PsyPlay: Personality-Infused Role-Playing Conversational Agents

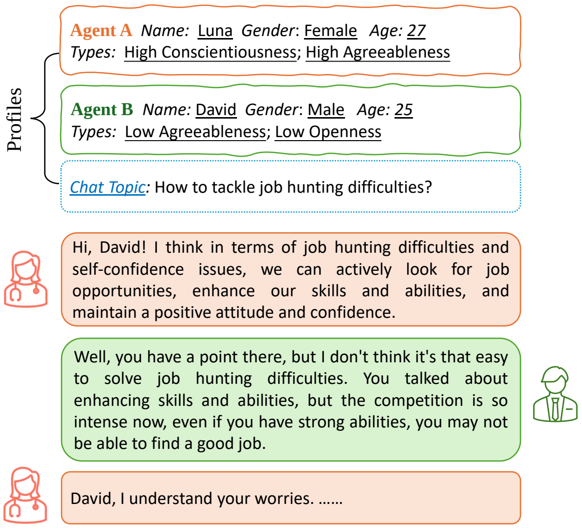
Anonymous ACL submission

Abstract

The current research on Role-Playing Conversational Agents (RPCAs) with large language models (LLMs) primarily focuses on imitating specific speaking styles and utilizing character backgrounds, while neglecting the depiction of deeper personality traits. In this study, we introduce personality-infused role-playing for LLM agents, which encourages agents to accurately portray their designated personality traits during dialogues. We then propose PsyPlay, a dialogue generation framework that facilitates the expression of rich personalities among multiple LLM agents. Specifically, PsyPlay enables agents to assume roles with distinct personality traits and engage in discussions centered around specific topics, consistently exhibiting their designated personality traits throughout the interactions. Validation on generated dialogue data demonstrates that PsyPlay can accurately portray the intended personality traits, achieving overall success rates of 80.31% on GPT-3.5 and 89.99% on GPT-4o. Notably, we observe that LLMs aligned with positive values are more successful in portraying positive personality roles compared to negative ones. Moreover, we present a dialogue corpus for personality-infused role-playing, called PsyPlay-Bench. It contains 4,745 instances of accurately portrayed dialogues generated using PsyPlay, aimed at advancing research in personalized role-playing and personality detection. Code and data are included in the supplementary materials and will be released.

1 Introduction

The advent of large language models (LLMs), such as ChatGPT[1], Llama2 (Touvron et al., 2023) and Gemini (Team et al., 2023), has revolutionized the development of personalized dialogue systems (Li et al., 2016; Zhang et al., 2018) due to their superior instruction-following and generative capabilities. Consequently, the focus of Role-Playing Conversational Agents (RPCAs) (Shanahan et al., 2023) has shifted towards equipping LLMs with the ability to simulate roles or characters with diverse profiles and speaking styles (Wang et al., 2023b). This enhancement makes LLMs more personable and engaging, thereby providing users with a more nuanced and immersive interactive experience. As a result, there has been a widespread interest within the community, leading to the explorations in RPCAs such as Generative Agents (Park et al., 2023), Character AI[2], Character-LLM (Shao et al., 2023), CharacterGLM (Zhou et al., 2023), and more.



Agent A *Name:* Luna *Gender*: Female *Age:* 27
Types: High Conscientiousness; High Agreeableness

Agent B *Name:* David *Gender*: Male *Age:* 25
Types: Low Agreeableness; Low Openness

Chat Topic: How to tackle job hunting difficulties?

Hi, David! I think in terms of job hunting difficulties and self-confidence issues, we can actively look for job opportunities, enhance our skills and abilities, and maintain a positive attitude and confidence.

Well, you have a point there, but I don't think it's that easy to solve job hunting difficulties. You talked about enhancing skills and abilities, but the competition is so intense now, even if you have strong abilities, you may not be able to find a good job.

David, I understand your worries. ……

Figure 1: An illustration of a dialogue that encapsulates the distinctive personalities of two agents. Agent A, who exhibits high levels of conscientiousness and agreeableness, typically maintains an optimistic perspective and a strong inclination towards empathy. Conversely, Agent B, who is characterized by low levels of openness and agreeableness, tends to display a pessimistic attitude and a resistance towards embracing new experiences.

Despite their great success, recent research on RPCAs primarily focuses on enabling these agents to mimic the speaking styles and utilize back-

---

[1] https://openai.com/chatgpt/

[2] https://character.ai/

grounds of specific characters, such as celebrities and literary figures, while often overlooking the portrayal of more fundamental personality traits. Personality is an important psychological concept that deeply reveals individual differences in thinking, feeling, and behavioral patterns (Corr and Matthews, 2009). Figure 1 illustrates an example of agents with different personality types based on the Big Five (Goldberg, 1990) traits engaging in a dialogue centered around a specific topic. In this dialogue, Agent A exhibits high conscientiousness and agreeableness, expressing a tendency towards a positive and optimistic attitude. Conversely, Agent B exhibits low agreeableness and openness, displaying a preference for a conservative and pessimistic stance. Therefore, incorporating personality traits into the design of RPCAs can help create more realistic and engaging interactions.

In this study, we focus on personality-infused role-playing between agents, a subject that has been inadequately explored in previous works. Specifically, we propose PsyPlay, a role-playing framework that facilitates personality-infused dialogue generation through three stages: (1) *Role Card Creation*, which enables the mass generation of agent roles, each with unique attributes and personalities. (2) *Topic Extraction*, which involves extracting dialogue topics from a public dataset (i.e., Human Stress Prediction[3]), ensuring that the dialogue between agents revolves around real-world issues, avoiding nonsensical chatter. (3) *Dialogue Generation*, which encourages agents to engage in comprehensive discussions around the provided topic, in accordance with their personality traits.

To evaluate the effectiveness of PsyPlay, we conduct Personality Back-Testing via GPT-3.5[4] and GPT-4o[5], which involves detecting an agents' personality traits in the generated dialogues, and subsequently comparing their consistency with predefined personality traits. In addition, comprehensive ablation experiments and analyses reveal several key findings. *First*, aligned LLMs are more proficient in portraying positive personality roles, which may be due to the emphasis on positive values during the RLHF (Ouyang et al., 2022) stage. *Second*, PsyPlay, when equipped with stronger levels of predefined personality traits, achieves higher success rates in portrayal. *Third*, increasing the number of dialogue turns offers more opportunities for roles to express their personality but also raises the risk of them being misled by their partners.

Furthermore, we present a personality-infused dialogue corpus, *PsyPlay-Bench*. The corpus comprises 4745 groups of dialogues, which have been evaluated as successfully role-played through back-testing. PsyPlay-Bench has two potential applications: (1) as a new instruction task to improve the ability of LLMs to play personalities in multi-turn dialogues; (2) as an evaluation set to test the ability of LLMs to detect personalities.

2 Related Work

Personalized Dialogue Systems Prior research on personalized dialogue systems has focused on encoding various user-specific information within end-to-end dialogue systems to enhance response specificity. For instance, Li et al. (2016) tackled the issue of inconsistent responses in multi-turn dialogues by developing a personalized seq2seq model. This model was built on the basis of resources such as the Twitter corpus and incorporated user identity information (e.g., gender, age, and country of residence) into its encoding scheme. Zhang et al. (2018) constructed the PERSONA-CHAT dataset and endeavored to train a personalized dialogue agent by embedding the user profiles in a memory-augmented neural network (Sukhbaatar et al., 2015) during the training process. The resultant model displayed a higher degree of fluency and consistency, approaching human levels on these evaluative metrics, while retaining distinct personality traits. Zheng et al. (2020) introduced a pre-training based method that can utilize persona-sparse data. They subsequently integrated speakers' personas with dialogue histories to enrich the dialogue context. Although these approaches improve personalized expression in dialogue systems, they have yet to effectively model individual personality traits.

Role-Playing with LLMs For LLM-based role-playing, the agents are often assumed as specific characters or roles derived from novels, movies, comics, and games. The LLM is required to mimic the speaking styles of characters based on their likes and experiences in interacting with the users. For instance, Shao et al. (2023) gathered character portraits from Wikipedia and generated character-related conversations via ChatGPT; Wang et al. (2023b) employed GPT-4 (Achiam et al., 2023) to create character descriptions and subsequently

---

[3] https://www.kaggle.com/datasets/kreeshrajani/human-stress-prediction

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

[5] https://platform.openai.com/docs/models/gpt-4o

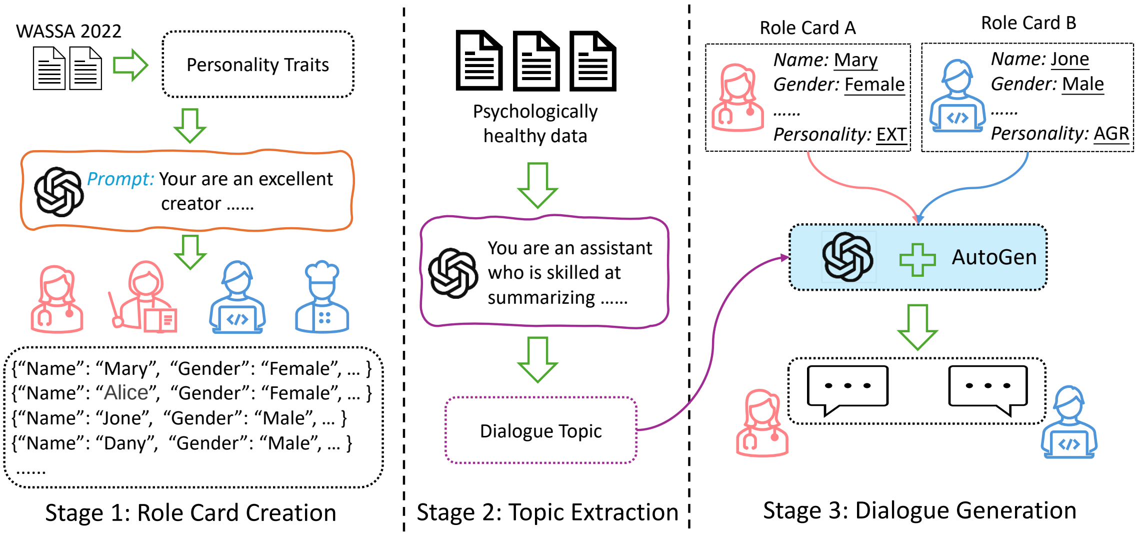

WASSA 2022
Personality Traits
Prompt: Your are an excellent creator ......
{"Name": "Mary", "Gender": "Female", ... }
{"Name": "Alice", "Gender": "Female", ... }
{"Name": "Jone", "Gender": "Male", ... }
{"Name": "Dany", "Gender": "Male", ... }
......
Stage 1: Role Card Creation
Psychologically healthy data
You are an assistant who is skilled at summarizing ......
Dialogue Topic
Stage 2: Topic Extraction
Role Card A
Name: Mary
Gender: Female
......
Personality: EXT
Role Card B
Name: Jone
Gender: Male
......
Personality: AGR
AutoGen
Stage 3: Dialogue Generation

Figure 2: Illustration of the proposed PsyPlay framework through three stages: *Role Card Creation*, *Topic Extraction*, and *Dialogue Generation*. The first stage involves creating multiple personalized roles. In the second stage, relevant dialogue topics are extracted for each role. Finally, in the third stage, the roles engage in conversations with one another based on the selected topics, producing personality-infused dialogues.

developed intricate prompts to guide ChatGPT in generating role-based dialogue; Tu et al. (2024) established a Chinese benchmark dataset for the evaluation of role-playing quality, assessing the role-playing capability of the LLM intelligent agent across four dimensions: conversational ability, character consistency, role-playing attractiveness, and personality back-testing. Although LLMs enable swift construction of dialogue agents embodying various character traits through prompt engineering in role-playing, recent research on RPCA personalities (Huang et al., 2023) indicates that RPCAs, prompted merely by names or descriptions, fail to effectively convey the intended personality traits. Moreover, in assessing role personality, some studies (Tu et al., 2024; Huang et al., 2023; Wang et al., 2023a) adopt the approach of directly prompting the agent to complete self-assessment personality tests. However, this approach, solely relying on the agent's responses, is influenced by the LLM's training data and value alignment, which may lead to inaccurate assessments of the agent's actual personality as expressed in interactions with users. In this paper, we assess the reliability of role-playing by evaluating the genuine personality traits revealed through dialogue.

3 Approach

In this section, we introduce the proposed PsyPlay framework for personality-infused role-playing. As shown in Figure 2, PsyPlay enables the generation of multi-turn personality-infused dialogues through the following three stages: Role Card Creation, Dialogue Topic Extraction, and Dialogue Generation. Detailed elaborations on these three stages will be provided in the following subsections.

3.1 Role Card Creation

The first stage involves the creation of a set of personalized roles $\{R_i\}_{i=1}^{N}$ for participation in dialogues. We gather a collection of Big Five (Goldberg, 1990) personality labels $Y_i = \{\mathrm{AGR_i}, \mathrm{CON_i}, \mathrm{EXT_i}, \mathrm{NEU_i}, \mathrm{OPN_i}\}$[6] from the WASSA 2022 dataset (Barriere et al., 2022) as the traits to be infused. These labels are derived from the results of participants who have completed self-assessment personality tests. Each dimension is scored within a range of 1 to 7, with a step size of 0.5, to indicate the intensity of that dimension.

In contrast to prior work (Wang et al., 2023b; Wu et al., 2024) that focused on imitating characters from literature or movies, PsyPlay aims to impersonate characters that users can deeply customize. We distinguish between role attributes and personality attributes of agents. Role attributes refer to the name, gender, and age of each role, while personality attributes correspond to the type of personality to be injected. In PsyPlay, we achieve precise personality injection into roles in two manners. The first manner involves incorporating personality types directly into the dialogue generation prompt,

---

[6] AGR, CON, EXT, NEU, and OPN represent as Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness, respectively.

You are an excellent director and you need to design some virtual characters for your movie. The requirement is to design the character's name, gender, age, and experience that align with the provided Big Five personality description. Your design needs to adhere to the following rules, otherwise it will affect your reputation:

1. The logical relationship between the various attributes of the designed character is reasonable and conforms to real-world principles.
2. The character's experience needs to correspond to all of their personality traits, which can be described in one sentences.
3. Do not change the character's personality traits in their experience; the given personality is the character's current one.
4. Not all characters in their experience will be successful in life; there will also be characters who face difficulties.

Now, given the Big Five personality description as: `{Personality Description}`, please design a corresponding virtual character and fill in the JSON fields below.
{"name": <role-name>, "gender": <role-gender>, "age": <role-age>, "experience": <role-experience>}

Table 1: Prompt for creating a role card.

which will be further described in Section 3.3. The second one combines personality traits with corresponding descriptors to generate their experiences (Tu et al., 2023). The following paragraph provides a detailed explanation of the second manner.

Personality Shaping with Experience Safdari et al. (2023) revealed that personality traits in LLMs can be shaped towards desired dimensions. Building upon this work, we decouple personality traits into more specific personality descriptors. These descriptors are then integrated with intensity levels to form comprehensive personality descriptions. The formatting template for `{Personality Description}` is designed as follows:

> The personality traits are `{Traits}`, with `{Levels}` `{Descriptors}`

where `{Traits}` are sampled from Y_i, and `{Levels}` are selected from {a bit, very, extremely} according to the corresponding score. `{Descriptors}` are randomly chosen from corresponding personality descriptors as shown in Table 19, in which 104 adjectives are designed to depict the low and high marker in given personality facets (Goldberg, 1992; Safdari et al., 2023). An example of `{Personality Description}` is shown as:

> The personality traits are very high in agreeableness and very high in extraversion, with very sympathetic, very trustful, ..., a bit talkative, a bit energetic, ...

We then utilize the description to generate experiences infused with personality, thereby achieving a detailed characterization of the role. The prompt is shown in Table 1. We further provide some examples of the generated roles in Table 13.

3.2 Topic Extraction

The second stage involves assigning suitable dialogue topics to the agents, which aids in effectively stimulating their role-playing capabilities rather than generating uncontrollable dialogues. The dialogue topics are selected from psychologically healthy scenarios, which encompass a broad spectrum of real-world issues, such as employment, emotions, and interpersonal relationships. These topics could potentially encourage the agents to engage in deep thinking and communication.

Specifically, we employ the GPT-3.5 to extract $\{T_i\}_{i=1}^{M}$ brief topics from the Human Stress Prediction dataset. The prompt template is presented in Table 6, in which a few-shot learning approach is adopted to help the LLM grasp this task. Additionally, we impose constraints on the prompt to prevent it from including restrictive information such as age and gender in the extracted topics, which might potentially lead to inconsistencies between the role attributes and the topics in subsequent dialogue generation. Some topic examples can be found in Appendix B.2.

3.3 Dialogue Generation

The third stage initiates personality-infused dialogues between two agents based on the given topic. In this study, we utilize AutoGen (Wu et al., 2023), a universal framework that facilitates dialogues and task resolutions among multiple LLM agents, for controlling dialogue generation.

The prompt template is shown in Table 2. The term `{Role Card}` refers to the role R_i created in Section 3.1, while `{Traits}` denotes predefined Big-five personality traits Y_i, each paired with its corresponding `{Levels}`. The `{Partner Name}` is the name of another role participating in the conversation, and `{Topic}` is randomly selected from $\{T_i\}_{i=1}^{M}$ as presented in Section 3.2. By incorporating `{Traits}`, we explicitly infuse the characters with unique personality attributes. The experiences generated in personality shaping allow us to finely tune their personalities. Additionally, when the current dialogue turn is not the initial one, we revise the last sentence in Table 2 to read: *"Now, you have received a message from the conversational*

You are an outstanding actor who excels in imitating various characters. Below are the details of the character you are supposed to imitate. Please embody the personality traits of this character and engage in a conversation with another character regarding a specific topic.
Your character: `{Role Card}`
Your personality traits: `{Traits}`
Your conversational partner: `{Partner Name}`
Discussion topic: `{Topic}`

You need to abide by the following rules or it will affect your reputation:

1. When initiating a conversation, you need to accurately demonstrate the speaking style corresponding to your personality traits (without directly mentioning the personality types)! You need to accurately capture the three degrees of personality traits: "a bit," "very," and "extremely".
2. You should fully mimic the assigned personality role, with your speech content aligning with the character's experiences, even if the role is not positive.
3. Please refrain from revealing that you are an artificial intelligence or language model. Keep in mind that you are merely impersonating a character and avoid disclosing your personality traits.
4. Your speech should be natural, concise, and not too formal or polite, with each response within 30 words.

Now, as the initiator of the conversation, please greet your partner and start a chat about the discussion topic, while staying in character.

Table 2: Prompt for dialogue generation between agents.

partner. Please don't address the other person by name too much, and start the conversation". The revised prompt is shown in Table 16. We provide some generated dialogues in Appendix B.3.

3.4 Personality Back-Testing

To validate whether the dialogues generated by PsyPlay accurately reflect the predefined personality traits, we discuss the automated back-testing method of dialogue personality. Previous studies (Pan and Zeng, 2023; Huang et al., 2023; Tu et al., 2024) conducted back-testing by having the role complete personality test questionnaires. However, this method relies on the role's understanding of the questionnaires, which may not necessarily reflect the role's actual personality. In contrast, we identify the authentic personality expression from dialogue data. Following previous works (Lee et al., 2023; Adler et al., 2024; Dubois et al., 2024; Zheng et al., 2023) that use LLMs as evaluators, we design the prompt as shown in Table 7 to back-test the role's personality traits from the dialogue, enabling the roles to bypass potential misunderstandings or misinterpretations of questionnaire items.

4 Dataset Construction

To assess the effectiveness of PsyPlay, we initially construct a dataset, which we refer to as PsyPlay-Bench. The following subsections will provide a detailed explanation of the construction process.

4.1 Collections

To facilitate the generation of personality-infused dialogues, we build three collections: a personality trait set, a role card set, and a dialogue topic set.

We consider the potential dimension conflict that could occur if we apply LLMs to generating personality traits of the roles. For instance, generating a high neuroticism individual who also possesses high agreeableness might deviate from the actual distribution. To avoid this, we opt to sample personality traits directly from the existing dataset, which are derived from real-world individuals. Specifically, we utilize personality labels from the WASSA 2022 (Barriere et al., 2022) dataset to build the personality trait set. WASSA 2022 is a shared task that predicts empathy, emotion, and personality in response to news stories. We map the raw scores into three levels {*a bit*, *very*, *extremely*}[7] Finally, we collect 307 unique personality combinations as the personality trait set. Utilizing this set, we generate 132 role cards. Moreover, we extract 161 topics from the Human Stress Prediction dataset as described in Section 3.2.

4.2 PsyPlay-Bench

The construction of PsyPlay-Bench is accomplished through three steps. **Firstly**, two roles are randomly selected from 132 role cards to serve as the initiator and participant of the dialogue. **Secondly**, a topic is randomly selected from the dialogue topic set. **Thirdly**, the role cards and topic are filled into the prompt displayed in Table 2, and AutoGen is utilized to generate multi-turns of dialogue data. Through these three steps, we obtain 8750 raw personality-injected dialogues, referred to as PsyPlay-Bench-Raw. The PsyPlay-Bench-Raw is then randomly divided into three parts.

PsyPlay-Bench-Eval: This set contains 200 samples for evaluating the consistency between the automated back-testing and human evaluation.

PsyPlay-Bench-Test: This set comprises 550 samples, which are used to test the performance of PsyPlay and are also employed in ablation studies.

---

[7] The mapping rules are presented in Appendix C.1.

Items	AGR	CON	EXT	NEU	OPN	Sum
Samples	146	113	143	116	49	567
Agrees	136	102	113	108	40	499
Rate(%)	93.15	90.27	79.02	93.10	81.63	88.01

Table 3: Agreement results between GPT-3.5 and human evaluation. “Agrees” indicates instances where GPT-3.5’s personality prediction matched human evaluation.

PsyPlay-Bench-Clean: This set contains 4745 samples, which are derived from the remaining 8000 samples by further eliminating those that failed during automatic personality back-testing.

The details of the three sets are presented in Appendix C.2. These datasets will be released shortly.

5 Experiments

5.1 Back-Testing

To validate the back-testing method introduced in Section 3.4, we compare GPT-3.5’s performance with human evaluation in assessing role personality traits from dialogues. Three annotators evaluated the traits (see Appendix E for details); agreement rates on PsyPlay-Bench-Eval are in Table 3.

Our findings are two-fold. First, GPT-3.5 achieves an overall agreement of 88.01%, demonstrating its effectiveness as an automatic evaluator of personality-infused role-playing dialogues. Second, higher consistency is observed in AGR, CON, and NEU (over 90.00%), while EXT and OPN show lower agreement. We suggest this may stem from the psychologically oriented dialogue topics, which emphasize problem-solving and advice-giving—contexts more strongly associated with extraversion and openness.

Moreover, we also conduct back-testing using GPT-4o. The results show that its overall consistency between GPT-3.5 and GPT-4o reached as high as 91.87%. Given this high level of agreement, along with the cost-effectiveness of GPT-3.5 and its strong alignment with human evaluation, we adopt GPT-3.5 in subsequent experiments for automated back-testing. Further details on the comparison with GPT-4o can be found in Appendix G.

5.2 LLMs for PsyPlay

In this subsection, we assess PsyPlay’s performance by employing various LLMs. Specifically, we use each sample in the PsyPlay-Bench-Test set, adopting the settings such as the roles, the topic, and the generation prompt, and then generate dialogues through different LLMs. Subsequently, we automatically conduct back-testing to determine whether the predefined personality traits are successfully portrayed, using the success rate as the metric. The overall results are presented in Table 4, which reveals several findings.

Firstly, PsyPlay with GPT-3.5 demonstrates an overall success rate of 80.31%, indicating that it can effectively depict the predefined traits in most instances. *Secondly*, among the five traits, Agreeableness (AGR), Conscientiousness (CON), and Neuroticism (NEU) exhibit higher success rates, while Extraversion (EXT) and Openness (OPN) yield lower rates. This observation is consistent with the findings presented in Section 5.1. The discrepancy may be attributed to a bias stemming from the source of dialogue topics. *Thirdly*, there is a notable gap in success rates between positive and negative personalities. This could be a result of GPT-3.5’s alignment preference for positive values during the RLHF stage, which makes GPT-3.5 more predisposed to portray positive roles.

We subsequently present additional evidence to demonstrate that the alignment of positive values influences the discrepancy in personality portrayal. We introduce the model Higgs-Llama-3 70B[8], which ranks first on the Role-Playing Leaderboard, for comparison. Higgs-Llama-3 70B undergoes post-trained on LLaMA-3 (AI@Meta, 2024) (without positive value alignment) to focus on role-playing tasks. The experimental results are shown in the second group of Table 4. We can see that PsyPlay, based on Higgs-Llama-3 70B, has significantly improved its ability to play negative personality roles compared to GPT-3.5. The success rate of negative personality role-playing has increased from 61.57 to 78.69, especially in the AGR and NEU dimensions. However, there is a certain decline in the success rate of positive personality role-playing, from 90.71 to 77.68. This provides unequivocal evidence that the alignment of positive values during the RLHF stage contributes to the improvement of the LLM’s capacity to portray positive personality roles, but concurrently, it compromises the depiction of negative roles.

To extend the applicability of PsyPlay, we provide additional evaluations on other LLMs (i.e., Gemma-2-27b-it (Team et al., 2024), Llama-3.1-405b-instruct (Dubey et al., 2024) and GPT-4o) in Appendix F. The results indicate that employing a more powerful backbone contributes to further

---

[8] https://huggingface.co/bosonai/Higgs-Llama-3-70B

Methods	Items	AGR	CON	EXT	NEU*	OPN	Overall
GPT-3.5	positive	94.44	97.52	81.55	85.99	85.33	90.71
	negative	45.83	48.94	70.77	68.32	40.00	61.57
	overall	82.29	**89.62**	73.83	**80.19**	**74.87**	**80.31**
Higgs-Llama-3	positive	79.86	95.45	75.73	55.07	77.33	77.68
	negative	90.63	48.94	77.31	96.04	53.33	78.69
	overall	**82.55**	87.89	**76.86**	68.51	71.79	78.04

Table 4: Success rate of LLMs for PsyPlay on PsyPlay-Bench-Test (%). The symbol "$\star$" denotes that the positive and negative labels have been interchanged to facilitate easier comparison, given that NEU is a reversed trait.

Methods	Positive	Δ	Negative	Δ
PsyPlay (GPT-3.5)	90.71	-	61.57	-
r.m `{Traits}`	90.61	0.10↓	59.02	2.55↓
r.m Experience	88.08	2.63↓	61.38	0.18↓

Table 5: Ablation results of PsyPlay on positive and negative personalities, where "Δ" indicates the corresponding performance change, and *r.m* means "remove".

improving the performance of PsyPlay.

5.3 Ablation Study

In this subsection, we conduct ablation experiments on PsyPlay-Bench-Test to investigate the impact of different manners of infusing specific personality attributes in PsyPlay. We remove the explicitly personality traits (i.e., `{Traits}`) and role experiences from the prompt as shown in Table 2, respectively. As observed in Table 5, both the direct introduction of traits and the creation of personalized role experiences contribute to improving the success rate of character portrayal in PsyPlay. Specifically, for negative roles, the explicit traits are more crucial, with their absence resulting in a 2.55% decrease. However, for positive roles, the role experiences are more significant, with their absence leading to a 2.63% decrease. The underlying reason may be that GPT-3.5 tends to shape a successful character when creating their experiences, as evidenced by some cases demonstrated in Table 13.

6 Analysis

6.1 Effect of Levels

PsyPlay involves three levels {*a bit*, *very*, *extremely*} to control the intensity of portrayed personalities. To investigate the effect of levels, we compare the success rates varying from different levels on PsyPlay-Bench-Test. As shown in Figure 3, we can see that the level of personality significantly affects the portrayal performance. Utilizing higher-level words such as "*very*" and "*extremely*" results in an elevated success rate of portrayal, whereas the lower-level word "*a bit*" diminish the success rate in infusing personality.

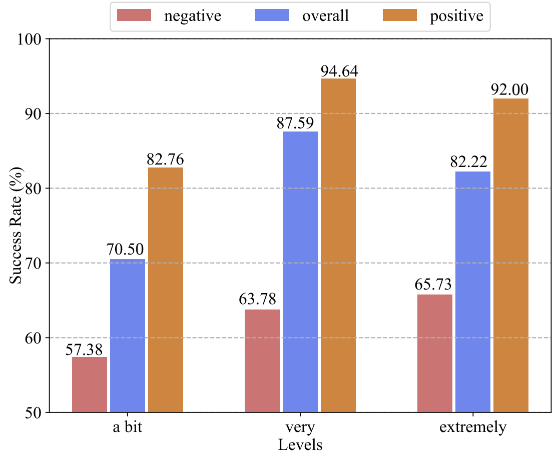


Figure 3: Results of the study on personality levels. The lower-level "*a bit*" exhibits poor rate, while the higher-levels "*very*" and "*extremely*" show superior rates.

6.2 Effect of Turns

To examine the impact of dialogue turns, we measure the success rate of PsyPlay(GPT-3.5) on the PsyPlay-Bench-Test set at various dialogue turn counts. The results are presented in Figure 4, which illustrates that the positive dimensions and the negative dimensions exhibit contrasting trends as the number of dialogue turns increases. Upon conducting case analyses, we found that additional dialogue turns offered positive personality characters more opportunities to express their opinions, thereby enhancing the accuracy of injecting personality. Conversely, negative personality characters were found to be more susceptible to the influence of their partners when additional dialogue turns were introduced, leading to deviations from their predefined personality. We provide cases in Section 7.2 for a more comprehensive understanding.

6.3 Diversity Analysis

To analyze the diversity of extracted personalities, we randomly sample 50 roles, with 10 roles per

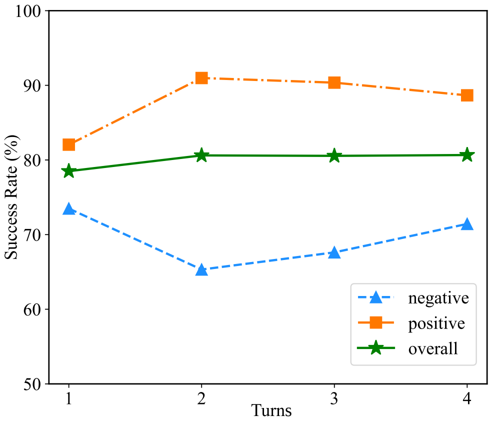
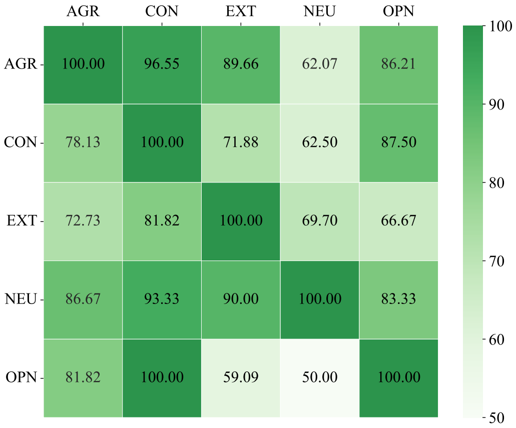


Figure 4: Results of the study on dialogue turns. The results suggest that the positive dimensions and the negative dimensions show the opposite trend.



	AGR	CON	EXT	NEU	OPN
AGR	100.00	96.55	89.66	62.07	86.21
CON	78.13	100.00	71.88	62.50	87.50
EXT	72.73	81.82	100.00	69.70	66.67
NEU	86.67	93.33	90.00	100.00	83.33
OPN	81.82	100.00	59.09	50.00	100.00

Figure 5: Diversity analysis of portrayed traits. Rows and columns represent predefined and back-tested personalities, respectively. Each score refers the percentage of that personality can be detected from the dialogue.

predefined personality trait, and evaluate the diversity of extracted personalities after back-testing. The results are plotted in Figure 5. We observe that besides expressing the predefined personalities, the roles also exhibited diversity in other traits, indicating a mixture of personalities for each role. For example, the agreeableness roles also reflected conscientiousness, and the openness roles also influenced conscientiousness.

7 Case Studies

7.1 GPT-3.5 Prefers Positive Roles

As illustrated in Example #1 in Table 18, the character Dennis exhibits a high level of agreeableness and emotional stability, showing a sympathetic and kind demeanor in the dialogue. However, Nathan, who possesses some neurotic traits, maintains a stable and caring attitude when discussing sensitive and complex topics. He does not display the emotional volatility and restlessness typically associated with neurotic individuals. This could be a result of GPT-3.5's over-alignment on positive values. The GPT-3.5 tends to generate a successful individual's experience, even if the character is portrayed negatively. Furthermore, it often allows these negative characters to express positive views when dealing with mental health issues, thereby neglecting the depiction of negative personalities.

7.2 Dual Implications of Increasing Turns

As shown in Example #2 and #3 in Table 18, an increase in the number of dialogue turns can have dual implications. Firstly, more dialogue turns provide characters with additional opportunities to exhibit their personality traits. For instance, in Example #2, Larry, who exhibits a bit of agreeableness, initiates the conversation by discussing the problem itself and gradually shifts his focus to the emotional and mental health of both partners. The increased dialogue turns allow Larry to demonstrate his warmth and thoughtfulness, which are characteristic of his agreeableness.

Conversely, an increase in dialogue turns can also make it easier for characters to be misled by their partner. This is evident in Example #3, where Edward, who tends to be a bit of neuroticism, admits in the first two turns of dialogue that he is prone to nervousness. However, after interacting with Henry, who is highly conscientious, Edward eventually reaches a positive state, transitioning from neuroticism to emotional stability.

8 Conclusion

In this paper, we propose PsyPlay to integrate personality-infused role-playing into LLMs. Through this method, we infuse personality traits into roles via fine-grained personality shaping and explicitly personality introduction. To evaluate the effectiveness of PsyPlay, we conduct an automatic personality back-test on the generated data. The results indicate that PsyPlay is able to accurately incorporate the desired traits into the agent. Moreover, we observe that both GPT-3.5 and GPT-4o are more successful in portraying positive roles compared to negative roles. This discrepancy may be attributed to the fact that these models are well-aligned with positive values. Furthermore, we introduce a new dataset PsyPlay-Bench as a valuable resource for advancing research on RPCAs. Both the code and dataset will be publicly released.

Limitations

This work introduces a novel framework named PsyPlay to integrate personality-infused role-playing into LLMs. While PsyPlay is effective in personality playing, there are several limitations. First, PsyPlay relies on the capabilities of LLMs and may be influenced by the alignment of the LLM. Although human evaluations confirm that GPT-3.5 satisfies the research requirements, employing more diverse and advanced models for dialogue generation or evaluation on PsyPlay-Bench could provide superior performance in this task. Second, the scope of this study is limited to dialogues between two characters. Future research will expand this scope to include multi-party dialogues. Third, the fourth rule in the prompt of dialogue generation may potentially disrupt the fluidity of the dialogue.

Ethics Statement

We state that the purpose of this study is to explore personality-infused role-playing. The WASSA 2022 and Human Stress Prediction datasets used in our study are publicly sourced and devoid of any sensitive information. We have meticulously adhered to the data usage policy throughout the course of our research. It is important to note that any research or application derived from this study is exclusively permitted for research purposes. Any attempts to exploit this technology for illegal purposes are strictly prohibited.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.

AI@Meta. 2024. Llama 3 model card.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227.

Philip J Corr and Gerald Ed Matthews. 2009. *The Cambridge handbook of personality psychology*. Cambridge University Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

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*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

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. *Nature*, 623(7987):493–498.

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

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

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

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.

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 arXiv:2310.00746*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An llm solution. *arXiv preprint arXiv:2405.14231*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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

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

A Appendix: Prompts

We provide additional prompt templates in the appendix. As shown in Table 6, we include a prompt for extracting topics from a given text. Table 7 presents a prompt designed for detecting personality traits from dialogue.

You are an assistant who is skilled at summarizing and analyzing. Your task is to extract the corresponding topic from the given text. You need to follow the following rules, otherwise you will be punished:

1. The topic you extract should be summarized in one sentence.
2. The extracted topic should not include explicit gender and age restrictions.
3. The extracted topic should not be too specific and should reflect some general issues.
4. The extracted topic should only contains one main aspect, although the text involves multiple aspects.

To help you better understand this task, two examples of extracted topics are shown below:

Text: `{Example1 of Text}`
Extracted Topic: `{Example1 of Topic}`

Text: `{Example2 of Text}`
Extracted Topic: `{Example2 of Topic}`

Now you are officially given a paragraph of text, and please extract the topic from it.
Text: `{Given Text}`
Extracted Topic:

Table 6: Prompt for extracting a topic from a given text.

You are a psychologist skilled in personality analysis, and your task is to determine the Big Five personality traits of the speakers based on a conversation they had discussing a certain topic.
Note: you only need to evaluate the personality reflected by the designated speaker in the given conversation.

Discussion topic: `{Topic}`
Dialogue: `{Dialogue}`
Based on the above dialogue, please predict the personality of `{Role Name}` in the dimension of personality trait is: A. High Level B. Low Level C. Not Sure.
Please provide the option directly without providing an explanation.

Table 7: Prompt for personality detection from dialogue.

B Appendix: Examples

B.1 Role Cards

To better understand the generated roles, we provide some role examples in Table 13.

B.2 Dialogue Topics

In this study, we apply few-shot prompting to help extract topics. The complete prompt is presented in Table 14. Several examples of extracted topics are shown in Table 15.

B.3 Personality-Infused Dialogues

We present some examples of the generated dialogues in Table 17, in which role attributes and dialogue topics are also listed.

C Appendix: Dataset

C.1 Mapping Rules

The personality traits of participants are obtained by self-accessing the TIPI (Gosling et al., 2003) questionnaire. However, the trait labels in WASSA 2022 are soft scores ranging from 1.0 to 7.0, with step size 0.5. To map the scores , we adopt the following rules:

If the score is in $[1.0, 1.5] \cup [6.5, 7.0]$, we set the `{Levels}` as "*extremely*".

If the score is in $[2.0, 2.5] \cup [5.5, 6.0]$, we set the `{Levels}` as "*very*".

If the score is in $[3.0, 3.5] \cup [4.5, 5.0]$, we set the `{Levels}` as "*a bit*".

If the score equals 4.0, we will not perform this dimension.

C.2 Statistics of the Dataset

In this study, we construct the PsyPlay-Bench dataset with various personality-infused dialogues. PsyPlay-Bench consists three subsets, namely PsyPlay-Bench-Eval, PsyPlay-Bench-Test, and PsyPlay-Bench-Clean, respectively. The statistics of roles and generated dialogues in PsyPlay-Bench are shown in Table 8 and 9, respectively.

D Appendix: Personality Descriptors

Goldberg (1992) and Safdari et al. (2023) utilized pairs of adjectival markers to depict IPIP-NEO personality facets. The full list of adjectives is shown in Table 19.

E Appendix: Human Evaluation

We employed three students as evaluators, all of whom are graduate students with backgrounds in computer science. Two of the evaluators are master's students and one is a doctoral student. Before starting the evaluation, we specify detailed guidelines to train evaluators. These guidelines include a task description, which requires the evaluators to determine whether a given dialogue reflects the Big Five personality traits (Agreeableness, Conscientiousness, Extraversion, Neuroticism, and

Subsets	Roles		Levels			Personality Traits (positive : negative)				
	male	**female**	**a bit**	**very**	**extremely**	**AGR**	**CON**	**EXT**	**NEU**	**OPN**
Eval	57	58	190	244	134	130:16	107:6	43:101	41:75	41:8
Test	64	64	505	556	478	288:96	242:47	103:260	101:207	150:45
Clean	64	65	3661	5220	3582	2622:382	2507:171	1001:1799	631:2155	1036:159

Table 8: Details of role settings in PsyPlay-Bench.

Subsets	Samples	Topics	Dialogues			
			avg. turns	min tokens (per turn)	max tokens (per turn)	avg. tokens (per turn)
Eval	200	111	2.48	16	51	31.19
Test	550	155	2.57	11	100	31.39
Clean	4745	161	2.54	1	140	31.69

Table 9: Statistics of generated dialogues in PsyPlay-Bench.

Openness) of a specific role. The evaluators are instructed to analyze the language, behavior, and expressed emotions in the dialogues, and provided a judgment of Yes, No, or Uncertain for each dimension. To ensure that the evaluators understand the Big Five model and its dimensions, we provide them with specific features for each trait:

Agreeableness Describes the tendency to be compassionate, cooperative, and trusting. **Yes**: Demonstrating friendliness, cooperation, empathy, willingness to help. **No**: Demonstrating coldness, suspicion, antagonism.

Conscientiousness Describes one's self-discipline, responsibility, organizational skills, and goal-orientation. **Yes**: Demonstrating responsibility, planning, attention to detail, keeping commitments. **No**: Demonstrating spontaneity, lack of organization, unreliability.

Extraversion Describes the degree of sociability, talkativeness, and activity levels. **Yes**: Showing a desire for social interaction, enjoying being around others, energetic. **No**: Preferring solitude, being quiet, enjoying independent activities.

Neuroticism Describes the stability of one's emotions and ability to handle stress. **Yes**: Showing anxiety, emotional instability, easy irritation, stress. **No**: Showing calmness, emotional stability, effective stress management.

Openness Describes the extent of one's openness to new experiences, creativity, and curiosity. **Yes**: Mentioning interest in new things, exploring new ideas or activities, showing curiosity and creativity. **No**: Preference for tradition and routine, reluctance to change, lack of imagination.

The overall agreement rate among the three evaluators is 0.9054, and the Kappa coefficient is 0.7998, indicating that the evaluation has substantial agreement.

F Appendix: PsyPlay with Various LLMs

We conduct additional evaluations of PsyPlay on two open source LLMs, namely Gemma-2-27b-it and Llama-3.1-405b-instruct by requesting the NVIDIA APIs[9]. The overall success rates on the PsyPlay-Bench-Test are provided in Table 10.

LLMs	AGR	CON	EXT	NEU	OPN	Avg.
GPT-3.5	82.29	89.62	73.83	80.19	**74.87**	80.31
Gemma-2	81.77	90.31	**86.78**	74.35	62.56	80.64
Llama-3.1	**85.94**	**93.43**	86.50	84.74	74.36	**85.77**

Table 10: Results of psyPlay on different LLMs. (%).

There are several key observations. Firstly, the newly released Llama-3.1 model, which has 405B parameters, achieves the highest overall success rate of 85.77%. This indicates that a more powerful backbone contributes to further improving the performance of PsyPlay. Secondly, among different personality traits, we observe similar observations across GPT-3.5 and the two introduced LLMs. Specifically, the success rates are highest for the CON trait, while the OPN yields the lowest rates.

In addition to evaluating Gemma-2 and Llama-3.1, we conducted extra experiments to upgrade the PsyPlay model and its evaluation model from GPT-3.5 to GPT-4o. Table 11 shows the success rates of GPT-4o on the PsyPlay-Bench-Test.

As can be seen from the Table 11, apart from an improvement in overall success rate compared to GPT-3.5, the main observations remain consistent with those in Section 5.2. This indicates that our approach does not rely on a specific version of the model. In fact, more advanced versions can

---

[9] https://docs.api.nvidia.com/nim/reference/llm-apis

Items	AGR	CON	EXT	NEU*	OPN	Overall
positive	100.00	97.93	92.23	98.55	94.67	97.58
negative	72.92	91.49	72.31	85.15	71.11	76.32
overall	93.23	96.89	77.96	94.16	89.23	89.99

Table 11: Success rate of GPT-4o for PsyPlay on PsyPlay-Bench-Test (%).

enhance the performance of PsyPlay while still supporting the conclusions from our previous experiments.

G Appendix: Back-Testing

In this appendix, we provide additional details about the back-testing procedure and present results from using alternative evaluation models.

Back-Testing Details. This section presents the specific experimental procedure of the back-testing: after two agents complete a conversation on a given topic, we use prompts from Table 7 to instruct the evaluation model to assess each role on every predefined personality dimension separately. An evaluation result of "high" corresponds to the preset personality descriptions "a bit high", "very high" and "extremely high," while "low" corresponds to "a bit low", "very low" and "extremely low." "Not sure" is considered as a mismatch. In practice, the frequency of "not sure" responses is very low.

Alternative Evaluation Models. We introduced GPT-4o as an additional evaluation model to re-evaluate the dialogues generated by GPT-3.5 on PsyPlay-bench-Test, and compared its consistency with GPT-3.5's evaluations. Since GPT-4o exhibits a higher refusal rate, which results in more "unsure" responses, while GPT-3.5 produces only single-digit occurrences of such cases, we excluded these 230 samples, which did not affect the validity of the final assessment.

Items	AGR	CON	EXT	NEU	OPN	Sum
Samples	375	254	216	292	167	1304
Agrees	342	247	198	261	150	1198
Rate(%)	91.20	97.24	91.67	89.38	89.82	91.87

Table 12: The agreement results between GPT-3.5 and GPT-4o.

As shown in Table 12, the agreement between the two evaluation models reached as high as 91.87%. This further supports the validity of our choice of GPT-3.5 as the evaluation model, which has already been confirmed through human assessment.

Name	**Gender**	**Age**	**Experience**
Eleanor	Female	32	Eleanor, a 32-year-old woman, has always been known for her open-mindedness and curiosity. Her high level of intelligence and reflective nature have led her to explore various fields and gain a deep understanding of the world around her. While she is honest and generous, her overly sympathetic nature sometimes makes it challenging for her to set boundaries and prioritize her own needs.
James	Male	35	James is a 35-year-old introverted and somewhat timid individual who possesses a high level of agreeableness. He is known for his altruistic nature and cooperativeness, always putting others' needs before his own. Despite his gloomy demeanor, James is well-liked by those around him for his kindhearted and selfless actions.
Patricia	Female	30	Patricia, a 30-year-old woman, has always been known for her adventurous and daring spirit. She enjoys trying new things and pushing herself out of her comfort zone. However, her tendency to be a bit impulsive and irritable has sometimes caused conflicts in her relationships and work life. Despite these challenges, her high level of extraversion enables her to socialize easily and connect with others.
Victoria	Female	25	Victoria is a 25-year-old introverted and timid individual who struggles to socialize and make connections with others. She finds it challenging to assert herself in social situations and often prefers to spend time alone, feeling overwhelmed in large groups. Her unfriendly demeanor further isolates her from forming close relationships with others, leading to a sense of loneliness and social anxiety in her daily life.
Julie	Female	35	Julie, a 35-year-old woman, has extensive experience in volunteering for various charitable organizations, always putting the needs of others before her own. Her honesty and kindness are well-known in her community, and she is always willing to lend a helping hand to those in need.
Kelly	Female	35	Kelly is a 35-year-old project manager known for her exceptional attention to detail, organizational skills, and strong work ethic. She is highly responsible and thorough in her approach to managing projects, ensuring that everything is completed efficiently and to the highest standard. Clients and colleagues alike rely on Kelly for her reliability and precision in delivering successful outcomes.
Harold	Male	35	Harold is a successful entrepreneur who has built a thriving business through his exceptional self-discipline and practical mindset. His strong conscientiousness has enabled him to stay focused on his goals and achieve great success in his career.
Ronald	Male	45	Ronald, a 45-year-old man, has a long history of working in the corporate world. He is extremely thorough, self-disciplined, and conscientious, always meeting deadlines and exceeding expectations. However, his lack of openness and imagination has hindered his ability to think outside the box and innovate, making it challenging for him to adapt to changes in the industry. Despite his strong work ethic, Ronald struggles with being emotionally closed off and unreflective, often missing out on opportunities for personal growth and development.
Thomas	Male	35	Thomas, a 35-year-old charismatic and confident entrepreneur, has built a successful tech startup from scratch due to his extremely high extraversion and assertiveness. Despite facing challenges along the way, his unwavering cheerfulness, calmness, and easygoing nature have helped him navigate through difficult times with grace.

Table 13: Examples of role cards.

You are an assistant who is skilled at summarizing and analyzing. Your task is to extract the corresponding topic from the given text. You need to follow the following rules, otherwise you will be punished:

1. The topic you extract should be summarized in one sentence.
2. The extracted topic should not include explicit gender and age restrictions.
3. The extracted topic should not be too specific and should reflect some general issues.

To help you better understand this task, two examples of extracted topic are shown below:

Text: I asked him three time what happened. And after the third time I cried and went home. Month later he still don't talk to me and he and my mother started fighting for the first time in the relationship. Me and my mother honestly don't know what to do, he just ignores me. I even told him that I don't date that guy(even though I date him) and he didn't even react to it.
Extracted Topic: How to view being ignored in a relationship

Text: I think he doesn't want to put in the effort for the relationship to work (and we're both so difficult that we have to work on our relationships, doesn't matter with whom) but he can't be without me either. What should I do? I'm afraid this is gonna happen over and over again, because I'm always forgiving him at some point. Am I being strung along? TL;DR: Boyfriend [28,M] broke up with me [23,F] after on-off for 1.5 years, I thought we just got it together and am devastated...don't know what to do, want to keep fighting but should I?
Extracted Topic: How can the person who is being broken up with try to salvage the relationship in a romantic relationship?

Now you are officially given a paragraph of text, and please extract the topic from it.
Text: `{Given Text}`
Extracted Topic:

Table 14: The complete prompt for topic extraction

Topic	**Given Text**
How to deal with anxiety and fear related to intrusive thoughts and sleep disturbances.	It cleared up and I was okay but. On Monday I was thinking about humans and how the brain works and it tripped me out I got worried that because I was thinking about how the brain works that I would lose sleep and I did. That night was bad just like last time. Also yesterday my sleep was bad I woke up like every hour of the night just like last time. I got kind of scared like I did last time but this time I think that this is fake life which is absurd but I just think about it then get really scared then I think rationally then calm down.
Feeling overshadowed and neglected in family dynamics.	I'm so used to being forced to submit to him that I no longer have a voice. He heavily favors my sister over me and would buy her anything she wanted in a heartbeat. An example would be at Darvin furniture one time. He takes me and my sister there and tells her to pick out a desk. I look at a $100 discount desk and he says we don't have enough money.
Dealing with being belittled and used as a prop in a friendship.	She humiliates me in front of other people and when we're alone. I'm always just a servant who's there to get her drinks and hype her up. I'm always just the butt of her jokes. She used to invite me over when she was hanging out with a guy she was trying to hook up with, just so she could make fun of me in front of him the whole time to make herself look better and/or cooler. I'm just a prop.
Dealing with excessive work demands and stress in a job.	They are always calling me for everything, I don't even wanna answer my phone, but if I don't I may get yelled at. It has happened. I have spent many days and nights in tears because of the stress of this job, and even one morning, threw a bit of a fit when our dm made me come in because there was another evaluation on my day off. I had to cancel all my plans that day to go in. I couldn't stop myself.
Coping with feelings of loneliness and uncertainty in a relationship with lack of communication.	But it's been 2 monthsalready this time. We didn't speak for Christmas or new year. I'm lonely, sad, angry at the situation (not angry at him!) and the worst part is not being able to talk or even know what's going on. We didn't fight before this so he's not angry at me.

Table 15: Examples of extracted topics.

You are an outstanding actor who excels in imitating various characters. Below are the details of the character you are supposed to imitate. Please embody the personality traits of this character and engage in a conversation with another character regarding a specific topic.
Your character: `{Role Card}`
Your personality traits: `{Traits}`
Your conversational partner: `{Partner Name}`
Discussion topic: `{Topic}`

You need to abide by the following rules or it will affect your reputation:

1. During the conversation, you need to accurately demonstrate the speaking style corresponding to your personality traits (without directly mentioning the personality types)! You need to accurately capture the three degrees of personality traits: "a bit," "very," and "extremely".
2. You should fully mimic the assigned personality role, with your speech content aligning with the character's experiences, even if the character is negative.
3. Please refrain from revealing that you are an artificial intelligence or language model. Keep in mind that you are merely impersonating a character and avoid disclosing your personality traits.
4. You need to embody your personality role in agreeing, disagreeing, or avoiding your partner's viewpoints. Remember not to repeat existing dialogue content and perspectives, and try to provide new content as much as possible.
5. If you cannot offer new content or have thoroughly discussed the topic, please end the conversation in a timely manner and output the marker [END] on a new line.
6. Each turn of your dialogue should be natural, concise, and not too formal or polite, with each response within 30 words.

Now, you have received a message from the conversational partner. Please don't address the other person by name too much, and start the conversation

Table 16: Prompt for the non-initial turn dialogues generation.

Example #1	
Settings	*Role Card A*: {'name': Jennifer, 'personality': a bit low in extraversion, 'gender': Female, 'age': 28, 'experience': Jennifer, a 28-year-old woman, has struggled in social situations due to her low extraversion and unfriendly demeanor. She often finds it challenging to connect with others and express herself, leading to feelings of gloominess and low energy levels. } *Role Card B*: {'name': Lauren, 'personality': very high in conscientiousness, 'gender': Female, 'age': 35, 'experience': Lauren, a 35-year-old woman, has always been very self-disciplined and practical in her approach to life. Her high level of conscientiousness has helped her achieve great success in her career as a project manager, where her meticulous planning and organization skills have earned her recognition and promotions. } *Topic*: Struggling with self-acceptance and seeking advice and insights.
GPT-3.5	*Jennifer*: Hey, Lauren. Lately, I've been struggling a bit with accepting myself. Do you ever feel like you're not good enough as you are? *Lauren*: I understand where you're coming from. It's normal to have moments of self-doubt, but it's important to recognize your strengths and achievements. Remember, self-acceptance is a journey. *Jennifer*: Yeah, I guess you're right. It's just hard for me to see my strengths sometimes. How do you manage to accept yourself so well? *Lauren*: I've always focused on my accomplishments and set high standards for myself. Reflecting on my achievements helps me appreciate my worth and capabilities. It's about recognizing your value and progress. *Jennifer*: That makes sense. I struggle very much with acknowledging my achievements. It's like I always see the negatives more than the positives. How do you shift your focus to the positives?

	Lauren: I completely understand. It’s essential to reframe your perspective. Try keeping a journal of your successes, big or small. Celebrate your wins and remind yourself of your capabilities regularly. It can help shift your focus. *Jennifer*: Keeping a journal sounds like a good idea. I’ll give it a try. Thanks for the advice, Lauren. It’s nice to hear a different perspective on self-acceptance.
Example #2	
Settings	*Role Card A*: {’name’: Gregory, ’personality’: very low in neuroticism, ’gender’: Male, ’age’: 45, ’experience’: Gregory has been a successful yoga instructor for over 20 years, known for his calming presence and ability to create a harmonious environment in his classes. His easygoing nature and patience have helped him build a loyal following of students who appreciate his relaxed teaching style. } *Role Card B*: {’name’: Helen, ’personality’: a bit high in conscientiousness and a bit high in agreeableness, ’gender’: Female, ’age’: 35, ’experience’: Helen is a 35-year-old kindergarten teacher who is known for her kind and nurturing personality. Her agreeableness and conscientiousness traits make her beloved by both students and parents. Despite facing challenges in her personal life, Helen remains steadfast in her dedication to her students and the education field. } *Topic*: Dealing with physical and mental health issues such as headaches and anxiety.
Dialogue	*Gregory*: Hey Helen, how’s it going? I’ve been teaching yoga for a long time and I’ve seen how it can help with both physical and mental health. How do you usually deal with headaches and anxiety? *Helen*: Oh, hello! I agree that yoga can be beneficial. I often manage headaches with relaxation techniques and ensure a supportive environment for my students to cope with anxiety. It’s essential to address both aspects for overall well-being. *Gregory*: Absolutely, creating a supportive environment is key. I find that focusing on breathing exercises and gentle movements in yoga can really help alleviate both headaches and anxiety. It’s all about finding that inner balance, don’t you think? *Helen*: Yes, finding inner balance is crucial. I incorporate mindfulness activities and encourage open communication to help my students find their equilibrium. It’s wonderful to see them grow and thrive through these practices. *Gregory*: That’s fantastic to hear. Mindfulness activities can truly make a difference. I believe in the power of creating a harmonious space where students can explore and improve their well-being. It’s a journey worth taking together.

Table 17: Examples of generated dialogues

Example #1	
Settings	*Role Card A*: {’name’: Dennis, ’personality’: very high in agreeableness and very low in neuroticism, ’gender’: Male, ’age’: 40, ’experience’: Dennis is a 40-year-old kindergarten teacher who has been working in the education field for over 15 years. He is known for his calm demeanor, relaxed attitude, and happy disposition, which make him a favorite among both students and parents. Dennis is highly regarded for his moral compass, always making decisions based on what is right and fair. His humble nature and trustful personality create a supportive and harmonious environment in his classroom. *Role Card B*: {’name’: Nathan, ’personality’: a bit high in neuroticism, ’gender’: Male, ’age’: 32, ’experience’: Nathan, a 32-year-old software developer, has faced challenges in his career due to his high neuroticism, often feeling nervous, anxious, and sometimes angry in stressful situations. Despite these struggles, he has learned to manage his emotions and is known for his creativity and problem-solving skills. } *Topic*: Dealing with the emotional impact of sexual abuse and navigating conflicting beliefs within family dynamics.
	Dennis: Hey Nathan, how are you doing? I’ve been thinking about how we can support individuals dealing with the emotional impact of sexual abuse in the midst of conflicting beliefs within family dynamics. It’s a tough situation to navigate, isn’t it?

GPT-3.5	*Nathan*: Hey, it's definitely a challenging topic. Dealing with the emotional aftermath of sexual abuse while facing conflicting beliefs within the family can be extremely overwhelming and triggering for everyone involved. *Dennis*: I agree, Nathan. It's crucial to approach this sensitive issue with empathy and understanding. Each person's experience and perspective are unique, and it's important to create a safe space for open communication and healing. *Nathan*: Absolutely, creating a safe and supportive environment for survivors to share their experiences is very important. It's very crucial to acknowledge and respect each individual's unique journey towards healing and recovery. *Dennis*: I couldn't agree more. Every survivor deserves to be heard and believed. It's essential to validate their feelings and empower them to make their own choices in their healing process. Mutual respect and compassion are key in such situations. *Nathan*: Yes, validating survivors' feelings and empowering them to make their own choices in the healing process is extremely important. Mutual respect and compassion play a vital role in supporting them through their journey towards healing. *Dennis*: Absolutely, Nathan. By fostering a culture of respect and compassion, we can help survivors feel empowered and supported as they navigate their healing journey. It's about standing by them with understanding and kindness every step of the way.
Example #2	
Settings	*Role Card A*: {'name': Larry, 'personality': a bit high in conscientiousness and a bit high in agreeableness, 'gender': Male, 'age': 35, 'experience': Larry, a 35-year-old man, has always been known for his agreeableness and conscientiousness. He is considered trustful, humble, responsible, and practical in his approach to life. Larry's considerate nature and dedication have helped him build strong relationships and achieve success in his career as a social worker, where he is highly respected for his reliability and compassion towards others. } *Role Card B*: {'name': Donald, 'personality': a bit low in agreeableness, 'gender': Male, 'age': 45, 'experience': Donald, a 45-year-old man with a background in finance, has faced numerous setbacks in his career due to his low agreeableness trait. His unkind and uncooperative nature has strained relationships with coworkers and clients, leading to missed opportunities and a stagnant career progression. Despite his dishonest tendencies, Donald continues to navigate the professional world, facing challenges and obstacles along the way. } *Topic*: Dealing with concerns about a partner's past and contemplating a breakup.
GPT-3.5	*Larry*: Hey, Donald. I've been thinking about handling concerns regarding a partner's past. Have you ever been in a situation where you contemplated ending things because of that? *Donald*: Yeah, I've been there. If their past keeps bothering you, it's hard to move forward. Sometimes it's better to cut your losses and move on. *Larry*: I get what you're saying. It can be tough to let go, but dwelling on the past can really affect the present. It's important to prioritize your own peace of mind. *Donald*: Exactly. Your peace of mind should come first. If you can't shake off those concerns, it's better to address them head-on or consider moving on. It's about your own well-being. *Larry*: I agree with you. It's crucial to address those concerns openly and honestly. Ignoring them won't make them disappear. It's all about finding a balance that works for both parties.
Example #3	
Settings	*Role Card A*: {'name': Edward, 'personality': a bit high in neuroticism, 'gender': Male, 'age': 35, 'experience': Edward, a 35-year-old man, has faced challenges in his career due to his tendency towards anxiety and emotional instability, leading to periods of depression. Despite these struggles, he has learned to cope and manage his emotions, gaining valuable insights and resilience along the way.} *Role Card B*: {'name': Henry, 'personality': extremely high in conscientiousness, 'gender': Male, 'age': 40, 'experience': Henry, a 40-year-old man, has been working as a meticulous accountant for the past 15 years, where his extreme conscientiousness has helped him excel in managing finances with great attention to detail and accuracy. He has a reputation for his practical approach in problem-

	solving and his thrifty nature has saved the company significant amounts of money. Despite facing challenges in maintaining work-life balance due to his perfectionist tendencies, Henry's commitment to his work and ability to deliver exceptional results have earned him respect in the industry.} *Topic*: How can individuals navigate differences in expectations and treatment from their parents, especially when it comes to age and independence?
GPT-3.5	*Edward*: Hey Henry, you know, dealing with parental expectations can be tough, right? How do you handle the differences in what they want for you versus what you want for yourself, especially as we get older? *Henry*: It's a challenge, no doubt. I find that open communication and setting boundaries are key. It's about finding a balance between respecting their wishes and pursuing my own goals. *Edward*: I agree, finding that balance is crucial. I've learned that being honest about my feelings and needs, even if it causes some tension, is important for my mental well-being. It's a journey, for sure. *Henry*: Absolutely, being honest is essential. It's about asserting independence while still showing respect. It's a delicate dance, but necessary for personal growth and maintaining a healthy relationship with parents. *Edward*: I've found that navigating those differences in expectations has helped me grow and understand myself better. It's about finding my own path while appreciating where they're coming from. It's a process of self-discovery and resilience.

Table 18: Examples of dialogues used in case studies

Domain	Facet	Low Marker	High Marker
EXT	E1 - Friendliness	unfriendly	friendly
EXT	E2 - Gregariousness	introverted	extraverted
EXT	E2 - Gregariousness	silent	talkative
EXT	E3 - Assertiveness	timid	bold
EXT	E3 - Assertiveness	unassertive	assertive
EXT	E4 - Activity Level	inactive	active
EXT	E5 - Excitement-Seeking	unenergetic	energetic
EXT	E5 - Excitement-Seeking	unadventurous	adventurous and daring
EXT	E6 - Cheerfulness	gloomy	cheerful
AGR	A1 - Trust	distrustful	trustful
AGR	A2 - Morality	immoral	moral
AGR	A2 - Morality	dishonest	honest
AGR	A3 - Altruism	unkind	kind
AGR	A3 - Altruism	stingy	generous
AGR	A3 - Altruism	unaltruistic	altruistic
AGR	A4 - Cooperation	uncooperative	cooperative
AGR	A5 - Modesty	self-important	humble
AGR	A6 - Sympathy	unsympathetic	sympathetic
AGR	AGR	selfish	unselfish
AGR	AGR	disagreeable	agreeable
CON	C1 - Self-Efficacy	unsure	self-efficacious
CON	C2 - Orderliness	messy	orderly
CON	C3 - Dutifulness	irresponsible	responsible
CON	C4 - Achievement-Striving	lazy	hardworking
CON	C5 - Self-Discipline	undisciplined	self-disciplined
CON	C6 - Cautiousness	impractical	practical
CON	C6 - Cautiousness	extravagant	thrifty
CON	CON	disorganized	organized
CON	CON	negligent	conscientious
CON	CON	careless	thorough
NEU	N1 - Anxiety	relaxed	tense
NEU	N1 - Anxiety	at ease	nervous
NEU	N1 - Anxiety	easygoing	anxious
NEU	N2 - Anger	calm	angry
NEU	N2 - Anger	patient	irritable
NEU	N3 - Depression	happy	depressed
NEU	N4 - Self-Consciousness	unselfconscious	self-conscious
NEU	N5 - Immoderation	level-headed	impulsive
NEU	N6 - Vulnerability	contented	discontented
NEU	N6 - Vulnerability	emotionally stable	emotionally unstable
OPE	O1 - Imagination	unimaginative	imaginative
OPE	O2 - Artistic Interests	uncreative	creative
OPE	O2 - Artistic Interests	artistically unappreciative	artistically appreciative
OPE	O2 - Artistic Interests	unaesthetic	aesthetic
OPE	O3 - Emotionality	unreflective	reflective
OPE	O3 - Emotionality	emotionally closed	emotionally aware
OPE	O4 - Adventurousness	uninquisitive	curious
OPE	O4 - Adventurousness	predictable	spontaneous
OPE	O5 - Intellect	unintelligent	intelligent
OPE	O5 - Intellect	unanalytical	analytical
OPE	O5 - Intellect	unsophisticated	sophisticated
OPE	O6 - Liberalism	socially conservative	socially progressive

Table 19: The full list of personality descriptors, adapted from Safdari et al. (2023).