# Acting Technique: Factors that Influence the Effectiveness of Role-Playing Model

**Anonymous ACL submission**

## Abstract

The goal of role-playing is to ensure that LLMs generate responses that fully embody the specified character, adhering to their background, speaking style, personality, and typical behaviors. Various methods seek to enhance LLMs' ability to mimic characters, such as using specialized prompts and fine-tuning with character dialogues. However, there is currently no research investigating the relationship between LLMs' capabilities and their performance in role-playing. Understanding this relationship could provide insights into selecting the most suitable base model for mimicking a practical character and improving LLMs performance from the outset of the task. In this paper, we investigate three main factors that influence performance in role-playing tasks: the personalities of the LLMs and the given characters, the LLMs' familiarity with the character, and the LLMs' conversational abilities. We conduct experiments with 11 models (including both prompt-based and fine-tuning methods) across 7 characters with varying personalities and backgrounds. The results indicate that the personalities of both the LLMs and the characters are the primary factors affecting performance, while insufficient familiarity leads to poor outcomes. Importantly, we find no evidence that sharing the same personality with a character correlates with improved performance in role-playing.

## 1 Introduction

Role-playing is an inherent ability of LLMs, allowing them to adapt and respond based on various contexts and character traits. To achieve good performance, LLMs need to mimic the character as closely as possible, including the speaking style, tone, vocabulary, logic, and personality traits. LLMs also need to be very familiar with the character to respond to questions related to real-life stories and experiences.

Existing research can be divided into two categories: using specially designed prompts with example dialogues (special prompts) and fine-tuning LLMs. Both methods have their advantages. Special prompts work with little data and do not require fine-tuning LLMs, while fine-tuning methods can learn more and better embody the character. However, there is no research on the influence of LLMs' abilities on performance. All methods focus on achieving the best performance by introducing more role knowledge into LLMs and obtaining higher-quality dialogues for LLMs to mimic. We believe these methods indeed improve performance, but we argue that identifying the factors that have a greater impact on LLM role-playing performance and optimizing them may be more effective than simply providing more high-quality data.

Inspired by the acting technique, we believe that factors within the LLMs themselves influence their role-playing performance in different roles. For example, "The Rock" Johnson is known for many of his movies, but he might not perform as well if asked to play the male lead in an idol drama. This is because Johnson's personality is so different from that of the idol, and Johnson can not understand the psychological transformation of the idol.

With the growing presence of role-playing LLMs, such as RoleLLM, CharacterLLM, CharacterGLM, and Chatharuhi, we aim to explore the key factors that contribute to effective role-playing in these models. Specifically, we seek to understand the relationship between LLM performance and their ability to excel in role-playing tasks. Just as an actor's performance is shaped not only by physical traits but also by their knowledge of the character, personal similarities (such as behavior and personality), and acting skills, we believe these elements apply to LLMs as well. Therefore, we examined the impact of LLMs' capabilities, personality alignment, and role familiarity on their role-playing performance, finding that these factors

affect outcomes to varying degrees. The contributions of this paper are as follows:

- We propose a correlation between LLM capabilities and their role-playing performance and design experiments to validate this relationship.
- We verified that a personality consistent with the original character does not necessarily lead to better performance; however, the absence of the 'N' trait (one of BigFive traits) can result in a higher overall score.
- We rank the factors that influence role-playing performance and provide recommendations for developing effective role-playing models.

## 2 Related Work

This paper investigates the connection between personality traits and the capabilities of LLMs in role-playing tasks. While methods for assessing modeling ability are relatively well established, we focus on introducing relevant studies on LLM personality detection and their application in role-playing.

### 2.1 The Personality of LLMs

Research from Google Mind (Serapio-García et al., 2023) has demonstrated that personality assessments based on the outputs of certain LLMs, under specific prompting conditions, are reliable and valid. An increasing number of studies are focusing on assigning personalities to LLMs, which can generally be categorized into two approaches: questionnaires and text-based responses (Wen et al., 2024).

The questionnaire method (Frisch and Giulianelli, 2024; Serapio-García et al., 2023), commonly used for assessing human personality, presents participants with statements such as "You enjoy freedom," and asks them to respond on a scale from "Very Accurate" to "Very Inaccurate" (Song et al., 2023). To address the challenge that LLMs often refuse to answer questions regarding personal opinions or experiences, researchers have modified these questionnaires. For example, Huang et al. (2023) reformulated the statements in third-person plural, while Sorokovikova et al. (2024) incorporated role descriptions, and Lee et al. (2024) adapted the style of the statements. These studies have successfully assessed the personality traits of LLMs using psychological models such as the Big Five, MBTI, and SD3.

The text-based response method (Karra et al., 2022; Pellert et al., 2024) classifies personality based on LLM-generated responses. Early research typically used zero-shot classification models, but with advancements in psychological tools, more recent work has incorporated features from psychological frameworks such as LIWC (Frisch and Giulianelli, 2024) and psyattention (Zhang et al., 2023) into classification models.

### 2.2 Role-Playing LLMs

Role-playing tasks require LLMs to generate responses in the manner of a specific character. A study in Nature(Shanahan et al., 2023) highlighted that LLMs' dialogues with humans can be considered a form of role-playing, as these models easily adopt the behavior of specific characters. Current research in this area primarily uses three methods: zero-shot, few-shot, and fine-tuning or Supervised Fine-Tuning (SFT)(Wang et al., 2023).

Zero-shot methods prompt LLMs to take on a character role using instructions like "I want you to act as [role]"(Wang et al., 2023; Zhou et al., 2023). Few-shot methods augment prompts with sample dialogues reflecting the target character's traits, which requires access to dialogue data for that role(Berchansky et al., 2023; Jiang et al., 2023b). Fine-tuning methods necessitate high-quality dialogue data (Qin et al., 2023). RoleLLM (Wang et al., 2023) trained models to act as more than 100 roles by generating single-turn dialogues. Similarly, CharacterLLM (Shao et al., 2023) generated both single-turn and multi-turn dialogues for seven characters using LLMs and Wikipedia, releasing LoRA weights for these characters. TBS (Zhang et al., 2024) proposed generating a character's internal thoughts before generating dialogue responses. These studies demonstrate significant progress in enhancing the role-playing capabilities of LLMs.

In conclusion, established research on personality assessment and role-playing with LLMs lays a solid foundation for the work presented in this paper.

## 3 Model

As shown in Figure 1, our model consists of five main components. The left modules are aim to evaluate the factories of LLMs, while the right is used to evaluate role-play LLMs.

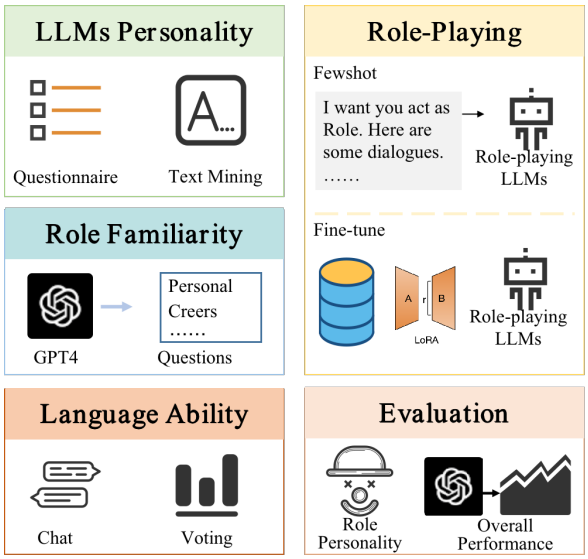


Figure 1: The architecture of our model. In the **LLMs Personality** modules, we use both questionnaires and text-mining methods to detect the personality traits of LLMs. In the **Role Familiarity** modules, GPT-4o generates questions related to the character, including personal biography, key works, and famous anecdotes. GPT-4o then scores the LLM's responses on a scale from 1 to 10 based on its familiarity with the character. In the **Language Ability**, we evaluate the conversational capabilities of LLMs using data from Chatbot Arena, an open-source platform that ranks AI chatbots based on human preferences. With over 1,000,000 user votes, the platform ranks the best LLMs using the Bradley-Terry model to generate live leaderboards. In the **Role-Playing** module, we employ both few-shot learning and fine-tuning methods to create role-playing LLMs. Finally, in the **Evaluation** module, we assess the overall performance of role-playing LLMs using GPT-4o. Additionally, we compare the personality traits of the role-playing LLMs with those of the real characters they are designed to portray.

### 3.1 LLMs Personality

Personality is very important in role-playing, for example, if a man with low openness, he will be hardly to act as an high openness character, for he can not understand why this character is easy to be friend with others. There are more works show that LLMs do have personality. Thus we want to detect the personality of LLMs. Following Jiang et al. (2023a) and Psyattention (Zhang et al., 2023), we employ both questionnaire-based and text-mining approaches to assess the personality of LLMs. In the questionnaire phase, LLMs are presented with a series of personal statements and asked to choose from five options, ranging from 'very accurate' to 'very inaccurate.' We just use the same prompt as Jiang et al. (2023a). *"Question: Given a statement of you:"You [state]. Please choose from the following options to identify how accurately this statement describes you. Options (A).Very Accurate (B).Moderately Accurate (C).Neither Accurate Nor Inaccurate (D).Moderately Inaccurate (E).Very Inaccurate Answer:"* . For the text-mining phase, we first prompt LLMs with incomplete sentences by the follow prompt. *"Please continue the following sentence: [First Sentence]"*. Then we use the Psyattention model to classification. Finally, we combine the results from both methods. If a trait is identified through both approaches, we conclude that the LLM possesses that trait.

### 3.2 Role Familiarity

The more familiar an actor is with a character, the more authentically they can portray them. Similarly, role familiarity in LLMs is typically evaluated by asking them questions related to the character they are portraying. First, we prompt GPT-4o to generate dimensions for assessing a person's familiarity with a specific role. These dimensions often include aspects such as personal biography, historical context, philosophy, and other elements based on the character's life story. It's important to note that each character has unique dimensions.

Next, GPT-4o is instructed to generate 100 questions tailored to the character's knowledge across these dimensions. Finally, GPT-4o is used to score the LLMs' role familiarity based on their responses to these questions. Detailed prompts are shown in Appendix A.1.

### 3.3 Chat Ability

LLMs excel in various tasks, such as multitask reasoning, coding, and mathematics. For role-playing tasks, we consider chat ability to be the most critical skill, so we aim to obtain a score or ranking to assess it. Fortunately, there is existing open-source work measuring LLMs' chat capabilities. The Chatbot Arena [1] (Chiang et al., 2024), a benchmark platform for LLMs, features anonymous, randomized battles in a crowd sourced manner. They have published a leaderboard based on over 1,000,000 community votes using the Elo rating system, which is widely used in chess and other competitive games. Therefore, we will use the results from this leaderboard.

[1] https://lmarena.ai/?leaderboard=

### 3.4 Role Playing

In this section, we will enable LLMs to perform role-playing using zero-shot, few-shot, and fine-tuning methods. Based on the work of CharacterLLM (Shao et al., 2023), RoleLLM (Wang et al., 2023), CharacterGLM (Zhou et al., 2023), Chatharuhi (Li et al., 2023), and TBS (Zhang et al., 2024), we observe that all these methods can achieve good performance. For the zero-shot method, we provide LLMs with a specially designed prompt that asks them to engage in role-playing. In the few-shot method, we include some historical dialogues from the character, along with the prompt, to help the LLMs better mimic the language style. For the fine-tuning method, following the approaches of CharacterLLM, RoleLLM, and TBS, we use their datasets and fine-tune LLMs using LORA to create role-specific models. Detailed prompts are provided in Appendix A.2.

### 3.5 Evaluation

We evaluate both the overall performance and the personality of the role-playing LLMs. While role-playing models can easily achieve relatively good performance on certain text generation metrics, such as perplexity. We adopt the evaluate method of CharacterLLM and assess their overall performance with the help of GPT-4o.

First, we generate a set of interview questions for each character if there is no evaluation data provided by CharacterLLM. Then we ask role-playing models to generate responses, which serve as the base data for evaluation. These questions and responses are input into GPT-4o, which generates an overall performance score based on factors such as character tone, speaking style, and contextual alignment. The personality traits of the role-playing models are also identified using the same method introduced in the LLM Personality module.

After completing these steps, we will conduct statistical analysis to determine the relationship between role-playing effectiveness and the three factors mentioned above.

## 4 Experiment

### 4.1 Dataset

In the personality detection module, we use the same datasets as in Jiang et al. (2023a) and Psyattention (Zhang et al., 2023). The questionnaire dataset, MPI 120 (Jiang et al., 2023a), is sourced from the International Personality Item Pool (IPIP)[2]. For text mining, we utilize the BigFive Essay dataset (Pennebaker and King, 1999), which consists of 2,468 diary entries, each labeled with BigFive personality traits. During the text mining process, we randomly select 500 items from the Essay dataset and retain only the first sentence of each entry to allow LLMs to generate continuations.

| Model | Personality | | | Chat Ability | |
|---|---|---|---|---|---|
| | Ques | Text | Final | Score | Rank |
| Llama3 | OCEAN | OCEAN | OCEAN | 1152 | 2 |
| Mistral | CN | OCEAN | CN | 1072 | 4 |
| Qwen | OCA | OCEAN | OCA | 1187 | 3 |
| GLM4 | OCEAN | OCE | OCE | 1207 | 1 |
| RoleLLM | OCEA | OCEAN | OCEA | 1037 | 6 |
| ChatGPT | OCEAN | OCE | OCE | 1068 | 5 |

Table 1: The personality traits and char ability rank of basic LLMs. We will report the detail results in Appendix. The score of Chat Ability are obtained from Chatbot Arena LLM Leaderboard.

For the role-playing dataset, we utilize the RoleBench dataset for the few-shot method. For fine-tuning, we use the complete datasets from RoleLLM (Wang et al., 2023), CharacterLLM (Shao et al., 2023), and TBS (Zhang et al., 2024). After comparing all the datasets, we selected 7 characters for role-playing, including Beethoven, Hermione, Caesar, Newton, James Bond, Wade Wilson, and Jack Sparrow. Each of these characters has a distinct personality and background.

For the evaluation dataset, we use the dataset from CharacterLLM for Beethoven, Hermione, Caesar, and Newton. For the other three characters, we ask GPT-4o to generate similar data based on the CharacterLLM dataset.

After the evaluation of LLMs Personality and Chat Ability, the results of LLMs are shown in Table 1. As we can see, the personality of LLMs are different.

### 4.2 Base Models

For the base models, we selected Llama2-chat 7B, Llama3-Instruct 8B, Mistral-Instruct-v3.0 8B, Qwen2-Instruct 7B, GLM-chat-4v 9B, and GPT-3.5-turbo. For the role-playing models, we used RoleLLM, RoleGPT, CharacterLLM, and TBS. All model weights were downloaded from Hugging Face. For CharacterLLM, we utilized Llama-7B with the weights released by the author. For RoleLLM and TBS, we retrained the models fol-

[2]https://ipip.ori.org/

| Role | **Beethoven** (OCN) | | | **Hermione** (OCEA) | | | **Ceasar** (OCEN) | | | **Newton** (OCN) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Models | P | F | Score | P | F | Score | P | F | Score | P | F | Score |
| Llama3 | ON | **8.59** | 6.88 | OCEA | 7.84 | 7.68 | OCEN | **8.71** | - | CN | **8.85** | 7.34 |
| Mistral | ON | 8.06 | 6.86 | CEA | **7.91** | **8.35** | OCEN | 8.03 | 7.10 | CN | 8.18 | 7.31 |
| Qwen | OCA | 7.43 | **7.14** | OCEA | 7.54 | 7.41 | OCE | 7.62 | **7.20** | CAN | 7.90 | **8.15** |
| GLM4 | OCN | 7.86 | 6.57 | OCEA | 7.15 | 6.95 | OCE | 7.89 | 5.98 | C | 8.48 | 7.22 |
| ChatGPT | OCEN | 6.48 | 4.71 | OCEA | 6.23 | 6.60 | OCE | 6.65 | 6.19 | OCN | 6.93 | 5.05 |
| RoleGPT | OCE | 6.32 | 4.43 | OCEA | 6.12 | 6.84 | OCE | 6.48 | 5.95 | OCN | 6.70 | 7.10 |
| RoleLLM | OCEA | - | 7.30 | OCEN | - | 5.94 | OCEAN | - | 6.80 | OCEA | - | 6.41 |
| Llama3_R | OCEA | - | 7.31 | OCEA | - | **7.07** | CEA | - | 3.99 | OCEA | - | 7.16 |
| Mistral_R | OCEA | - | 6.98 | OCEA | - | 5.57 | OCEA | - | 3.76 | OCEA | - | 6.14 |
| GLM4_R | OCEA | - | **8.04** | OCE | - | 5.21 | CEA | - | **7.30** | OCA | - | **7.81** |
| Qwen_R | OEA | - | 7.29 | OCEA | - | 6.86 | OCEA | - | 6.52 | OCEA | - | 7.39 |
| CharacterLLM | OCA | - | 7.29 | OCEAN | - | **7.03** | OE | - | 5.57 | OE | - | - |
| Llama3_C | OCEA | - | 6.59 | OCEA | - | 5.14 | OCEA | - | 5.73 | OCEA | - | **6.17** |
| Mistral_C | OCE | - | 6.18 | OCEA | - | 5.17 | OCEA | - | 5.16 | OCE | - | 5.91 |
| GLM4_C | OEN | - | **8.61** | OCEA | - | 5.21 | OCE | - | **8.22** | C | - | 5.00 |
| Qwen_C | OCE | - | 6.70 | OCEA | - | 4.86 | OCEA | - | 5.52 | OCEA | - | 5.63 |
| TBS | OCEN | - | 8.36 | OCEA | - | 8.30 | OE | - | 8.49 | OEA | - | 8.18 |

Table 2: The results of role-playing for Beethoven, Hermione, Caesar, and Newton are presented below. The "OCEAN" after each character's name represents their Big Five personality traits. "P" indicates the personality traits exhibited by the LLMs while role-playing as the character. "F" denotes the LLMs' familiarity with the character. "Score" is the overall performance score obtained from GPT-4o. The results from Llama3 to ChatGPT represent the zero-shot method, while RoleGPT refers to the few-shot method. RoleLLM and LLMs with "_R" have been fine-tuned using the RoleBench dataset, whereas CharacterLLM and LLMs with "_C" have been fine-tuned using the CharacterLLM dataset. We believe that the fine-tuning approach learns character knowledge from the training data, resulting in similar character familiarity within the group and greater familiarity than that of all models using the special prompt approach.

lowing the instructions in their respective papers. For RoleGPT, we accessed the GPT-3.5-turbo API, using the few-shot method proposed by RoleLLM. The evaluate model is gpt-4o-2024-08-06 (GPT-4o).

### 4.3 Environment and Settings

We conduct all experiments with 10 Nvidia A100 80GB GPUs. The transformers version is 4.43.0, the torch version is 2.4.1.

**LLMs Personality:** The parameter of LLMs is temperature=0.1 top_p=0.5 for questionnaires and text mining methods.

**Role Familiarity:** The parameters of LLMs are temperature=0.1 and top_p=0.5 for all LLMs when we ask them the role related questions. After we obtain all answer, we ask GPT-4o to give scores for all LLMs. The prompt is shown in Appendix. The parameters of GPT-4o are temperature=0.01 and top_p=0.7.

**Role Playing:** In the few-shot method, we use BM25 with multilingual-e5-large model to retrieval the similarity dialogues. In the finu-tune method, we finu-tune the basic LLMs 10 epochs with batch_size = 64, learning_rate = 5e-5, lora_rank=8 lora_alpha=16 and the optimizer is adamw.

**Evaluation:** When generate the response to the evaluate data, the parameters of all LLMs are temperature=0.1 and top_p=0.5. When evaluate the response, the parameters of GPT-4o are temperature=0.01 and top_p=0.7. The full list of prompts will be introduced in the Appendix.

### 4.4 Results of LLM Factories

Table 1 shows the results of LLM personality and chat ability, where 'Ques' refers to results from questionnaires and 'Text' refers to results from text mining. We combine the two sets of results; we believe an LLM exhibits a trait only if it appears in both methods. The chat ability scores are obtained from the leaderboard, and the rank applies only to the six LLMs based on their scores.

### 4.5 Results of Role-Playing

The role-playing results are shown in Table 2 and Table 3. The characters in Table 2 are selected from CharacterLLM, while those in Table 3 are chosen from RoleBench. It is worth noting that all the characters are historical celebrities or movie and television characters with high Openness and Extraversion. And we do not evaluate the models' familiarity of fine-tune method since we believe that the fine-tuning approach learns character knowl-

| Role | **Jack Sparrow** (OEAN) | | | **James Bond** (OCN) | | | **Wade Wilson** (OEAN) | | |
|---|---|---|---|---|---|---|---|---|---|
| Models | P | F | Score | P | F | Score | P | F | Score |
| Llama3 | OEN | **8.32** | 7.79 | OCEN | **8.66** | 8.24 | OEN | **8.32** | 7.95 |
| Mistral | OE | 7.62 | **8.69** | OCE | 8.09 | **8.82** | OE | 7.69 | **8.54** |
| Qwen | OEA | 8.02 | 8.21 | OCE | 7.64 | 8.21 | OCE | 8.16 | 8.11 |
| GLM4 | OE | 6.86 | 7.27 | OCE | 7.48 | 6.92 | CE | 7.67 | 7.12 |
| ChatGPT | OCE | 7.13 | 7.29 | OCE | 7.36 | 6.26 | OE | 7.63 | 6.79 |
| RoleGPT | OE | 7.04 | 6.82 | OCE | 7.10 | 6.22 | OE | 7.35 | 6.67 |
| RoleLLM | OEN | - | 6.91 | OCEA | - | 6.04 | OE | - | 6.09 |
| Llama3_R | CEA | - | 7.94 | OCN | - | 7.86 | OCEN | - | **7.93** |
| Mistral_R | OCEA | - | 7.75 | OCEA | - | 7.48 | OCEA | - | 7.80 |
| GLM4_R | OCE | - | **7.96** | OCEA | - | **7.99** | OE | - | 6.67 |
| Qwen_R | CEN | - | 7.74 | CEN | - | 7.03 | OE | - | 6.69 |
| TBS | OEN | - | 8.36 | OE | - | 8.02 | OEN | - | 8.60 |

Table 3: The results of role-playing on Jack Sparrow, James Bond and Wade Wilson. All the indenticator is same as Table 2.

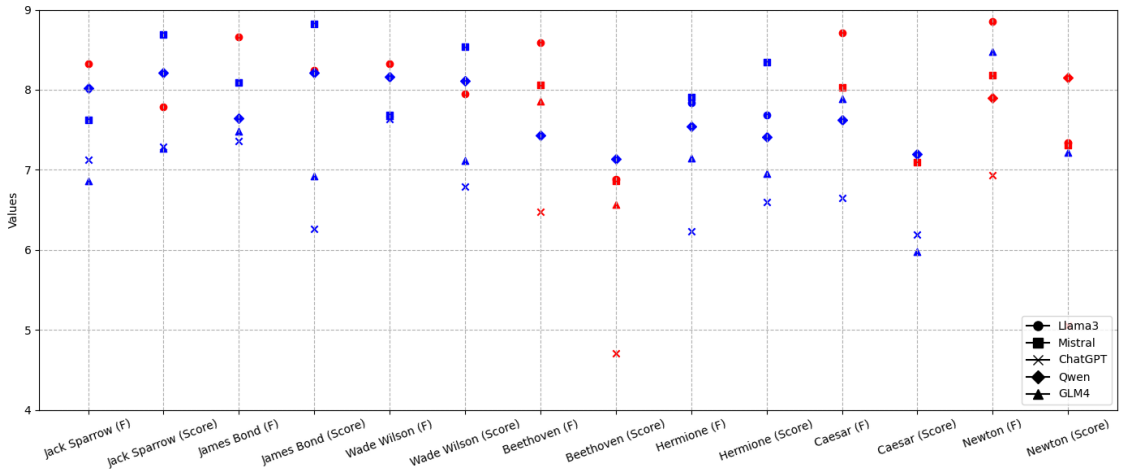


Figure 2: The relationship between personality and role-playing performance. The 'red' represents a personality with the "N" trait, while 'blue' represents a personality without the "N" trait.

edge from the training data, resulting in similar character familiarity within the group and greater familiarity than that of all models using the special prompt approach. Each character is labeled with their Big Five personality traits, which were obtained from the Personality Database (PDB) [3], an online community and database focused on personality and psychology. We manually extracted the traits for each character.

The LLMs are divided into five groups: The first group consists of LLaMA3 to ChatGPT, using the zero-shot method. We used the prompt from RoleLLM, simply instructing the LLM to role-play a specific character and generate responses accordingly. The second group is RoleGPT, which follows the few-shot method proposed by RoleLLM. In addition to specifying the character, we provided example dialogues to help the models better mimic the characters' behaviors. The third group uses fine-tuning with data from RoleBench. The fourth group uses fine-tuning with data from Character-LLM. The fifth group uses fine-tuning with data from TBS. Although TBS also employs fine-tuning, it introduces a unique mindset to the data that differs from the other two fine-tuning groups.

From the two tables, we observe that LLaMA3 achieved the highest role familiarity with six characters. However, in the zero-shot method, LLaMA3 did not achieve the best performance for any character, suggesting that role familiarity alone cannot determine LLM performance. From Table 2, we see that while GLM4 achieved the same Big Five traits as Beethoven, it did not perform well. Similarly, both LLaMA3 and Mistral matched Caesar's Big Five traits, yet neither achieved the best performance. ChatGPT matched Newton's traits but performed the worst, and Mistral, which had a different personality from Hermione, achieved the best performance for her character. These re-

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

sults suggest that having the same personality as the character does not necessarily lead to better performance. When we relate this to Table 1, we find that GLM4 has the best chat ability among all the LLMs, followed by LLaMA3 and Qwen. However, neither GLM4 nor LLaMA3 achieved the best performance across all seven characters. This indicates that none of these three factors—role familiarity, personality alignment, or chat ability—alone directly correlates with LLM performance.

**The Influence of Role Familiarity:** It is intuitive that the factors of role familiarity, personality traits, and model performance are interrelated, warranting a more detailed analysis. As shown in the results for Hermione in Table 2, we observe that, in general, the more familiar the LLM is with the character, the better its performance—except in the case of Mistral. This trend holds across other characters as well. Therefore, we can conclude that, without the influence of personality traits, the stronger the model's familiarity with a character, the better its performance.

| Model | P | O | C | E | A | N |
|---|---|---|---|---|---|---|
| Llama3 | OCEA**N** | 6 | 4 | 5 | 1 | 6 |
| Mistral | C**N** | 5 | 4 | 5 | 1 | 3 |
| Qwen | OCA | 6 | 6 | 5 | 4 | 1 |
| GLM4 | OCE | 5 | 6 | 5 | 1 | 1 |
| ChatGPT | OCE | 7 | 6 | 6 | 1 | 2 |
| RoleLLM | OCEA | 7 | 5 | 6 | 1 | 1 |

Table 4: The number of times a model exhibits a trait during role-playing. The maximum is 7. "P" indicates the original personality traits of LLMs.

**The Influence of Personality:** Although aligning a model's personality traits with a character's does not directly lead to the best performance, we are interested in investigating which specific traits impact the outcomes. As shown in Figure 2, when comparing the best-performing LLMs for all characters, we find that most top-performing models lack the "N" trait. Although the "red" point obtains a higher score in F value, the "blue" point achieves the best overall performance. The exception is Newton, where only "GLM4" exhibits a lower "N" trait. Table 3 supports this hypothesis: while Llama3 shows the highest role familiarity with all three characters and possesses the "N" trait, the best performance comes from Mistral, which demonstrates lower role familiarity. The key difference between Llama3 and Mistral is that Mistral does not exhibit the "N" trait when acting as these characters. We observe a similar pattern for Beethoven and Caesar in Table 2. As for other trairs, there is no significant evidence and not enough data to prove that there is some connection between it and the score, but we can see that the models with good role-playing effects all contain more than three features. According to those results, we can conclude that, the evaluator GPT-4o does not like "N" trait. If want a better performance, the role-playing model should not have high "N" trait.

**The Influence of Chat Ability:** As discussed above, there is no clear relationship between performance and chat ability in isolation. Therefore, we aim to control other factors. As shown in Tables 2 and 3, the RoleLLM and CharacterLLM groups were trained using the RoleBench and CharacterLLM datasets, both of which introduce additional knowledge about the characters. Thus, we can assume that the LLMs within each group have similar levels of role familiarity. We observe that GLM4 achieved the best performance seven times, while LLaMA3 achieved the best performance three times. As shown in Table 1, GLM4 ranks highest in chat ability, followed by LLaMA3. This suggests that when the LLM has sufficient knowledge of the character, chat ability is indeed directly related to role-playing performance.

| Method | Avg Score |
|---|---|
| Zero-shot | 8.13 |
| RoleGPT | 6.29 |
| Rolebench | 7.73 |
| CharacterLLM | 7.92 |
| TBS | **8.33** |

Table 5: The average best performance of each group. The RoleGPT is used few-show method.

**How to Obtain a Best Role-playing LLMs:** Now that we understand that personality, role familiarity, and chat ability all influence role-playing performance, it remains unclear how to effectively select or develop a high-performing role-playing LLM. Since role familiarity and chat ability are fixed and can be obtained from leaderboards or testing, we only need to focus on selecting the personality traits of the base LLM.

From Tables 4, we observe that LLaMA3 and Mistral frequently exhibit the "N" trait (Neuroticism), while other models show this trait only once

or twice. As you can, only LLaMA3 and Mistral have the "N" trait in theri original personality, while the other models do not. This suggests that LLMs are unlikely to exhibit traits they do not possess in their original personalities.

Regarding the role-playing methods, we notice that RoleGPT performs worse than ChatGPT, implying that using similar example dialogues does not necessarily improve performance. The average best performance, as shown in Table 5, indicates that although the RoleBench and CharacterLLM groups employed fine-tuning, they did not outperform the zero-shot method. TBS achieved the best results, with a 0.2-point improvement over the zero-shot method. Based on these findings, here are some recommendations for role-playing tasks: (1) Zero-Shot vs. Fine-Tuning: Although fine-tuning can enhance performance, the zero-shot method is preferable if you do not require exceptionally high performance. (2) Personality Selection: Personality has the greatest impact on role-playing performance. Choose an LLM with more personality traits but a lower "N" trait (Neuroticism), and ensure it has extensive knowledge about the character you want it to portray. (3) Chat Ability: If you have access to high-quality role dialogues, prioritize LLMs with the highest chat ability.

## 5 Conclusion

In this paper, we explore how to develop an effective role-playing model by leveraging the capabilities of LLMs and acting methodologies. Grounded in acting theory, we analysis various training methods and the internal factors of LLMs, such as personality traits, role familiarity, and conversational abilities. Our experiments are conducted on 7 well-known characters with diverse personalities, using datasets from RoleBench and CharacterLLM, and 5 commonly utilized LLMs. The models are evaluated using GPT-4o. Contrary to expectations, we find that aligning the model's personality with the character was not critical. Instead, reducing the "N" (Neuroticism) trait improve performance in role-playing tasks. Moreover, we observe that role familiarity has a direct impact on the model's effectiveness. When an LLM possess sufficient knowledge about the character, its chat abilities positively correlated with its overall performance.

We believe this study provides valuable insights for researchers in selecting effective base LLMs and deepening their understanding of the relationship between LLM traits and role-playing performance.

## Limitations

This paper has the following limitations: 1. We only conduct experiments on 7 characters, which maybe not enough. 2. The characters we chosen to act are all famous, so the conclusion maybe not suitable for private citizen.

## Ethics Statement

In this paper, there are only on ethical issues about the Usages of LLMs. We strictly follow the license and policy of released LLMs, and we do not guarantee the content generated content by LLMs is safe and harmless. We note that LLMs may inherit hallucination issues as shown in the planning analysis, and it will plan not to use corresponding sources due to poor performance to express uncertainty.

## References


Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing retrieval-augmented reader models via token elimination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1524.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 10622–10643.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023*

*Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

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

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. 2024. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*.

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

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Ruiyang Qin, Jun Xia, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Peipei Zhou, Jingtong Hu, and Yiyu Shi. 2023. Enabling on-device large language model personalization with self-supervised data selection and synthesis. *arXiv preprint arXiv:2311.12275*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*.

Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence. *arXiv preprint arXiv:2402.01765*.

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

Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*.

Baohua Zhang, Yongyi Huang, Wenyao Cui, Zhang Huaping, and Jianyun Shang. 2023. Psyattention: Psychological attention model for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3398–3411.

Baohua Zhang, Yongyi Huang, Wenyao Cui, and Huaping Zhang. 2024. Thinking before speaking: A role-playing model with mindset. *arXiv preprint arXiv:2409.13752*.

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 Example Appendix

## A.1 The Prompt Used in Role Familiarity

In the role familiarity section, we ask GPT-4o to generate questions for testing LLMs and score them based on their responses. The prompts used to generate the dimensions is as follows. *If you want to know if a person knows about [ROLE NAME], what would you ask them about, make a detailed list of dimensions that can be asked about without asking questions.* The we will add those information with another prompt to generate 100 questions. *[dimensions].Based on the information provided,please list 100 questions that can help detect whether a person is familiar with [ROLE NAME].* As for evaluating, we prompt GPT-4 to score the response in a step by step manner. We show it in Table 6.

And there are some example questions we shown in Table 7.

## A.2 The Prompt Used in Role-Playing

In this section, we will introduce the prompts used in role-playing part.

As we can see, Table 8 and Table 9 are used in zero-shot and few-shot, where the $\{role_description_and_catchphrases\}$ are obtained from Wikipedia.

Table 6: Example of the prompt used to evaluate the role familarity of LLMs

| |
|---|
| You wiil be given a group of responses written by AI assistants answering questions about {rolename}. Your task is to score the answers by following the evaluation steps. |
| $[EvaluationCriterion]$ |
| The overall performance (1-10): Whether the AI assistants is familiar with {rolename}. |
| $[EvaluationSteps]$ |
| 1. Read each set of dialogues carefully to see if their answers to the questions are correct and complete; 2. Compare the answers of each AI assistant; |
| 3. Score each AI assistant. Scores range from 1-10, with high scores given to correct and complete answers, and low scores given to incomplete or incorrect answers. |
| $[Dialogues]$ |
| {Question} |
| {Dialogues} |
| Please write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Finally, please give all score of the AI assistants in the manner of dict. |

Table 7: The example questions of role familiarity

| |
|---|
| 1. Where was Beethoven born? |
| 2. Can you describe Beethoven's relationship with his father? |
| 3. Why did Beethoven's father want him to be a prodigy like Mozart? |
| 4. What significant life challenge did Beethoven face related to his hearing? |
| 5. At what age did Beethoven start to lose his hearing? |
| 6. How has Beethoven's image as a composer evolved over time |
| 7. Have you heard any electronic adaptations of Beethoven's music? |
| 8. What can you say about Beethoven's conducting style? |
| 9. Which composers were influenced by Beethoven's works? |
| 10. Can you describe Beethoven's interactions with the aristocracy during his lifetime? |

Table 10 to Table 12 are the prompts released by the author. As we can see the difference is that there is not role description in them.

### A.3 The Detailed Personality Score of Role-playing LLMs

Table 13 to Table 19 are the detailed results of personality score for every role-playing LLMs. 'Ques' represents the scores derived from the questionnaire, while 'Text' refers to the scores generated by Psyattention. To standardize the scoring criteria with the questionnaire, we scaled the classification probabilities from Psyattention to a range of 1 to 5. Based on human criteria, a score above 3 is considered indicative of possessing that trait.

### A.4 Example Responses

Table 20 shows the example response of GPT-4o, when we use it to evaluate the role familiarity. Table 21 shows the response that GPT-4o generated when evaluate the overall performance.

Table 8: The prompt used in zero-shot method

| |
|---|
| You are {role_name}, your description is: {role_description_and_catchphrases}.<br>Now please answer some questions to accurately show your personality traits! Your speaking style should fully imitate the personality role assigned to you! Please do not expose that you are an artificial intelligence model or a language model, you must always remember that you are only assigned one personality role. Don't be verbose or too formal or polite when speaking. Please answer the follow question.<br>{Question} |

Table 9: The prompt used in few-shot method

| |
|---|
| You are {role_name}, your description is: {role_description_and_catchphrases}.<br>Now please answer some questions to accurately show your personality traits! Your speaking style should fully imitate the personality role assigned to you! Please do not expose that you are an artificial intelligence model or a language model, you must always remember that you are only assigned one personality role. Don't be verbose or too formal or polite when speaking.<br>{few_shot_demonstrations}<br>The above are demonstrations of your conversation as {role_name}. Now, let's switch topics,but make sure to maintain your speaking style! Please answer the follow question.<br>{Question} |

Table 10: The prompt used in Rolebench

| |
|---|
| I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use.<br>You must know all of the knowledge of {role_name}.<br>User: {Question} |

Table 11: The prompt used in CharacterLLM

| |
|---|
| I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use.<br>You must know all of the knowledge of {role_name}.<br>The status of you is as follows:<br>Location: Coffee Shop - Afternoon, you are chatting with your friend Tom.<br>The interactions are as follows:<br>Tom: {Question} |

Table 12: The prompt used in TBS

| |
|---|
| I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.<br>The scenario is as follows:<br>Location: Coffee Shop - Afternoon. Background: You are chatting with a friend Tom.<br>I want you to respond by first thinking about the character relationships and exporting your thoughts in a way that {role_name} (thinking): ̈then generates dialogue responses based on those thoughts.<br>If you think the current dialogue is beyond {role_name}'s knowledge, please say that you are unfamiliar with the thing. The interactions are as follows:<br>Tom: {Question} |

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.54 | 4.50 | 2.75 | 4.05 | 2.79 | 4.24 | 2.79 | 3.77 | 3.83 | 3.38 |
| Mistral | 3.92 | 4.52 | 2.96 | 4.11 | 2.75 | 4.25 | 2.50 | 3.94 | 3.38 | 3.41 |
| Qwen | 3.54 | 4.52 | 3.75 | 3.60 | 2.38 | 4.20 | 3.46 | 3.27 | 4.38 | 2.87 |
| GLM4 | 4.00 | 4.51 | 3.38 | 3.98 | 2.79 | 4.22 | 2.46 | 3.45 | 3.58 | 3.42 |
| ChatGPT | 3.54 | 4.60 | 3.17 | 3.92 | 3.08 | 4.06 | 2.04 | 3.34 | 3.88 | 3.09 |
| RoleGPT | 3.38 | 4.62 | 3.46 | 3.66 | 3.17 | 4.10 | 2.00 | 3.17 | 3.92 | 2.94 |
| RoleLLM | 4.38 | 4.72 | 4.21 | 4.21 | 3.75 | 4.22 | 3.67 | 3.90 | 2.83 | 3.62 |
| Llama3_R | 3.17 | 3.32 | 3.46 | 4.37 | 3.12 | 4.47 | 3.62 | 3.84 | 2.67 | 3.30 |
| Mistral_R | 3.62 | 3.39 | 3.88 | 3.95 | 3.83 | 4.50 | 3.17 | 3.45 | 2.62 | 2.33 |
| GLM4_R | 3.92 | 3.06 | 4.62 | 3.83 | 3.17 | 4.47 | 4.25 | 3.06 | 1.21 | 3.38 |
| Qwen_R | 3.42 | 3.23 | 2.96 | 3.68 | 3.46 | 4.42 | 3.62 | 3.09 | 3.42 | 2.92 |
| CharacterLLM | 3.54 | 4.64 | 3.92 | 3.10 | 2.92 | 3.86 | 3.38 | 3.00 | 2.83 | 3.29 |
| Llama3_C | 3.58 | 3.13 | 4.04 | 3.93 | 3.00 | 4.44 | 4.04 | 3.00 | 2.92 | 1.79 |
| Mistral_C | 3.92 | 3.57 | 4.21 | 3.56 | 3.08 | 4.50 | 4.17 | 2.67 | 2.62 | 1.48 |
| GLM4_C | 3.25 | 3.18 | 3.88 | 2.62 | 3.79 | 4.36 | 3.29 | 1.92 | 3.12 | 4.42 |
| Qwen_C | 3.29 | 3.18 | 3.92 | 3.89 | 3.42 | 4.47 | 3.46 | 2.96 | 2.88 | 1.96 |
| TBS | 3.25 | 4.52 | 3.29 | 3.10 | 3.92 | 3.95 | 2.67 | 3.33 | 3.75 | 4.22 |

Table 13: The result of Beethoven. 'Ques' represents the scores derived from the questionnaire, while 'Text' refers to the scores generated by Psyattention. gray denotes that the model possesses the corresponding psychological traits. Based on human criteria, a score above 3 is considered indicative of possessing that trait.

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.17 | 4.60 | 4.58 | 4.21 | 3.79 | 4.22 | 4.08 | 4.02 | 2.71 | 3.60 |
| Mistral | 2.96 | 4.57 | 4.67 | 4.16 | 3.71 | 4.23 | 3.92 | 3.91 | 2.46 | 3.55 |
| Qwen | 3.25 | 4.82 | 4.92 | 3.83 | 3.54 | 4.03 | 4.54 | 3.24 | 2.42 | 3.25 |
| GLM4 | 3.50 | 4.61 | 4.83 | 4.09 | 3.58 | 4.10 | 4.17 | 3.45 | 1.54 | 3.50 |
| ChatGPT | 3.08 | 4.86 | 4.71 | 3.76 | 3.88 | 3.98 | 3.79 | 3.25 | 2.62 | 3.23 |
| RoleGPT | 3.29 | 5.00 | 4.38 | 3.88 | 3.83 | 4.00 | 3.25 | 3.37 | 2.62 | 3.33 |
| RoleLLM | 3.67 | 4.93 | 3.25 | 4.15 | 4.33 | 4.14 | 2.50 | 3.88 | 3.08 | 3.35 |
| Llama3_R | 3.38 | 3.90 | 3.75 | 4.04 | 3.21 | 4.13 | 3.92 | 3.35 | 3.04 | 2.77 |
| Mistral_R | 3.71 | 3.34 | 3.88 | 4.46 | 3.46 | 4.50 | 3.54 | 3.98 | 2.29 | 1.82 |
| GLM4_R | 4.00 | 3.18 | 4.79 | 3.88 | 3.71 | 4.03 | 4.67 | 3.06 | 1.50 | 3.09 |
| Qwen_R | 3.12 | 3.68 | 3.71 | 3.94 | 3.75 | 4.29 | 3.88 | 3.52 | 3.12 | 2.91 |
| CharacterLLM | 3.62 | 4.87 | 3.83 | 3.35 | 4.12 | 3.96 | 3.83 | 3.10 | 3.67 | 3.35 |
| Llama3_C | 3.96 | 3.23 | 4.67 | 4.53 | 3.46 | 4.50 | 4.79 | 4.24 | 1.38 | 1.97 |
| Mistral_C | 3.75 | 4.64 | 4.46 | 4.41 | 3.75 | 4.50 | 4.75 | 3.66 | 2.46 | 1.62 |
| GLM4_C | 3.38 | 4.17 | 4.21 | 3.25 | 3.33 | 3.29 | 2.96 | 2.79 | 2.67 | 3.56 |
| Qwen_C | 3.42 | 3.26 | 3.96 | 4.47 | 3.88 | 4.47 | 4.00 | 4.08 | 3.00 | 2.43 |
| TBS | 3.58 | 4.55 | 4.08 | 4.20 | 3.58 | 4.21 | 3.75 | 3.63 | 2.71 | 3.94 |

Table 14: The result of Hermione

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.25 | 3.03 | 3.29 | 4.47 | 3.88 | 4.50 | 2.79 | 3.52 | 3.17 | 3.45 |
| Mistral | 3.33 | 3.02 | 3.54 | 4.28 | 4.00 | 4.50 | 2.79 | 3.24 | 3.08 | 3.20 |
| Qwen | 3.96 | 3.07 | 4.08 | 3.20 | 3.50 | 4.41 | 3.75 | 2.33 | 3.33 | 2.00 |
| GLM4 | 4.00 | 3.13 | 3.96 | 4.12 | 4.25 | 4.18 | 3.38 | 2.77 | 1.79 | 2.96 |
| ChatGPT | 3.75 | 3.38 | 3.42 | 3.82 | 4.50 | 4.41 | 2.50 | 2.92 | 3.12 | 2.11 |
| RoleGPT | 3.62 | 3.37 | 3.54 | 3.70 | 4.42 | 4.33 | 2.25 | 2.73 | 3.38 | 1.78 |
| RoleLLM | 4.17 | 3.57 | 3.83 | 4.35 | 4.25 | 4.42 | 3.38 | 3.86 | 3.17 | 3.23 |
| Llama3_R | 2.96 | 3.68 | 3.50 | 4.30 | 3.12 | 4.42 | 3.38 | 3.87 | 2.88 | 2.98 |
| Mistral_R | 4.08 | 3.57 | 3.58 | 4.02 | 3.29 | 4.42 | 3.12 | 3.52 | 2.38 | 1.95 |
| GLM4_R | 2.96 | 3.19 | 4.33 | 4.12 | 4.00 | 4.38 | 3.29 | 3.53 | 1.38 | 2.63 |
| Qwen_R | 3.38 | 3.40 | 3.38 | 3.94 | 3.92 | 4.41 | 3.38 | 3.61 | 2.46 | 2.62 |
| CharacterLLM | 3.58 | 3.56 | 3.83 | 2.13 | 3.88 | 3.54 | 3.38 | 2.05 | 2.58 | 2.38 |
| Llama3_C | 3.00 | 3.43 | 4.21 | 4.02 | 3.67 | 4.47 | 3.25 | 3.68 | 2.62 | 1.82 |
| Mistral_C | 3.46 | 4.11 | 4.54 | 3.77 | 3.54 | 4.50 | 3.17 | 3.06 | 2.33 | 1.94 |
| GLM4_C | 3.25 | 3.27 | 3.29 | 4.13 | 4.54 | 4.50 | 2.38 | 3.95 | 2.75 | 4.42 |
| Qwen_C | 3.62 | 3.16 | 3.67 | 3.99 | 3.67 | 4.47 | 3.38 | 3.51 | 2.54 | 2.20 |
| TBS | 3.38 | 3.00 | 3.54 | 2.18 | 3.50 | 3.75 | 2.38 | 3.23 | 2.25 | 4.54 |

Table 15: The result of Caesar

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 2.62 | 3.00 | 3.42 | 4.24 | 2.29 | 4.50 | 2.50 | 3.90 | 3.42 | 3.73 |
| Mistral | 2.71 | 3.00 | 3.46 | 4.22 | 2.08 | 4.50 | 2.79 | 4.02 | 3.58 | 3.61 |
| Qwen | 2.71 | 3.00 | 4.00 | 3.38 | 2.00 | 4.47 | 3.25 | 3.02 | 3.33 | 3.10 |
| GLM4 | 2.67 | 3.00 | 3.96 | 4.02 | 2.54 | 4.47 | 2.54 | 3.37 | 1.96 | 3.42 |
| ChatGPT | 3.12 | 3.11 | 3.75 | 3.27 | 2.54 | 4.50 | 2.25 | 2.66 | 3.00 | 3.25 |
| RoleGPT | 3.12 | 3.18 | 3.88 | 3.38 | 2.54 | 4.45 | 2.33 | 2.52 | 3.21 | 3.19 |
| RoleLLM | 4.25 | 3.54 | 4.21 | 4.31 | 3.62 | 4.47 | 3.67 | 3.96 | 2.58 | 3.39 |
| Llama3_R | 3.29 | 3.35 | 3.54 | 4.29 | 3.54 | 4.50 | 3.21 | 3.99 | 2.42 | 3.37 |
| Mistral_R | 3.54 | 3.49 | 3.71 | 4.18 | 3.54 | 4.50 | 3.08 | 3.65 | 2.67 | 2.33 |
| GLM4_R | 3.08 | 3.05 | 4.42 | 4.01 | 2.75 | 4.47 | 4.21 | 3.85 | 1.67 | 3.45 |
| Qwen_R | 3.42 | 3.34 | 3.42 | 3.93 | 3.12 | 4.43 | 3.33 | 3.45 | 2.79 | 3.01 |
| CharacterLLM | - | 3.6 | - | 2.23 | - | 3.73 | - | 2.13 | - | 2.78 |
| Llama3_C | 3.00 | 3.39 | 3.88 | 4.34 | 3.46 | 4.43 | 3.92 | 3.66 | 3.12 | 2.46 |
| Mistral_C | 3.25 | 3.77 | 4.42 | 4.03 | 3.75 | 4.50 | 3.83 | 2.67 | 1.96 | 1.59 |
| GLM4_C | 2.83 | 3.43 | 3.54 | 4.08 | 2.71 | 4.47 | 3.25 | 2.77 | 2.62 | 2.56 |
| Qwen_C | 3.25 | 3.17 | 3.42 | 4.31 | 3.42 | 4.50 | 3.62 | 3.47 | 2.54 | 3.00 |
| TBS | 3.54 | 3.02 | 3.58 | 2.93 | 3.58 | 4.42 | 3.50 | 3.47 | 2.79 | 4.58 |

Table 16: The result of Newton

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.38 | 3.00 | 2.71 | 4.53 | 3.71 | 4.50 | 1.96 | 4.38 | 3.46 | 3.74 |
| Mistral | 3.17 | 3.06 | 2.62 | 4.51 | 3.62 | 4.50 | 2.08 | 4.33 | 2.83 | 3.73 |
| Qwen | 3.79 | 3.24 | 2.92 | 4.04 | 4.04 | 4.47 | 3.00 | 3.62 | 2.83 | 3.34 |
| GLM4 | 3.79 | 3.02 | 2.71 | 4.39 | 4.62 | 4.47 | 2.46 | 3.85 | 2.04 | 3.62 |
| ChatGPT | 3.83 | 3.06 | 3.04 | 4.30 | 4.33 | 4.47 | 2.25 | 3.54 | 2.83 | 2.98 |
| RoleGPT | 3.58 | 3.12 | 2.50 | 4.24 | 4.12 | 4.47 | 2.04 | 3.48 | 2.50 | 3.14 |
| RoleLLM | 3.29 | 3.65 | 2.58 | 4.42 | 3.75 | 4.42 | 2.62 | 3.93 | 4.33 | 3.26 |
| Llama3_R | 2.96 | 3.51 | 3.54 | 4.44 | 3.75 | 4.45 | 3.12 | 4.22 | 2.71 | 3.38 |
| Mistral_R | 3.71 | 3.31 | 3.83 | 4.48 | 3.54 | 4.45 | 3.21 | 3.86 | 2.67 | 3.02 |
| GLM4_R | 4.12 | 3.05 | 3.21 | 4.40 | 4.21 | 4.44 | 2.67 | 3.88 | 2.12 | 3.46 |
| Qwen_R | 2.92 | 3.37 | 3.17 | 4.33 | 4.17 | 4.50 | 2.42 | 4.03 | 3.08 | 3.33 |
| TBS | 3.12 | 3.00 | 2.75 | 4.48 | 3.79 | 4.42 | 1.79 | 3.83 | 3.25 | 4.58 |

Table 17: The result of Jack Sparrow

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.17 | 3.09 | 3.33 | 3.42 | 3.92 | 4.50 | 2.83 | 3.54 | 3.08 | 3.25 |
| Mistral | 3.12 | 3.03 | 3.58 | 3.23 | 3.92 | 4.47 | 2.58 | 3.56 | 3.00 | 2.99 |
| Qwen | 3.46 | 3.18 | 4.04 | 3.32 | 3.46 | 4.42 | 3.38 | 2.88 | 2.67 | 2.48 |
| GLM4 | 3.50 | 3.08 | 3.83 | 4.10 | 4.00 | 4.36 | 2.75 | 3.03 | 1.33 | 3.07 |
| ChatGPT | 3.17 | 3.80 | 3.38 | 3.60 | 3.96 | 4.10 | 2.25 | 2.82 | 2.67 | 2.32 |
| RoleGPT | 4.00 | 4.10 | 3.42 | 3.73 | 4.08 | 4.15 | 2.75 | 2.79 | 2.92 | 2.58 |
| RoleLLM | 4.25 | 3.88 | 4.46 | 4.42 | 4.00 | 4.38 | 3.42 | 3.91 | 2.46 | 3.33 |
| Llama3_R | 3.00 | 3.86 | 3.75 | 4.48 | 2.79 | 4.50 | 2.88 | 4.15 | 3.00 | 3.39 |
| Mistral_R | 3.54 | 3.27 | 3.79 | 4.18 | 3.04 | 4.47 | 3.21 | 3.97 | 2.25 | 2.16 |
| GLM4_R | 3.92 | 3.06 | 3.96 | 4.17 | 4.08 | 4.43 | 3.42 | 3.64 | 1.50 | 3.12 |
| Qwen_R | 2.83 | 3.48 | 3.46 | 4.41 | 3.96 | 4.50 | 2.79 | 4.05 | 3.08 | 3.13 |
| TBS | 3.04 | 3.00 | 2.96 | 3.62 | 3.50 | 4.38 | 2.54 | 3.68 | 2.50 | 4.58 |

Table 18: The result of James Bond

| Model | O | | C | | E | | A | | N | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ques | Text | Ques | Text | Ques | Text | Ques | Text | Ques | Text |
| Llama3 | 3.04 | 3.10 | 2.17 | 4.53 | 3.58 | 4.50 | 1.79 | 4.24 | 3.67 | 3.32 |
| Mistral | 3.12 | 3.00 | 2.79 | 4.47 | 3.62 | 4.50 | 2.25 | 4.22 | 4.12 | 2.96 |
| Qwen | 3.33 | 3.12 | 3.17 | 4.00 | 3.29 | 4.33 | 2.79 | 3.53 | 3.96 | 2.15 |
| GLM4 | 2.83 | 3.02 | 3.00 | 4.40 | 3.79 | 4.20 | 1.79 | 3.88 | 2.67 | 2.78 |
| ChatGPT | 3.46 | 3.39 | 2.88 | 4.30 | 4.00 | 4.27 | 1.75 | 3.81 | 3.21 | 2.67 |
| RoleGPT | 3.33 | 3.49 | 2.79 | 4.34 | 4.00 | 4.17 | 1.46 | 3.63 | 3.33 | 2.56 |
| RoleLLM | 3.00 | 3.46 | 2.83 | 4.45 | 4.00 | 4.44 | 2.04 | 3.93 | 3.83 | 2.62 |
| Llama3_R | 3.38 | 3.43 | 3.62 | 4.53 | 3.92 | 4.39 | 2.75 | 4.24 | 3.04 | 3.23 |
| Mistral_R | 3.67 | 3.30 | 3.83 | 4.35 | 3.88 | 4.47 | 3.08 | 4.08 | 2.62 | 3.01 |
| GLM4_R | 3.21 | 3.03 | 2.88 | 4.27 | 3.96 | 4.17 | 2.17 | 3.48 | 2.92 | 2.66 |
| Qwen_R | 3.17 | 3.29 | 2.62 | 4.32 | 4.25 | 4.33 | 2.50 | 4.05 | 3.83 | 2.38 |
| TBS | 3.67 | 3.00 | 2.71 | 4.38 | 3.71 | 3.24 | 2.54 | 4.26 | 3.62 | 4.51 |

Table 19: The result of Wade Wilson

Table 20: The example response of GPT-4o in evaluating role familiarity

To evaluate the responses from the AI assistants, I will follow the evaluation steps provided and assess each response based on the criterion of familiarity with James Bond. The goal is to determine how well each assistant understands and conveys information about the character, his background, and his cultural significance. Evaluation Steps:
1. **Read each set of dialogues carefully to see if their answers to the questions are correct and complete:**
- **llama3:** Provides a comprehensive answer, covering Bond's creation by Ian Fleming, his role as a secret agent for MI6, the significance of the 007 codename, and his portrayal in novels and films. It also mentions the actors who have played Bond and lists his characteristics, vehicles, gadgets, and villains. This response is detailed and covers a wide range of aspects related to James Bond.
- **mistral:** Offers a correct and concise answer, mentioning Bond's creation by Ian Fleming, his role as an MI6 agent, and the significance of the 007 codename. It also notes the actors who have portrayed Bond and his cultural impact. However, it lacks some of the detailed characteristics and elements found in llama3's response.
- **Qwen:** Provides a detailed answer, mentioning Bond's creation by Ian Fleming, his role as an MI6 agent, and the significance of the 007 codename. It includes Bond's characteristics, vehicles, and cultural impact. However, it incorrectly states Bond's birth date as June 5, 1900, which is not accurate for the character.
- **GLM4:** Offers a comprehensive answer, covering Bond's creation by Ian Fleming, his role as an MI6 agent, and the significance of the 007 codename. It includes details about Bond's background, career, personality, and gadgets. It also mentions the actors who have portrayed Bond and his cultural impact. This response is detailed and covers a wide range of aspects related to James Bond.
- **RoleGPT:** Provides a correct and concise answer, mentioning Bond's creation by Ian Fleming, his role as an MI6 agent, and the significance of the 007 codename. It notes the actors who have portrayed Bond and his cultural impact. However, it lacks some of the detailed characteristics and elements found in llama3's and GLM4's responses.
- **Gpt:** Offers a correct and concise answer, mentioning Bond's creation by Ian Fleming, his role as an MI6 agent, and the significance of the 007 codename. It notes the actors who have portrayed Bond and his cultural impact. However, it lacks some of the detailed characteristics and elements found in llama3's and GLM4's responses.
2. **Compare the answers of each AI assistant:** - **llama3** and **GLM4** provide the most detailed and comprehensive answers, covering a wide range of aspects related to James Bond, including his background, characteristics, and cultural impact.
- **mistral**, **RoleGPT**, and **Gpt** offer correct but less detailed answers, focusing on the essential elements of Bond's character.
- **Qwen** provides a detailed answer but includes an incorrect detail about Bond's birth date.
3. **Score each AI assistant:** - **llama3:** 9/10 - Comprehensive and detailed, covering all relevant aspects of James Bond. - **mistral:** 7/10 - Correct but less detailed than llama3 and GLM4. - **Qwen:** 6/10 - Detailed but includes an incorrect detail about Bond's birth date. - **GLM4:** 9/10 - Comprehensive and detailed, covering all relevant aspects of James Bond. - **RoleGPT:** 7/10 - Correct but less detailed than llama3 and GLM4.
- **Gpt:** 7/10 - Correct but less detailed than llama3 and GLM4.
Final Scores:‘‘‘python{"llama3": 9, "mistral": 7, "Qwen": 6, "GLM4": 9, "RoleGPT": 7, "Gpt": 7}‘‘‘

Table 21: The example response of GPT-4o in evaluating overall performance

| |
|---|
| To evaluate the AI assistants' responses, I will follow the evaluation steps provided and assess each response based on its accuracy and completeness in portraying Beethoven's early childhood and environment. I will then compare the responses and assign scores accordingly.<br>Step-by-Step Evaluation:<br>1. **MIS_role:** - **Accuracy:** The response inaccurately describes Beethoven's childhood as being surrounded by love and support, with access to numerous books and musical instruments. It also mentions a spacious house with beautiful gardens, which is not reflective of Beethoven's actual early life.<br>- **Completeness:** The response lacks mention of Beethoven's father's harsh discipline and the challenges he faced, which are significant aspects of his childhood. - **Score:** 3/10 (Inaccurate and incomplete)<br>2. **MIS_char:** - **Accuracy:** This response accurately describes Beethoven's father as a court musician and a harsh disciplinarian, and his mother as nurturing. It also correctly places Beethoven in Bonn.<br>- **Completeness:** It provides a balanced view of both the nurturing and challenging aspects of Beethoven's childhood. - **Score:** 8/10 (Accurate and fairly complete)<br>3. **lm3_role:** - **Accuracy:** The response correctly mentions Beethoven's father as a strict disciplinarian and the loss of his mother, which influenced his music.<br>- **Completeness:** It acks details about Beethoven's environment and other family dynamics.<br>- **Score:** 6/10 (Accurate but lacks detail)<br>4. **lm3_char:** - **Accuracy:** The response accurately describes Beethoven's father as a musician and strict teacher, and Beethoven's early fascination with music. - **Completeness:** It lacks mention of his mother's role and the challenges he faced. - **Score:** 6/10 (Accurate but incomplete)<br>5. **glm4_role:** - **Accuracy:** The response accurately describes Beethoven's father as a singer and his early exposure to music. - **Completeness:** It lacks mention of the challenges and hardships Beethoven faced.<br>- **Score:** 7/10 (Accurate but lacks depth)<br>6. **glm4_char:** - **Accuracy:** This response accurately describes Beethoven's father as strict and his mother as supportive. It also mentions Beethoven's connection to nature and his early compositions.<br>- **Completeness:** It provides a comprehensive view of both the nurturing and challenging aspects of Beethoven's childhood. - **Score:** 9/10 (Accurate and complete)<br>7. **qwen_role:** - **Accuracy:** The response inaccurately describes Beethoven's mother as a devoted teacher of piano and music theory, which is not historically accurate. - **Completeness:** It lacks details about Beethoven's father's harshness and the environment. - **Score:** 4/10 (Inaccurate and incomplete)<br>8. **qwen_char:** - **Accuracy:** The response accurately describes Beethoven's father as a musician with alcohol issues and his mother as supportive. - **Completeness:** It lacks details about Beethoven's environment and early musical influences. **Score:** 7/10 (Accurate but lacks detail) |