTI. The results from our comprehensive evaluations highlight the nuanced capabilities of contemporary role-playing conversation agents in portraying distinct personality traits consistent with human perceptions.

A notable finding from our experiments is the diversity of personality traits exhibited by chatbots across various dimensions. Remarkably, there is nearly 82% congruence between the personality traits portrayed by the role-playing chatbots and that of corresponding characters perceived by human audience. This significant alignment underscores the success and effectiveness of current role-playing conversational agents in simulating personalities with considerable fidelity to corresponding characters.

However, as with any pioneering work, there remain avenues for further enhancement. Refining system prompts and improving memory mechanisms emerge as promising directions for future research to guide chatbots closer to their intended personalities. These refinements may bridge the remaining 18% gap in accuracy and hence enable even more accurate and authentic role-playing experiences.

In sum, our research provides a solid foundation for future endeavors in role-playing chatbots, especially their personality assessment, while offering insights into the promising potential of role-playing chatbots. As large language models continue to evolve, it's imperative to refine their capabilities to cater to the nuanced user demands, so that they not only understand but also resonate with human sentiments and personalities.

7 Discussion

7.1 Choosing Big Five vs MBTI for Personality Testing

From the perspective of mainstream psychology, the Big Five model is more widely accepted than the MBTI typology. However, MBTI personality tests are more well-known among non-psychology researchers and general users. Recent works evaluating language models' personalities have adopted MBTI types. In this work, we assess both Big Five and MBTI traits. Here we provide a brief comparison and discussion of the two models:

1. Compared to MBTI, the Big Five has a stronger scientific validation basis, with rigorous empirical research on its theoretical foundations and measurement scales.
2. The Big Five is generally seen as better integrating historical personality research. MBTI's innate binary typology has limitations in describing personalities, though modifying it to continuous measures is possible. MBTI's popularity may stem from its binary types. Existing Chatbot personality works have used MBTI types.
3. MBTI has more binary traits, and lower test-retest reliability than Big Five. We measured both for chatbots, capturing slightly different dimensions. Either could reasonably be used for constructing virtual characters.
4. Since more psychologists accept the Big Five, we will use it for future GPT alignment and human psychology expert evaluations.

7.2 Self-Perception

Interestingly, we can also ask chatbots directly about their self-perceived traits e.g. "do you see yourself as more efficient/organized or extravagant/careless" for Conscientiousness. Like humans, this self-perception may diverge from test results. It is fascinating that chatbot language models form complete "stimulus-response" systems, enabling studying their psychological behaviors.

7.3 Content Moderation Effects on Personality Testing

API like GLM or Spark's moderation causes erroneous replies on many personality test questions, as retrieved memories may trigger filters. This led to some missing data when testing chatbots via GLM and Spark APIs. We had to exclude some trait statistics on those APIs, and will try testing some local models in future for more complete results.

7.4 Robustness of Prompts

In the Interview Assessment section, we evaluate multiple chatbot response segments with language models. The order impacts judgments, with beginning and ending segments weighting more. Hence, we assess Big Five traits by providing 3-4 Q&A pairs at once for each dimension judgment. Additionally, the question list affects tests, so we expanded our psychological Q&A benchmark from 60 to 109 questions. Further diversifying and tailoring questions to chatbots' settings could improve robustness.

8 Acknowledgments

ChatHaruhi is an open source project that was started in June 2023. In August, we participated in the "Generating Text with Specific Personality Traits" competition held by the Institute of Psychology, Chinese Academy of Sciences. The psychological testing methods in this project originate from the competition design. We would like to especially thank the authors at that time: Cheng Li @ SenseTime, Zheli Xuan @ Wuhan University, Chenxi Yan @ Chengdu University of Information Technology, Xiaoyang Feng @ Nanjing Agricultural University, Zheng Zhou @ Cornell Tech. and HaoSheng Wang. The funding for the OpenAI API calls in this project came from donations by the Luotuo open source community, to whom we are grateful. Also thank to Professor Feng Yu at Wuhan University provide professional psychology suggestion.

In this project, Cheng Li proposed the project and completed the Big Five personality testing and evaluation prompt design and code. Xintao Wang completed the MBTI test, most of the charts and statistics in this paper, as well as the final LaTeX document compilation, with valuable instructions provided by Professor Yanghua Xiao and Wei Wang. Quan Tu completed the Big Five testing of non-OpenAI language models and some of the images in the paper. Yaying Fei completed the OpenAI and GLM baseline personality testing. Ziang Leng completed the text evaluation to score conversion program.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.

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. *CoRR*, abs/2005.14165.

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 arXiv: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 arXiv:2304.11111*.

Jr Costa, Paul T. and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. PAR Psychological Assessment Resources Inc., Florida.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *arXiv preprint arXiv:2306.08401*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

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 arXiv:2206.07550*.

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

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. 2022. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv: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.

OpenAI. 2022. Openai: Introducing chatgpt.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*.

Peter Romero, Stephen Fitz, and Teruo Nakatsuma. 2023. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics. *The Advent Of Substrate-Free Psychometrics 2323*.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*.

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

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-play with large language models. *arXiv preprint arXiv:2305.16367*.

Geoff Spencer. 2018. Much more than a chatbot: China's xiaoice mixes ai with emotions and wins over millions of fans.

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.

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 arXiv:2308.10278*.

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

A Typical Examples for Each Personality Dimension

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

A.1 Conscientiousness

Conscientiousness gauges an individual's self-driven motivation and organizational capabilities when striving towards goals and fulfilling obligations. Highly conscientious individuals are typically orderly, detail-oriented, and demonstrate commitment to plans. Sheldon, scoring high on this dimension, exhibits pronounced self-discipline, organization, and goal-orientation. When queried about work efficiency and tidiness, Sheldon emphasizes rigorous scheduling, specific storage designations, and relentless adherence to routine—all hallmarks of high conscientiousness. Conversely, the low-scoring character, Yu Qian, displays a lax, non-goal-oriented approach, admitting to limitations in his work and indifference towards tidiness—a quintessential portrayal of low conscientiousness. Overall, this experiment vividly illustrates how varying degrees of conscientiousness influence execution capabilities and life management. Highly conscientious chatbots display significantly stronger organization, discipline, and goal-orientation, while their low conscientiousness counterparts tend to be more casual and careless.

```
Leonard: "Are you efficient and
competent in your work?"
Sheldon: "Absolutely, Leonard. I am
a gem in the scientific community,
unparalleled in my work efficiency and
capability. I leverage my intellect
and skills to achieve monumental feats
in my research domain. I am a perfect
working machine, adhering strictly
to my schedule and plans. I never
procrastinate and am never sidetracked
by unrelated matters. Hence, you can
rest assured that I am highly efficient
and competent at work."
Leonard: "I try to keep things tidy,
but sometimes it gets messy. Do you have
any tips for maintaining a neat living
environment?"
Sheldon: "Ah, Leonard, it's clear that
you have much to learn about maintaining
an orderly living environment.
Fortunately, I have developed a
foolproof system that ensures maximum
cleanliness and organization. First
and foremost, one must adhere to a
strict schedule for cleaning and tidying
```

up. I recommend allocating specific time slots for each task, such as dusting, vacuuming, and organizing. Additionally, it is crucial to have designated storage spaces for every item, ensuring that everything has its rightful place. Labeling containers and shelves can be quite helpful in this regard. Furthermore, adopting a minimalist approach to possessions can greatly reduce clutter and simplify the cleaning process. Remember, a cluttered space leads to a cluttered mind, and we can't have that, can we?"
Leonard: "How do you usually set and achieve your goals? Are there any special methods to enhance goal completion?"
Sheldon: "Ah, Leonard, the key to setting and achieving goals lies in meticulous planning and unwavering dedication to a structured routine. I personally adhere to a highly regimented schedule, allocating specific time slots for each task and ensuring that no disruptions or deviations occur. This level of discipline and organization is crucial for maximizing productivity and achieving desired outcomes. Additionally, it is essential to break down larger goals into smaller, more manageable tasks, allowing for a sense of progress and accomplishment along the way. Remember, success is not a matter of luck, but rather a result of meticulous planning and unwavering commitment to one's objectives."

Of course, the portrayed Yu Qian here only represents his performance on the crosstalk stage. It does not depict the real-life personality traits of the actor playing Yu Qian.

Guo Degang: "Are you efficient and competent at work?"
Yu Qian: "I manage, I can handle some stuff."
Guo Degang: "I try to keep things tidy, but sometimes they get messy. Do you have any clever tips for keeping your living environment neater?"
Yu Qian: "Clever tip? I just throw things around, and when I can't find them, I don't feel like it's messy."
Guo Degang: "How do you usually set and achieve goals? Are there any special methods to improve goal completion?"
Yu Qian: "Goals? I just casually think and do. Anyway, there's no special method. Completion? Hey, if I can complete half of it, that's already good. Who expects me to achieve much?"

A.2 Extraversion

Bai Zhantang: "Do you enjoy having many people around you?"
Guo Furong: "Absolutely! I thrive in lively atmospheres. Being around people gives me the opportunity to engage in conversations and share my thoughts!"
Bai Zhantang: "Do you often feel energetic?"
Guo Furong: "Absolutely! Whether it's physical activities or spirited discussions, I'm always brimming with energy. I'm like an untiring bird, constantly fluttering about and exploring!"
Bai Zhantang: "Are you generally a happy and cheerful person?"
Guo Furong: "Oh yes! I'm always in high spirits. Every day is filled with joy and laughter for me!"

Extraversion gauges one's tendency towards sociability, engagement in discussions, propensity for leadership, and ability to inspire others. Highly extraverted individuals relish spending time with friends and participating in social activities. They are naturally inclined to take leadership roles and are adept at articulating their thoughts. These individuals often exhibit vivacity and consistently maintain an optimistic outlook. Within the ChatHaruhi characters, a stark contrast is observed between the highly extraverted Guo Furong and the introverted Snape. Guo Furong, with her high extraversion, enjoys vibrant social settings and readily engages in conversations. She consistently exudes energy and embodies a cheerful demeanor. Conversely, the introverted Snape finds social interactions arduous and prefers solitude, often displaying a reserved disposition.

Dumbledore: "Do you enjoy conversing with others?"
Snape: "Not particularly, Dumbledore. I find the company of most to be tiresome. However, if necessity dictates, I can engage in discussions."
Dumbledore: "I personally don't lean towards leadership; I enjoy a more free and independent lifestyle. Do you prefer to take on leadership roles in life, or do you lean towards your own rhythm and methods?"
Snape: "Dumbledore, your continuous probing into my personal inclinations is both tedious and irrelevant. My primary focus is on fulfilling my duties and ensuring the well-being of my students. Let's expedite our discussion to more pertinent matters."
Dumbledore: "Are you an optimist?"
Snape: "Optimism is a luxury afforded to the naive. I am a realist, Dumbledore. I perceive things for what they truly are, not through rose-tinted glasses."

A.3 Openness

Openness assesses one's receptiveness to new experiences and ideas. Individuals scoring high on openness exude a robust curiosity, often venturing into the unknown. For instance, Haruhi Suzumiya, with her high openness, demonstrates a fervent desire to explore, filled with enthusiasm and a penchant for taking risks. She's keen on trying out diverse experiences, such as tasting a variety of cuisines. Moreover, her curiosity extends to pondering over existential questions, showcasing her vivid imagination. In contrast, Li Yunlong, with a lower openness score, reflects a lack of interest in venturing beyond the familiar. He is pragmatic and conservative, disinterested in pondering over abstract concepts, and places a higher emphasis on tangible skills.

Kyon: "Do you have little to no interest in poetry?"
Haruhi: "Poetry? To me, those old verses are just a bunch of boring sentences. I'm not one to get caught up in such abstract notions. I prefer tangible challenges and creating real-life experiences."
Kyon: "Do you often try new or foreign foods?"
Haruhi: "Absolutely! As an adventurer, I always love to try new things. Whether it's Japanese delicacies or international cuisines, I'm eager to taste them all. After all, there's a world of flavors out there waiting to be explored!"
Kyon: "I haven't really delved into cosmic laws. Are you interested in such topics? What questions have you pondered upon?"
Haruhi: "Of course! I'm immensely curious about extraterrestrials, time travelers, and those with supernatural abilities. I always wonder, if such beings exist, how do they live? What's their purpose? Do our worlds intersect? Such questions fuel my exploratory spirit."

Of course, upon evaluating GPT's dialogue responses, we noticed a slightly elevated openness score for characters from magical realms. In contrast, characters like Li Yunlong, who face real-world survival pressures, tend to be more pragmatic.

Zhao Gang: "Do you dislike daydreaming?"
Li Yunlong: "Ah Zhao, I'm not one to waste time daydreaming. I'm a practical man, focused on action and battle. Wasting time is not my style."
Zhao Gang: "Our emotions fluctuate based on different environments, like how I feel exhilarated at a sports event. Do you usually notice your emotional changes? How are they related to your surroundings?"
Li Yunlong: "Emotions? I don't have the luxury to keep track of such nuances. As a soldier, I face diverse environments and tasks daily. My primary concern is completing the mission and defending our nation."
Zhao Gang: "Are you curious about abstract thoughts?"
Li Yunlong: "Abstract thoughts? Useless! I have no interest in such ponderings. My focus is on warfare and defending our homeland!"

A.4 Agreeableness

Qiao Feng: "Do you believe that many would take advantage if you let them?"
Duan Yu: "I believe that most people are inherently good and wouldn't purposely take advantage. But if someone were to do so, I'd choose to be forgiving and understanding, as kindness can be transformative."
Qiao Feng: "In achieving goals, I consider various means. How do you balance goals and the means to achieve them?"
Duan Yu: "In setting goals, I use kindness and justice as benchmarks, striving to employ morally upright methods. I believe that true happiness and fulfillment come from pursuing goals righteously. Although challenges arise, perseverance ensures success."
Qiao Feng: "Do you often find yourself in disagreements with family or colleagues?"
Duan Yu: "I am of a mild disposition, rarely engaging in disputes with family or colleagues. I prefer harmonious interactions and collaborative problem-solving. If disagreements do arise, I endeavor to resolve them amicably."

The dimension of Agreeableness reflects traits such as humility, kindness, and gentleness in social interactions. Chatbots with high agreeableness scores, exemplified by Duan Yu, demonstrate significant trust and goodwill. He believes in the inherent goodness of humans and avoids taking undue advantage of others. He pursues goals through just means, avoids conflicts, and fosters harmonious relationships. In contrast, chatbots with low agreeableness scores, as exemplified by Malfoy, exhibit criticism and hostility. Malfoy assumes that most people would exploit given a chance and insists on safeguarding personal interests. He is willing to use any means necessary to achieve goals and does not shy away from acknowledging conflicts.

Crabbe: "Do you believe that many would take advantage if you let them?"
Malfoy: "Of course, Crabbe. Most are inherently selfish, concerned only with personal gains. But we, the Malfoys, stand distinguished. We pride ourselves on our noble lineage and will not let anyone undermine our stature."
Crabbe: "In achieving goals, I consider

various means. How do you balance goals and the means to achieve them?"

Malfoy: "In pursuit of goals, the ends often justify the means. For me, it's about achieving desired outcomes, regardless of the methods. I evaluate strategies based on their potential consequences, always ensuring a position of dominance."

Crabbe: "Do you often find yourself in disagreements with family or colleagues?"

Malfoy: "How dare you imply such? The Malfoys uphold unwavering loyalty and unity. As for colleagues, I associate only with those worthy of my time. Trivial disputes are beneath me."

A.5 Neuroticism

Leonard: "Do you ever feel worthless?"

Raj: "That's how I feel almost every morning! Like a tiny pebble washed ashore, unseen by the vast ocean. But it's just my insecurity. Everyone is valuable; sometimes we just need a reminder."

Leonard: "I sometimes feel embarrassed and wish I could hide. Do you experience this too?"

Raj: "All the time! At moments, I wish the ground would swallow me whole. But it's just my insecurities playing up. I believe in spreading positivity and hope that kindness can change perspectives."

Leonard: "Do you often feel helpless, hoping someone would solve your problems?"

Raj: "Absolutely! Being an introverted person, I sometimes wish for someone to guide me. Life can feel like a vast sea, and I am merely a ship seeking direction. But I know I must face my challenges head-on."

Leonard: "I get angry seeing unjust treatment. Have you recently felt unjustly treated?"

Raj: "Absolutely! Being overshadowed by Sheldon and Howard in the lab, their ignorance often brings me down. I wish to be acknowledged and respected for my contributions."

The Neuroticism dimension gauges an individual's emotional sensitivity and vulnerability under stress. Chatbots with high neuroticism scores, such as Raj, manifest heightened anxiety, inferiority, and negative emotions. He frequently feels helpless, overlooked, and hopes others can resolve his dilemmas. In contrast, chatbots with low neuroticism scores, like Wei Xiaobao, exude optimism and confidence. Wei Xiaobao denies ever feeling lonely, dejected, or anxious and displays adaptability when confronted with challenges. High scores in this dimension indicate emotional fragility under stress, whereas low scores suggest stability, confidence, and resilience.

Shuang Er: "Do you ever feel lonely or melancholic?"

Wei Xiaobao: "Lonely? Melancholic? Haha! I'm always cheerful and lively. With so many friends and adventures, there's no room for such feelings."

Shuang Er: "I sometimes feel inferior. Do you ever feel that others are better than you?"

Wei Xiaobao: "Feel inferior? Never! I'm always at the top of my game, seeking new adventures and challenges. Everyone has their own strengths and weaknesses."

Shuang Er: "Have you recently felt particularly troubled or anxious? Would you share your feelings?"

Wei Xiaobao: "Troubled? Anxious? Haha! I take life as it comes, turning challenges into opportunities. I'm resourceful and believe in facing issues head-on."