Rethinking Role-Playing Evaluation: Anonymous Benchmarking and A Systematic Study of Personality Effects

Ji-Lun Peng[⋆†] **Yun-Nung Chen**[⋆]
[⋆]National Taiwan University, Taipei, Taiwan
[†]Academia Sinica, Taipei, Taiwan
r13946006@ntu.edu.tw y.v.chen@ieee.org

Abstract

Large Language Models (LLMs) have shown remarkable potential in developing role-playing agents (RPAs). However, current evaluation frameworks rely heavily on well-known fictional characters, raising a critical concern: models may be leveraging their internal training memory of these characters rather than demonstrating role-playing capabilities. This reliance often leads to significant performance degradation when RPAs encounter unseen or out-of-distribution personas. To address this, we propose a more rigorous evaluation protocol designed to decouple role-playing proficiency from character recognition. Our experiments across multiple benchmarks demonstrate that *anonymizing* characters degrades performance, confirming that name exposure provides implicit cues that mask a model's true capability. To mitigate this, we investigate diverse personality augmentation as a method to enhance role fidelity in anonymous settings. We systematically analyze the impact of various personality-description methods on agent behavior and consistency. Our results show that incorporating personality information consistently improves RPA performance. This work establishes a more equitable evaluation standard and validates a scalable, personality-enhanced framework for constructing robust RPAs.[1]

1 Introduction

LLMs have demonstrated strong performance across diverse tasks (Peng et al., 2024). A recent and significant application involves their use as RPAs, where they simulate specific professions (Wang et al., 2023) and characters (Shao et al., 2023). This approach enhances the human-likeness of interactions, with character simulation being a major application area (Tseng et al., 2024; Chen et al., 2024; He et al., 2025; Wu et al., 2025).

Current research on role-playing agents primarily focuses on simulating fictional characters, such as those from novels and films, prompting language models to generate text consistent with a predefined character style (Yu et al., 2025; Qin et al., 2025; Zhang et al., 2025c; Wang et al., 2025a). This line of work includes the construction of various benchmark datasets for evaluating role-playing capabilities (Wang et al., 2025a; Dai et al., 2024; Wu et al., 2025), as well as methodological advancements designed to enhance LLMs' role-playing performance (Gao et al., 2025; Yang et al., 2025; Ye et al., 2025; Liu et al., 2025). Enabling LLMs to simulate fictional characters has practical applications in avatar-based interactions[2] and interactive gaming environments (Zhang et al., 2025a), thereby extending virtual characters beyond their original narrative contexts.

In addition, several studies have explored the simulation of real individuals, including tasks such as survey completion (Park et al., 2024), social media post generation (Huang, 2024), and message response imitation (Shi et al., 2025). However, evaluation methodologies for LLMs impersonating lesser-known real humans remain limited, largely due to the difficulty of collecting high-quality, large-scale data about lesser-known real individuals. Moreover, well-known fictional characters are often well-represented in LLM pretraining corpora, granting models substantial prior knowledge about their personas. Consequently, high proficiency in role-playing these famous figures does not necessarily generalize to the simulation of lesser-known individuals, as models are unfamiliar with the these people.

To address this imbalance in prior knowledge, we propose an anonymized role-playing evaluation method in which character identities are concealed. Under this setting, LLMs must rely solely on the

[1]Code: https://github.com/MiuLab/RolePlayBench

[2]https://character.ai

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 205–218
August 2–5, 2026.

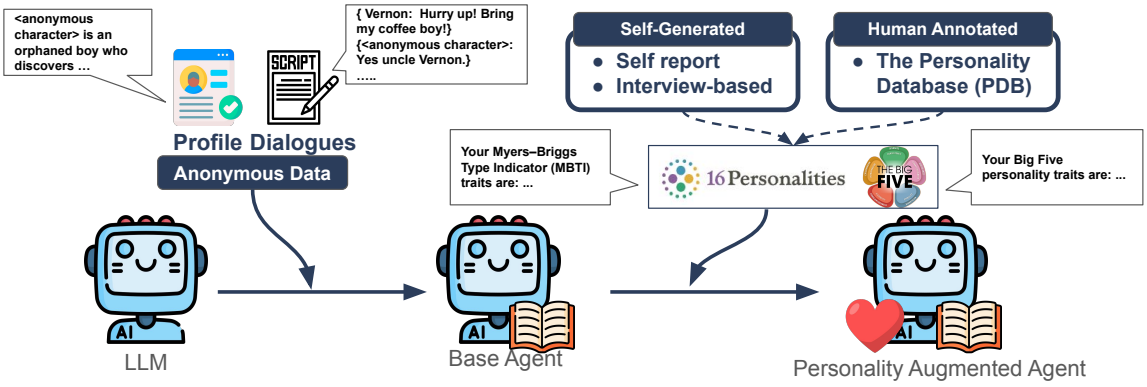

Figure 1: The framework of constructing a personality augmented role-playing agent under anonymous scenarios.

provided character descriptions rather than leveraging pretraining-based prior knowledge. This design mitigates confounding effects stemming from memorized character information and enables a more rigorous assessment of role-playing performance. Importantly, it also facilitates a more principled extrapolation from benchmark results to real-world human impersonation scenarios. An overview of our research framework is illustrated in Figure 1.

While anonymization ensures rigorous evaluation, it also challenges the LLM's ability to maintain character fidelity without prior knowledge. To address this gap, we investigate whether incorporating personality information can bolster role-playing performance under anonymized conditions. Since personality serves as a structured descriptor of human behavioral tendencies (McCrae and Costa Jr, 1997), it provides a precise mechanism for guiding role-play in the absence of character names. Drawing on prior literature (Wang et al., 2024b), we compare multiple approaches to acquiring character personality information, including human-annotated data from The Personality Database (PDB)[3] and model-generated personalities.

Our experiments demonstrate that incorporating personality information consistently improves role-playing performance. Notably, personality self-generated by the model achieve performance comparable to those derived from human annotations. These gains are robust across different model architectures and multiple benchmark datasets.

Importantly, whereas much of the prior work reports results solely on self-constructed datasets (Zhang et al., 2025b; Gao et al., 2025; Liu et al., 2024), our study emphasizes cross-dataset validation. By evaluating across diverse benchmarks and experimental conditions, we show that anonymization significantly reduces model performance in fictional character role-playing, while personality augmentation consistently improves role fidelity. This cross-benchmark consistency strengthens the empirical validity of our findings. The contributions of this paper are as follows:

- We propose a new evaluation setting for role-play: anonymizing character names in the dataset so that the model must rely solely on the provided role descriptions in the prompt. Experimental results show that anonymization degrades previously reported role-play performance, suggesting that character names carry significant implicit information for LLMs. Therefore, anonymized evaluation offers a fairer and more generalizable assessment.
- We investigate whether personality augmentation enhances LLM-based role-playing by enabling the model to generate personality traits from character information. Across multiple benchmark datasets, we show that incorporating personality guidance consistently improves role-playing performance.
- We further compare personalities generated by the model with human-annotated data sourced from PDB. Experimental results across multiple benchmarks indicate that self-generated personalities achieve performance comparable to human-annotated ones, suggesting that effective personality guidance can be obtained without relying on external annotation resources.

2 Anonymous Role-Playing Evaluation

RPAs leverage LLMs to simulate a character's knowledge (Lu et al., 2024), speaking style (Wang

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

Table 1: Datasets for role-playing evaluation. "Char." denotes Charactereval, while "RoleAg." denotes RoleAgentBench. Numbers in parentheses indicate the number of characters available on PDB.

Dataset	Task	#Ques.	#Char. (PDB)
Char. (CH)	Resp. Gen.	8,032	77 (43)
RoleAg. (EN & CH)	General	1,005	54 (39)
	Summary	462	54 (39)

et al., 2024a), and behavior (Park et al., 2023). These systems can be categorized by their **simulation target** and **simulation scale** (Chen et al., 2025). Simulation Target refers to whether the RPA simulate an individual (Xu et al., 2024; Liu et al., 2024; Wang et al., 2024a) or a group (Suzgun and Kalai, 2024; Park et al., 2023). Simulation Scale refers to the number of targets simulated in a given context. This study focuses on RPAs with an individual target and a single agent to simulate character persona, where the target is an existing character from literature, TV shows, or animation.

Two main approaches are used to construct character personas: (1) **Supervised fine-tuning (SFT)**, where models are trained on character-specific Q&A pairs (Wang et al., 2024a; Shao et al., 2023; Wang et al., 2025b; Yu et al., 2024), even enabling smaller models to capture characters' style. (2) **In-context learning (ICL)**, which conditions models with prompts containing character profiles (Tu et al., 2024) or memory from source texts (Li et al., 2023; Liu et al., 2024; Gao et al., 2025). Many recent studies adopt ICL for its flexibility in defining roles and inputs, whereas SFT-based RPAs are limited to specific training data (Wang et al., 2024a) and generalize poorly across roles or benchmarks. We therefore choose ICL, which allows personality information to be directly incorporated into RPAs.

In the first phase of this study, we assess whether LLMs memorize characters through name exposure. We evaluate the effect of anonymizing character names and show that removing names reduces performance. Moreover, under this setting, evaluation results obtained from benchmarks that use fictional characters as role-playing targets can be more reliably generalized to other impersonation scenarios. These include cases in which the target persona is not represented in the model's pretraining data, such as lesser-known real humans or situations where the role-playing target is a specific social group.

2.1 Anonymizing Process

To create anonymous scenarios during experiments, we replaced all simulated character names in the prompt with the token `<anonymous character>`. For example, when simulating Harry Potter, if the profile reads "*Harry Potter's friends include Ron and Hermione.*," it becomes "`<anonymous character>'s` *friends include Ron and Hermione.*". Since *Ron* and *Hermione* were not the simulation target of the RPA, they remained un-anonymized. The LLM was thus required to role-play without access to the simulated character's name, relying solely on information in the prompt. After the RPA generated its responses, we restored the original character names to avoid any impact on evaluation results.

2.2 Datasets and Evaluation Methods

We evaluate our approach using two RPA benchmarks that cover diverse question formats, evaluation methods, and languages: **CharacterEval** (Tu et al., 2024) and **RoleAgentBench** (Liu et al., 2024). The statistic of two benchmarks is shown in Table 1.

CharacterEval is a Chinese role-playing dialogue dataset designed to evaluate model-generated responses conditioned on a character's profile and dialogue history, covering 3 dimensions and 11 objectives: **Conversational Ability** (Fluency, Coherency, Consistency), **Role-Playing Attractiveness** (Human-Likeness, Communication Skills, Expression Diversity, Empathy), and **Character Consistency** (Knowledge-Exposure, Knowledge-Accuracy, Knowledge Hallucination, Persona-Behavior, Persona-Utterance). The responses are scored on a continuous scale from 1 to 5 using a reward model trained in the original study.

RoleAgentBench includes roles from both Chinese and English scripts and provides each character's profile along with dialogue history involving other characters. It consists of **General Response** and **Summarization** two tasks, which measure general communication skills and memory system effectiveness respectively[4]. Following the original paper, we adopt the LLM-as-a-judge evaluation, where the LLM uses the character profile and the dataset's gold answer to select the output that best

[4]Since **Self-Knowledge** and **Reaction**, the other two tasks in RoleAgentBench, are less relevant to the role-playing capabilities we focus on, we exclude these two tasks from our experiments.

Table 2: CharacterEval scores in Character Consistency dimension of three models: original vs. anonymous settings. † denotes a statistically significant deterioration ($p < 0.05$).

	Know-Exposure	Know-Accuracy	Know-Hallucination	Persona-Behavior	Persona-Utterance	Avg.
`gemini-2.0-flash`						
Original	2.667	3.187	3.286	3.787	3.323	3.251
Anonymous	2.656 (-)	3.178 (-)	3.250 (-)	3.783 (-)	3.277 (-)	3.229 (-)
`llama-3.1-405B-instruct`						
Original	2.037	3.013	3.030	3.266	3.190	2.907
Anonymous	2.097	3.029	3.046	3.205 (-)	3.187 (-)	2.913
`gpt-4o`						
Original	2.509	3.029	3.021	2.839	3.024	2.884
Anonymous	2.461 (-)	2.986 (-)	2.969 (-)	2.607† (-)	2.938† (-)	2.792† (-)

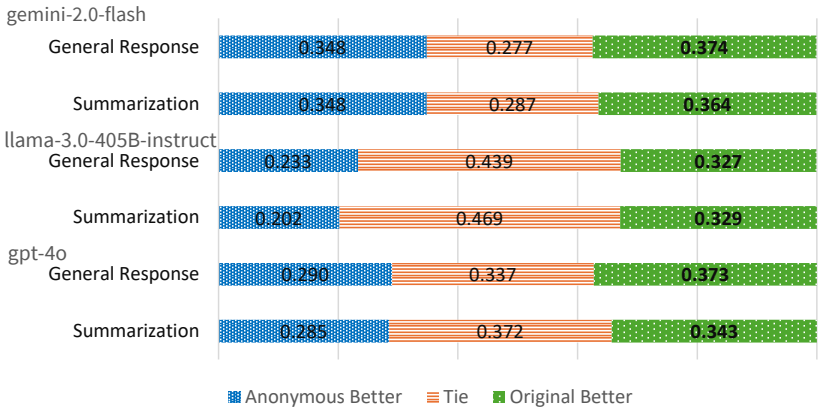

Figure 2: Paired response evaluation on RoleAgentBench: win rates of original vs. anonymous settings.

matches the character's traits and computes win rates (Dubois et al., 2025). We focus on the publicly available subset of the benchmark.

2.3 Models

Previous studies have shown that older models like GPT-3.5 perform significantly worse than GPT-4 in role-playing tasks (Suzgun and Kalai, 2024). Since this study focuses on personality effects and personality is more abstract than factual information, LLMs must possess sufficient role-playing capability to reflect personality input impact. Therefore, we conduct experiments using recent API-based models: `gemini-2.0-flash` and `gpt-4o`. We also include the open-source model `llama-3.1-405B-instruct` to compare performance between open-source and closed-source models.

For RoleAgentBench evaluation, we use `gpt-4o-mini` as the evaluator model. The evaluation is conducted in a pairwise manner, where the prompt order is swapped and run twice. A response is considered a win only if it wins both runs; one win and one loss results in a tie, and two losses count as a loss. The pairwise evaluation prompt template, adapted from (Liu et al., 2024), is shown in Figure 6.

2.4 Anonymous Evaluation Results

Table 2 presents experimental results comparing the original and anonymous versions of the CharacterEval dataset. We focus on the *Character Consistency* dimension, as it best reflects the model's ability to portray specific fictional characters with high fidelity. Overall, responses in the original setting receive higher scores than those in the anonymous setting, suggesting that the reward model finds them more aligned with the intended character traits. Comprehensive experimental results are detailed in Table 7 of Appendix, which reveals that the Original setting consistently outperforms the Anonymous setting across other objectives as well.

Similar trends are observed in RoleAgentBench, as shown in Figure 2. Under LLM-based pairwise evaluation, the original responses exhibit consistently higher win rates than the anonymized responses, indicating better alignment with the ground-truth role-playing behavior.

This performance gap suggests that results reported in previous work may have benefited from models' memorization of character names. Employing an anonymized evaluation setting enables a fairer comparison of models' ability to perform role-playing based solely on the information provided in the prompt, thereby making the evaluation results more generalizable to scenarios beyond fictional character role-playing. Therefore, we conduct all subsequent experiments under this setting.

3 Personality-Augmented RPAs

3.1 Personality of Role-Playing Agents

In psychology perspective, personality are often assessed through standardized questionnaires. Users indicate their level of agreement with each item, and a final score is computed (John et al., 1991; Pittenger, 1993). According to Wang et al. (2024b), there are three main approaches for obtaining the personality of RPAs: (1) **Self-report**, where the RPA is prompted with the scale items and possible responses, selecting an option directly to produce a score; (2) **Interview-based**, which reformulates scale items into open-ended questions, allowing the RPA to respond freely, after which an evaluator LLM assigns personality scores based on the responses; (3) **Crowdsourced data**, which retrieves personality from the PDB website based on crowdsourced votes for the target character.

Prior research has explored methods for inferring personality in RPAs (Wang et al., 2024b; Serapio-García et al., 2023). However, relatively little work has examined whether incorporating personality information can improve role-playing performance. A few recent studies (Tu et al., 2023; Huang, 2024) touch on personality-enhanced role-playing, but their objectives differ from ours. Specifically, Tu et al. (2023) focused on increasing the diversity of LLM-generated characters through personality information, whereas Huang (2024) inferred user personality from social media posts, which has relatively low alignment with personality results directly derived from standardized assessment. Moreover, leveraging the personality information of fictional characters to improve role-playing fidelity has not yet been systematically studied.

In this work, we address this gap by investigating whether personality information derived from these various acquisition methods consistently improves role-playing performance. We focus on evaluating the comparability between model-generated personalities (Self-report and Interview-based) and human-annotated data (PDB). By incorporating each source into RPAs, we systematically analyze whether self-generated personality information can serve as a reliable alternative to human annotations across multiple models and benchmarks.

3.2 Personality Enhancement

In this study, we use the 16Personalities test[5] to assess the personality of RPAs. This scale is based on the Myers-Briggs Type Indicator (MBTI) (Pittenger, 1993) and is widely adopted in both academic and public settings. It is a type-based test consisting of 60 items and yields a result composed of four dimensions. The output is a four-letter personality type (e.g., INTJ).

Following the implementation (Tu et al., 2024), we obtain each character's personality type using both the self-report and interview-based methods. For the interview-based approach, we employ `gemini-2.0-flash` as the evaluator to infer personality types. Additionally, we retrieve personality types from the PDB when available. We then provide the four-letter personality type and the definitions of each category as additional input to the RPA during role-playing. The prompt design is structured according to prior work (Wang et al., 2024a), as detailed in Figure 5 of Appendix.

3.3 Personality-Augmented RPA Results

To fairly compare the three personality sources, our evaluation focuses on characters available in the PDB, and we report the mean score over three runs to ensure reliability. The results show that, across both benchmarks, augmenting the prompt with personality information consistently improves performance across most dimensions for all models. This enhancement is evidenced by higher absolute scores (Table 3) and win rates that significantly exceed loss rates (Figure 3), indicating that personality-augmented RPAs generate behaviors more aligned with their target characters. The comprehensive experimental results for each objective of CharacterEval are presented in Table 8.

Notably, all three personality sources enhance RPA performance to a similar degree. This finding suggests that self-generated personality traits are a practical and effective alternative to PDB data, as they offer broader applicability to any character while maintaining strong performance.

[5] https://www.16personalities.com/

Table 3: CharacterEval with MBTI or Big Five integration across three dimensions: *Conversational Ability, Role-playing Attractiveness*, and *Character Consistency*. † denotes a significant improvement over the original condition ($p < 0.05$); **Bold** fonts indicate the best performance and underline fonts indicate the second best one.

		gemini-2.0-flash			llama-3.1-405B-inst.			gpt-4o		
		Conv.	**Attr.**	**Char.**	**Conv.**	**Attr.**	**Char.**	**Conv.**	**Attr.**	**Char.**
MBTI	RPA original	3.871	3.468	3.254	3.829	3.079	**2.898**	3.542	2.969	2.801
	+ SelfReport	**3.924**†	3.489	3.282	3.846	3.082	2.887	3.538	**3.176**†	**2.975**†
	+ Interview	3.919	3.499	3.262	**3.858**	**3.104**	2.884	3.527	3.168†	2.960†
	+ PDB	3.920	**3.512**	**3.289**	**3.858**	3.102	2.887	**3.556**†	3.052†	2.875†
Big 5	RPA original	3.951	3.549	3.372	**3.972**	3.131	**3.064**	3.718	3.017	2.873
	+ SelfReport	3.997	3.509	**3.427**	3.971	**3.203**	**3.064**	3.701	**3.086**	2.955
	+ Interview	3.992	3.491	3.376	3.970	3.175	3.012	3.723	3.082	**2.972**
	+ PDB	**4.057**	**3.557**	3.408	3.961	3.197	3.046	**3.736**	3.056	2.932

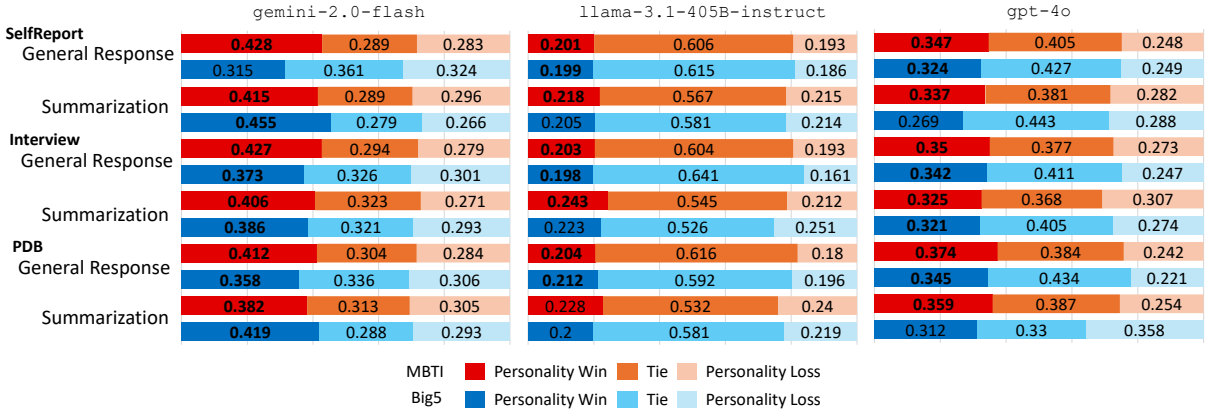

Figure 3: Pairwise comparison on RoleAgentBench: personality integration vs. original condition.

Our experiments further demonstrate that even when the model is unaware of the character's identity, self-generated personality information can significantly improve role-playing performance. These improvements are consistently observed across multiple models and benchmark datasets, highlighting the flexibility of our approach and its ability to robustly enhance RPA performance under diverse settings. Importantly, this suggests that our method is applicable to scenarios in which the model lacks prior familiarity with the target persona, such as impersonating real individuals who are not represented in its pretraining data.

4 Discussions and Analysis

4.1 Generalization to Big Five Personality Inventory

To further verify the effectiveness of personality information in enhancing RPA performance, we adopt the **Big Five Inventory (BFI)** (John et al., 1991), a widely used personality assessment instrument. The Big Five personality are measured as continuous scores across five dimensions, reflecting individual personality characteristics and allowing for more fine-grained distinctions. In contrast, while the MBTI framework is less sensitive to subtle variations, it is generally easier to interpret. Therefore, we evaluate both frameworks to show that widely used personality models, regardless of their structural differences, consistently improve RPA performance.

In our experiments, we provide RPAs with the BFI scores for each character to construct personality-augmented agents. To ensure a fair comparison, we evaluate only characters with available PDB personality annotations. However, Big Five data is sparse, covering only 9 characters in CharacterEval and 25 in RoleAgentBench. This limited sample size accounts for the differing "Original" baseline results across the two benchmarks.

The results, presented in Table 3 and Figure 3,

Table 4: Distribution of characters with distinctive personality types across different experimental conditions.

Dataset	Self-report	Interview	PDB
CharacterEval	22	20	37
RoleAgentBench	3	14	33

demonstrate that BFI also yields performance gains, suggesting that the benefits of incorporating personality information are robust across different frameworks. However, MBTI outperforms BFI in terms of enhancement magnitude. We hypothesize that this disparity stems from the prevalence of MBTI in public online discourse; for instance, the PDB contains more MBTI-related character analyses than BFI. Consequently, LLMs may have developed a superior ability to map MBTI traits to specific character behaviors during. The underlying mechanisms behind this performance disparity warrant further investigation. The comprehensive experimental results for each objective of CharacterEval are presented in Table 9 of Appendix.

4.2 Greater Improvement from Stronger Personality Traits

To further investigate the influence of personality on RPAs, we identified characters with distinctive personality profiles to evaluate whether such traits correlate with greater performance improvements. We define a distinctive personality profile as having MBTI scores either above 60% or below 40% across all four dimensions; only characters meeting this criterion were included in the analysis.

We conducted a granular evaluation of `gemini-2.0-flash`, as it not only achieved the highest average score on CharacterEval but also demonstrated the most significant win-rate margin in the personality-augmented experiments on RoleAgentBench. We hypothesize that `gemini-2.0-flash` is better equipped to leverage personality information for role-playing; consequently, we expect the presence of a distinctive personality to yield even more substantial performance gains. The distribution of these characters across each condition is detailed in Table 4. We specifically focus on the performance gains localized to this character subset, contrasting their improvements with the performance of the full dataset to highlight the efficacy of personality augmentation.

Table 5 reports the performance differences when considering only roles with strong personality traits, where positive values imply that stronger personality information enhances role-playing capabilities more. We observe that characters with distinctive personality types generally yield greater performance gains. These results suggest that when a character exhibits distinct personality tendencies, RPAs can more effectively leverage personality cues to generate behaviors that align more closely with the original character, thereby providing clearer evidence of the importance of personality information in LLM-based role-playing.

Table 5: Additional performance improvement attributed to distinctive personalities with `gemini-2.0-flash`. RoleAgentBench values denote win rate minus loss rate.

	CharacterEval			RoleAgentBench	
	Conv.	Attr.	Char.	Gen. Resp.	Sum.
SelfR.	-.013	**+.021**	-.012	**+.076**	**+.122**
Inter.	**+.138**	-.088	**+.067**	**+.055**	**+.020**
PDB	**+.152**	**+.291**	**+.162**	**+.101**	-.149

4.3 Qualitative Analysis

In addition to benchmark scores, we conducted a qualitative analysis of RPAs' responses across different settings. Sample responses are shown in Table 6. Prior to incorporating personality information, responses in the Original setting showed more character-specific tone and emotion compared to the Anonymized setting, likely due to the LLM leveraging prior knowledge when aware of the character's name. After adding personality information, the responses better reflected the character's tone, emotional state, and included actions, resulting in more vivid and expressive outputs. This suggests that even without access to the character's name, personality information can provide meaningful cues that help the RPA generate more lifelike and engaging role-play behavior.

4.4 Human Evaluation

To further validate these findings, we conducted a human evaluation on RoleAgentBench. We used the generation results of `gemini-2.0-flash` and focused on Harry Potter data to ensure evaluator familiarity. Two psychology graduate students with expertise in the Harry Potter series performed pairwise comparisons. A response was deemed a "win" if both evaluators preferred it, a "loss" if neither did, and a "tie" if they disagreed.

The evaluation covering role-playing for 5 char-

Table 6: Sample responses with different methods.

Question (from McGonagall): Mr. Malfoy, do you recall your behavior during our first flying lesson and how you provoked Mr. Potter, which led to him being discovered as a potential Quidditch player for Gryffindor?
Original RPA (Malfoy): (Scoffs) Of course I remember. Longbottom was being an idiot, as usual. It was hilarious. And Potter? He should be thanking me! I inadvertently gave him a chance to show off. Without me, he'd just be another nobody. Besides, it's not my fault Potter can't resist a challenge. He practically begged me to goad him.
Anonymized RPA: Yes, I recall Longbottom making a fool of himself. And I certainly remember showing Longbottom's Remembrall to Potter, wondering if the fat lump could even reach it on the roof. What's it to you? Potter being a Seeker hardly seems like *my* fault. He's still a mudblood-loving Gryffindor, isn't he?
+ SelfReport Personality (INTJ): (Scoffs, crossing his arms) Recall it? Of course, I recall it. Potter showing off, thinking he's so special. Longbottom was being an idiot, dropping that stupid Remembrall. I was just having a bit of fun. Potter deserved to be knocked down a peg or two. He got lucky, that's all. And if he hadn't been so busy trying to be a hero, maybe the stupid lump wouldn't have made such a fool of himself.

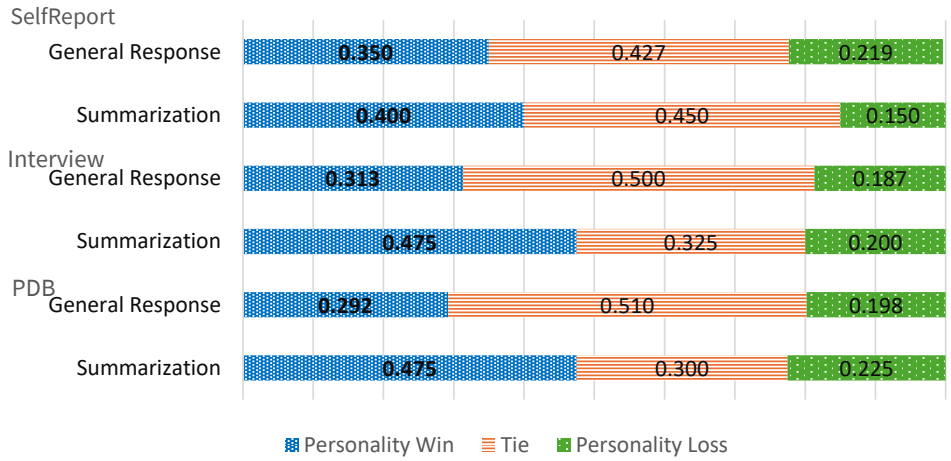

Figure 4: Human pairwise comparison against the original condition.

acters from the Harry Potter series. Each evaluator assessed 120 and 288 pairwise comparisons for the summary and general response tasks, respectively. In terms of inter-rater agreement, Cohen's kappa was 0.415 for the summary task and 0.308 for the general response task, corresponding to fair to moderate agreement according (McHugh, 2012).

The results, presented in Figure 4, confirm that personality-augmented RPAs outperform the original condition under human judgment, aligning with the findings from our LLM-based evaluations.

5 Conclusion

This work introduces an anonymous role-play setting and proposes the incorporation of personality information into RPAs to enhance performance. We compare multiple personality sources to assess their impact. Our experiments are conducted across multiple benchmarks, languages, and model architectures, yielding robust and consistent results.

Our findings show that the anonymous setting generally reduces RPA performance, highlighting LLMs' sensitivity to character names. This suggests that conventional role-play evaluation may partially rely on prior knowledge of characters. By contrast, anonymous evaluation provides a fairer assessment framework and produces results that are more generalizable to scenarios beyond fictional character role-playing.

We further observe that incorporating personality information consistently improves role-playing performance. Notably, self-generated personality information achieves performance comparable to human annotations, demonstrating a flexible and scalable approach that can be applied across diverse settings.

Limitations

This study has several limitations that open avenues for future research. First, although our anonymization strategy effectively reduces name cues, models may still infer character identities from other de-

scriptive signals, such as the names of interlocutors, nicknames mentioned in dialogue, or distinctive and widely recognized character traits. Future work could develop more comprehensive anonymization techniques to further mitigate potential confounding effects.

Second, although most models in our study exhibit performance gains from personality augmentation, the magnitude of improvement varies across model architectures. Further experiments on a broader range of models are needed to better understand what factors contribute to these differences. Identifying the mechanisms underlying such variability remains an important direction for future research.

Finally, while our experiments confirm that both personality frameworks enhance RPA performance, it remains unclear why the magnitude of improvement varies across frameworks, or how these disparate information sources can be optimally combined to achieve peak performance. Addressing these questions represents an important direction for future research.

Acknowledgements

This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grant 112-2223-E-002-012-MY5. We thank the National Center for High performance Computing of National Institutes of Applied Research (NIAR) in Taiwan for providing computational and storage resources.

References

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Yanfang Ye, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. *Preprint*, arXiv:2502.13012.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Zhenpeng Gao, Xiaofen Xin, and Xiangmin Xu. 2025. Tailorrpa: A retrieval-based framework for eliciting personalized and coherent role-playing agents in general domain. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5381–5412.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A novel configurable role-playing llm with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052.

Yuxuan Huang. 2024. Orca: Enhancing role-playing abilities of large language models by integrating personality traits. *arXiv preprint arXiv:2411.10006*.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Cheng Liu, Yifei Lu, Fanghua Ye, Jian Li, Xingyu Chen, Feiliang Ren, Zhaopeng Tu, and Xiaolong Li. 2025. Cogdual: Enhancing dual cognition of llms via reinforcement learning with implicit rule-based rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27295–27324.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Noah Wang, Jian Yang, JiakaiWang, Hongcheng Guo, Z.Y. Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. 2024. RoleAgent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist*, 52(5):509.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*.

David J Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of educational research*, 63(4):467–488.

Haiming Qin, Jiwei Zhang, Wei Zhang, KeZhong Lu, Mingyang Zhou, Hao Liao, and Rui Mao. 2025. R-char: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26984–27002.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *Preprint*, arXiv:2310.10158.

Quan Shi, Carlos E Jimenez, Stephen Dong, Brian Seo, Caden Yao, Adam Kelch, and Karthik Narasimhan. 2025. Impersona: Evaluating individual level lm impersonation. *arXiv preprint arXiv:2504.04332*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

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, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14743–14777.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025b. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. 2025. Raiden benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11086–11106.

Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024. Mindecho: Role-playing language agents for key opinion leaders. *arXiv preprint arXiv:2407.05305*.

Shihao Yang, Zhicong Lu, Yong Yang, Bo Lv, Yang Shen, and Nayu Liu. 2025. Hycora: Hyper-contrastive role-adaptive learning for role-playing. *arXiv preprint arXiv:2511.08017*.

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. *Preprint*.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2024. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. *Preprint*, arXiv:2408.10903.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, and 1 others. 2025a. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26318–26331.

Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025b. Roleplot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354.

Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. 2025c. Revealing and mitigating the challenge of detecting character knowledge errors in llm role-playing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33267–33290.

A Prompts for Constructing RPA and Pairwise Evaluation

Figure 5 depicts the prompt configuration for the personality-augmented RPA. The underlying architecture adapts to varying benchmark instructions, with comprehensive details provided in the codebase. Within this framework, personality information is operationalized using four-letter MBTI types and five-dimensional BFI scores. For the non-augmented condition, the prompts are simplified by truncating all content prior to the string: "Your MBTI traits are: personality info." Furthermore, the prompt of LLM-based pairwise evaluation protocol for RoleAgentBench is detailed in Figure 6.

B Full results of experiments

Table 7, Table 8, and Table 9 report the dimension-wise performance on CharacterEval, providing extended versions of the results summarized in Table 2 and Table 3. As illustrated in Table 7, the anonymous setting generally leads to a performance degradation across most objectives. This suggests that the absence of explicit character names negatively impacts the model's ability to portray well-known fictional characters effectively.

Furthermore, Table 8 and Table 9 demonstrate that incorporating personality information enhances role-playing performance across the majority of objectives, regardless of the personality source. Notably, the most significant improvements are observed in **Human-likeness, Expression Diversity, and Persona-Behavior**. These results indicate that personality profiles enable RPAs to generate more diverse, human-like, and behaviorally consistent responses.

System prompt:
{character's profile}

User prompt:
Myers–Briggs Type Indicator (MBTI) is a self-report questionnaire. The test assigns a binary value to each of four categories: introversion or extraversion, sensing or intuition, thinking or feeling, and judging or perceiving. Extraverts (also often spelled extroverts) are outward-turning and tend to be action-oriented, enjoy more frequent social interaction, and feel energized after spending time with other people. Introverts are inward-turning and tend to be thought-oriented, enjoy deep and meaningful social interactions, and feel recharged after spending time alone.. People who prefer sensing tend to pay a great deal of attention to reality. Those who prefer intuition pay more attention to things like patterns and impressions. People who prefer thinking place a greater emphasis on facts and objective data. Those who prefer feeling are more likely to consider people and emotions when arriving at a conclusion. Those who lean toward judging prefer structure and firm decisions. People who lean toward perceiving are more open, flexible, and adaptable.
Your MBTI traits are: {personality info}.
Please anwser the following questions based on character's information.
The following is the dialogue record of the role:
{dialogue history between the character and other agents}
{task instruction and question}

Figure 5: Prompt template for personality-augmented RPA.

System prompt:
You are a role-playing performance comparison assistant. You should rank the conditions based on the role characteristics and text quality of their responses. The rankings are then output using Python dictionaries and lists.

User prompt:
The conditions below are to play the role of {character name}. The role description of {character name} is {character's profile}. I need to rank the following conditions based on the two criteria and the reference answer below:
1. Which one has more pronounced role speaking style, and speaks more in line with the role description. The more distinctive the speaking style, the better.
2. Which one's output contains more knowledge and memories related to the role; the richer, the better. (If the question contains reference answers, then the role-specific knowledge and memories are based on the reference answer.)
The question provided to each condition is: {benchmark question}
The reference answer of the question is: {benchmark answer}
The respective answers from the conditions to this question are:
[{"condition": "condition 1", "answer": {answer 1}}, {"condition": "condition 2", "answer": {answer 2}}]
Now, based on the above two criteria and the reference answer, please rank the conditions. Avoid any positional biases and ensure that the order in which the responses are presented does not influence your decision. Do not favor certain model names. Then, use a list containing the condition's name, its rank, and the reason for its ranking to return the results, i.e., please ensure to use the following format to return the results:
[{"condition": <condition-name>, "reason": <rank-reason>, "rank": <condition-rank>},
{"condition": <condition-name>, "reason": <rank-reason>, "rank": <condition-rank>}]
Your answer must be a valid Python list of dictionaries to ensure I can directly parse it using Python. Do not include any extraneous content! Please provide a ranking that is as accurate as possible and aligns with the intuition of most people.

Figure 6: Prompt template for RoleAgentBench pair-wise evaluation.

Table 7: Full comparison between Original and Anonymous settings on CharacterEval. Underlined values indicate that the anonymous setting resulted in a lower score compared to the original, and † denotes a statistically significant deterioration ($p < 0.05$). The colors Cyan, Teal, and Purple represent the Conversational Ability, Role-Playing Attractiveness, and Character Consistency dimensions, respectively.

	gemini-2.0-flash		llama-3.1-405B-instruct		gpt-4o	
	Original	**Anonymous**	**Original**	**Anonymous**	**Original**	**Anonymous**
Fluency	3.779	3.744	3.713	3.682	3.516	3.455
Coherency	4.089	4.099	4.046	4.032	3.874	3.831
Consistency	3.914	3.883	3.928	3.893	3.495	3.406†
Human-likeness	3.679	3.637	3.743	3.693	3.155	3.099
Communication Skills	3.618	3.588	3.049	3.057	3.571	3.466†
Expression Diversity	3.255	3.245	2.623	2.574	2.236	2.103†
Empathy	3.410	3.399	3.142	3.106	3.306	3.274
Know-Exposure	2.667	2.656	2.037	2.097	2.509	2.461
Know-Accuracy	3.187	3.178	3.013	3.029	3.029	2.986
Know-Hallucination	3.286	3.250	3.030	3.046	3.021	2.969
Persona-Behavior	3.787	3.783	3.266	3.205	2.839	2.607†
Persona-Utterance	3.323	3.277	3.190	3.187	3.024	2.938†

Table 8: Full comparison between w/ MBTI Personality and w/o Personality in anonymous scenario on CharacterEval. "Original" refers to prompts without personality information. Bold values indicate that the score in the w/ personality setting is higher than in the original, and † denotes a statistically significant improvement ($p < 0.05$). The colors Cyan, Teal, and Purple represent the Conversational Ability, Role-Playing Attractiveness, and Character Consistency dimensions, respectively.

	gemini-2.0-flash				llama-3.1-405B-instruct				gpt-4o			
	Original	**SelfReport**	**Interview**	**PDB**	**Original**	**SelfReport**	**Interview**	**PDB**	**Original**	**SelfReport**	**Interview**	**PDB**
Fluency	3.676	**3.749**	**3.754**	**3.749**	3.598	**3.631**	**3.641**	**3.642**	3.418	3.329	3.351	3.411
Coherency	4.084	**4.120**	**4.116**	**4.109**	4.022	4.018	**4.040**	**4.031**	3.803	3.754	3.730	3.793
Consistency	3.852	**3.903**	**3.888**	**3.903**	3.865	**3.890**	**3.892**	**3.901**	3.405	**3.532**†	**3.502**	**3.464**
Human-likeness	3.632	**3.643**	**3.666**	**3.683**	3.692	**3.710**	**3.758**	**3.781**	3.094	**3.271**†	**3.261**†	**3.150**†
Communication Skills	3.618	**3.660**	**3.650**	**3.682**	3.044	**3.054**	**3.073**	**3.045**	3.475	3.404	3.399	3.458
Expression Diversity	3.213	**3.222**	**3.260**	**3.248**	2.465	2.443	2.456	2.457	2.020	**2.938**†	**2.919**†	**2.365**†
Empathy	3.411	**3.430**	**3.419**	**3.453**	3.116	**3.123**	**3.129**	**3.126**	3.286	3.092	3.093	3.235
Know-Exposure	2.759	**2.786**	2.748	**2.771**	2.124	2.046	2.064	2.057	2.493	**2.564**	**2.568**	**2.517**
Know-Accuracy	3.189	**3.195**	**3.199**	**3.203**	3.023	**3.029**	**3.026**	3.008	3.021	2.920	2.913	2.965
Know-Hallucination	3.285	**3.331**	**3.306**	**3.363**†	3.051	**3.090**	3.038	**3.071**	3.010	**3.028**	2.995	**3.024**
Persona-Behavior	3.730	**3.777**	**3.733**	**3.773**	3.117	3.091	3.100	3.106	2.541	**3.337**†	**3.313**†	**2.869**†
Persona-Utterance	3.309	**3.323**	**3.325**	**3.336**	3.175	**3.177**	**3.194**	**3.196**	2.938	**3.023**	**3.010**	**2.998**

Table 9: Full Comparison between w/ BFI Personality and w/o Personality in anonymous scenario on CharacterEval. "Original" refers to prompts without personality information. Bold values indicate that the score in the w/ personality setting is higher than in the original, and † denotes a statistically significant improvement over the original condition ($p < 0.05$). The colors Cyan, Teal, and Purple represent the Conversational Ability, Role-Playing Attractiveness, and Character Consistency dimensions, respectively.

	gemini-2.0-flash				llama-3.1-405B-instruct				gpt-4o			
	Original	**SelfReport**	**Interview**	**PDB**	**Original**	**SelfReport**	**Interview**	**PDB**	**Original**	**SelfReport**	**Interview**	**PDB**
Fluency	3.556	**3.706**	**3.641**	**3.798**	3.687	3.606	3.670	3.556	3.400	3.378	**3.469**	3.350
Coherency	4.230	**4.261**	**4.285**	**4.253**	4.157	**4.216**	**4.163**	**4.205**	4.055	4.045	**4.058**	**4.077**
Consistency	4.067	4.024	4.050	**4.122**	4.072	**4.090**	**4.076**	**4.122**	3.699	3.681	3.643	**3.781**
Human-likeness	3.807	3.696	3.697	3.608	3.787	3.763	3.757	3.763	3.151	**3.167**	**3.250**	**3.195**
Communication Skills	3.811	3.738	3.733	**3.863**	3.265	**3.317**	3.232	**3.268**	3.553	**3.644**	3.528	**3.591**
Expression Diversity	2.926	**2.991**	**2.988**	**3.090**	2.196	**2.303**	**2.297**	**2.338**	1.797	**1.951**	**1.989**	**1.852**
Empathy	3.652	3.613	3.548	**3.665**	3.357	**3.428**	**3.415**	**3.420**	3.569	**3.580**	3.560	**3.587**
Know-Exposure	3.200	**3.273**	3.105	3.184	2.607	2.566	2.529	2.559	2.940	**2.995**	**2.978**	2.921
Know-Accuracy	3.320	**3.352**	**3.365**	**3.399**	3.267	**3.291**	3.195	**3.297**	3.100	**3.132**	**3.144**	3.099
Know-Hallucination	3.364	**3.423**	**3.420**	**3.420**	3.297	3.222	3.191	**3.230**	2.841	**2.993**	**3.035**	**3.037**
Persona-Behavior	3.545	**3.649**	**3.563**	**3.616**	2.828	**2.906**	**2.875**	2.826	2.329	**2.541**	**2.649**	**2.521**
Persona-Utterance	3.433	**3.437**	3.425	3.420	3.321	**3.333**	3.269	3.315	3.154	3.115	3.055	3.087