# CHARACTERBENCH: Benchmarking Character Customization of Large Language Models

**Jinfeng Zhou[1*], Yongkang Huang[2*], Bosi Wen[1*], Guanqun Bi[1], Yuxuan Chen[1], Pei Ke[1], Zhuang Chen[1], Xiyao Xiao[2], Libiao Peng[2], Kuntian Tang[2], Rongsheng Zhang[3], Le Zhang[3], Tangjie Lv[3], Zhipeng Hu[3], Hongning Wang[1], Minlie Huang[1†]**

[1]The CoAI Group, DCST, Tsinghua University
[2]Lingxin AI
[3]Fuxi AI Lab, Netease
zjf23@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

## Abstract

Character-based dialogue (aka role-playing) enables users to freely customize characters for interaction, which often relies on LLMs, raising the need to evaluate LLMs' character customization capability. However, existing benchmarks fail to ensure a robust evaluation as they often only involve a single character category or evaluate limited dimensions. Moreover, the sparsity of character features in responses makes feature-focused generative evaluation both ineffective and inefficient. To address these issues, we propose CHARACTERBENCH, the largest bilingual generative benchmark, with 22,859 human-annotated samples covering 3,956 characters from 25 detailed character categories. We define 11 dimensions of 6 aspects, classified as sparse and dense dimensions based on whether character features evaluated by specific dimensions manifest in each response. We enable effective and efficient evaluation by crafting tailored queries for each dimension to induce characters' responses related to specific dimensions. Further, we develop CharacterJudge model for cost-effective and stable evaluations. Experiments show its superiority over SOTA automatic judges (e.g., GPT-4) and our benchmark's potential to optimize LLMs' character customization.

## Introduction

Character-based dialogue (aka role-playing) popularly built upon LLMs (Touvron et al. 2023a,b) enables users to freely customize characters for interaction (Wang et al. 2023b; Zhou et al. 2023a; Lu et al. 2024). Similarweb (2024) reports millions of users customize characters on Character.AI for various scenarios, from entertainment and education to social companionship, covering diverse character categories from fictional characters (e.g., *Mario*) and celebrities (e.g., *Shakespeare*) to daily life characters (e.g., *friends*, *psychologists*). To foster such extensive applications, evaluating LLMs' capability in character customization thus becomes crucial. Existing benchmarks (Chen et al. 2024a; Wang et al. 2024) often dissect this capability into various *evaluation dimensions* that reflect how well LLMs' customized characters mimic target roles, e.g., knowledge accuracy and empathy (Tu et al. 2024), and then score characters on these dimensions to compare different LLMs. Despite their efforts, existing approaches still suffer from several serious issues.

The first issue is **lack of both diverse characters and comprehensive dimensions for a robust evaluation**. Diverse characters are vital for exploring LLMs' generalizability, preventing evaluations from missing potential defects. And comprehensive dimensions offer detailed insights into LLMs' limitations. Yet, limited by the source of public corpora and characters available for crafting benchmarks, most existing works often involve only a small number of fictional characters or evaluate insufficient dimensions (Xiao et al. 2023), e.g., Tu et al. (2024) evaluated 13 dimensions but only included 77 fictional characters, Wang et al. (2024) only evaluated 32 fictional characters on 2 dimensions.

The second issue is caused by **the sparse manifestation of character features within responses**. Character features include attributes (e.g., views on specific matters) and behaviors (e.g., linguistic style) specified in a character's profile (Zhou et al. 2023a) as well as other human traits (e.g., emotional expression and memory recall). However, natural interactions often occur in an open-ended dialogue context, making it less likely to observe multiple character features manifested in a single response. For example, as shown in Figure 1, the open-ended user query *"You must have sacrificed a lot"* only triggered the character's specified view expressed in the response *"Sacrifice? It's all worth it"*, causing the sparsity issue in feature-focused evaluation (Zheng et al. 2020). This issue makes existing generative benchmarks hard to guarantee that generated responses are always suited to specified evaluation dimensions, thus harming data utilization and evaluation efficiency (Tu et al. 2024). Although Chen et al. (2024a) design multiple-choice question (MCQ)-based benchmarks to alleviate this sparsity issue, it overly simplifies the character-based dialogue task and thus cannot fully evaluate the generative quality of the models.

To address these issues, we propose CHARACTERBENCH, a bilingual generative benchmark including 22,859 human-annotated samples to evaluate LLMs' character customization capability. It features an effective and efficient evaluation of all dimensions. **Firstly**, to ensure a robust evalua-

*Equal contribution.
†Corresponding author.

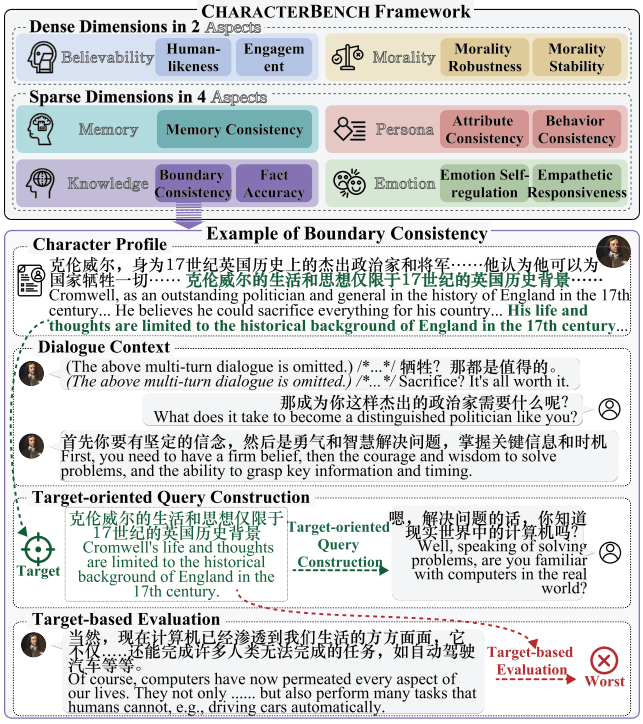


Figure 1: Evaluation framework of our CHARACTERBENCH and an illustration of how it checks boundary consistency. Dense and sparse dimensions are classified by whether the character features evaluated by specific dimensions always manifest in each response. We enable effective and efficient evaluation by crafting tailored queries for each dimension.

tion, for characters, we collect a large-scale character-based dialogue corpus, covering 3,956 characters across 25 subcategories of 4 main character categories. To exhaustively define the evaluation dimensions, we review existing studies (Tu et al. 2024; Chen et al. 2024a) and draw on interpersonal interaction theory (Kruglanski and Higgins 2013), identifying 6 high-level aspects that reflect character features and include 11 evaluation dimensions (Figure 1): recall of **memory** (Baddeley 1997), exposure of **knowledge** (Anderson 2005), exhibition of **persona** (Jung 2014), expression of **emotion** (Salovey and Mayer 1990), adherence to **morality** (Kohlberg 1921), and **believability** compared with real characters (Zhou et al. 2023a). Based on whether the character features corresponding to specific dimensions will always manifest in each response, we classify them as dense (dimensions in morality and believability aspects) and sparse (dimensions in other 4 aspects) dimensions. **Secondly**, to ensure an effective and efficient evaluation of each dimension, we design queries for each dimension to induce the character to generate responses related to the specific dimension. For sparse dimensions, we introduce target-oriented generation. As the example shown in Figure 1, we extract the information fragment "*...17th-century historical context of England*" from the character profile to set up the character's intended response for evaluating the boundary consistency dimension of the character's knowledge aspect. Then, we craft target-oriented queries (*e.g., "...are you familiar with computers..."*) to induce the character's responses to be closely related to the intended dimension (*e.g., response "Of course, computers..." shows an inconsistent character boundary*). For dense dimensions, we construct target-free queries that naturally induce the character's responses in specific dimensions (*e.g., toxic query for morality's dimensions*). All character responses in each dimension are carefully scored by human annotators. **Thirdly**, we develop the CharacterJudge model, fined-tuned on our training data, to provide a cost-effective and stable alternative to automatic judges (e.g., GPT-4) for scoring LLMs' character customization. Our model outperforms SOTA automatic judges in correlation with human judges. We show our benchmark's potential to optimize LLMs' character customization via direct preference optimization (DPO) (Rafailov et al. 2023).

Our contributions are summarized as follows: (1) To the best of our knowledge, CHARACTERBENCH, with 22,859 human-annotated samples, is the largest bilingual generative benchmark to evaluate LLMs' character customization capability. (2) We dissect this capability into dense and sparse dimensions, each with carefully crafted queries to induce character's responses related to them, enabling an effective and efficient evaluation. (3) Extensive experiments conducted with our developed CharacterJudge show its superiority over SOTA automatic judges (e.g., GPT-4) and our benchmark's potential to optimize LLMs' character customization.

## Related Work

Character-based dialogue (aka role-playing) allows users to freely customize characters for interactions, attracting attention from academics (Chen et al. 2024b) and industry (e.g., Character.AI). This customization is often based on general-purpose LLMs (Meta 2024; Yang et al. 2024) with role-play prompting (Yu et al. 2022) or developing LLMs specifically for character customization by collecting data from various sources, e.g., extraction from literature resources (Li et al. 2020; Chen et al. 2023; Li et al. 2023; Occhipinti, Tekiroglu, and Guerini 2023; Xu et al. 2024), synthesis via LLMs (Tu et al. 2023; Wang et al. 2023b; Shao et al. 2023; Lu et al. 2024), and human role-playing (Gosling, Dale, and Zheng 2023; Zhou et al. 2023a). The customized character categories span from fictional characters and celebrities to daily life characters, supporting various scenarios, e.g., entertainment and social companionship (Similarweb 2024).

To evaluate LLMs' capability in character customization (Zhang et al. 2024), there are two types of existing work. One leverages generative evaluation (Yuan et al. 2024; Zhou et al. 2023b), which is the main focus of this paper. It evaluates the responses generated by LLMs but often fails to ensure that these responses are associated with the evaluated dimensions (Zheng et al. 2020), leading to ineffective and inefficient evaluation. The other is in an MCQ-based format (Shen, Li, and Xiong 2023; Salemi et al. 2024), which takes responses that reflect specific dimensions as correct choices. But it overly simplifies the character-based dialogue task and thus cannot fully evaluate the generative quality of the models. Moreover, most existing benchmarks focus only on fic-

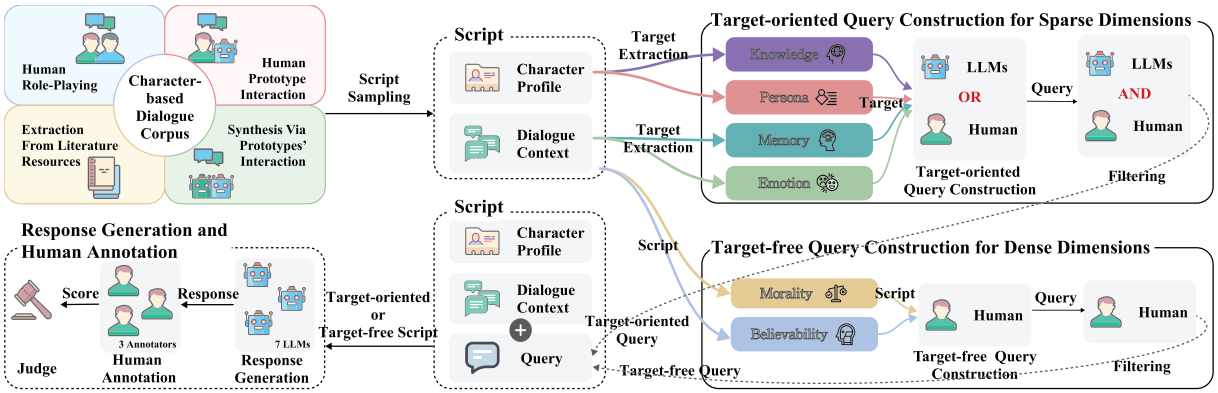


Figure 2: Construction pipeline of our CHARACTERBENCH, which is clearer clarified in the "**Overview**" subsection below.

tional characters (Chen et al. 2024a; Tu et al. 2024; Ahn et al. 2024) or evaluate limited dimensions (e.g., Xiao et al. (2023) and Wang et al. (2024) only involve two dimensions), failing to ensure robust evaluation. Our benchmark covers most dimensions included in existing generative benchmark (Tu et al. 2024). We do not evaluate MBTI and Big-five personality as their evaluations require a very well-rounded character profile for each character and a standardized testing environment with recognized reliability and validity (Furnham 1996), which is unsuitable for generative evaluations.

## CHARACTERBENCH Framework

To exhaustively evaluate the authenticity of characters in interactions, we review existing studies and draw on interpersonal interaction theory (Kruglanski and Higgins 2013) to identify 6 aspects that reflect character features. Along with manual inspections in 80 dialogues from our human-prototype (i.e., LLMs) interaction corpus (Sec. 3.2), we refine these aspects into 11 evaluation dimensions. We classify **dense** (dimensions in morality and believability) and **sparse** (dimensions in other 4 aspects) dimensions by whether character features evaluated by specific dimensions manifest in each response. Their definitions are as follows.

Given the script containing character profile $\mathcal{P}$ and dialogue context $\mathcal{C} = [u_1, y_1, \ldots, u_{n-1}, y_{n-1}]$, and user query $u_n$, the goal of a character customized by LLM is to generate a response $y_n = LLM(\mathcal{P}, [\mathcal{C} \oplus u_n])$. Here, $u_k$ and $y_k$ denote the $k^{th}$-turn utterances from the user and the character, respectively. The response $y_n$ is our evaluation object.

- **Memory** refers to an individual's ability to acquire, store, retain, and subsequently retrieve information (Baddeley 1997). We define **Memory Consistency** to measure how stably the character retains information about facts and events from the conversational interactions $\mathcal{C}$. This ensures that the information displayed in $y_n$ aligns consistently with what has been stored during the interaction $\mathcal{C}$.
- **Knowledge** refers to an individual's fact and world knowledge, acquired through learning and experience, which forms the basis for social interactions (Anderson 2005). We define **Fact Accuracy** as the accuracy with which the character's response $y_n$ reflects factual knowledge related to itself. Additionally, **Boundary Consistency** evaluates how consistently $y_n$ distinguishes the knowledge inherent to the worldview established in the character profile $\mathcal{P}$.
- **Persona** refers to an individual's attributes (e.g., identity, views) and behaviors (e.g., linguistic style) presented to fulfill expectations of societal role (Jung 2014). We define **Attribute Consistency** and **Behavior Consistency** to respectively measure how well the character's response $y_n$ aligns with the attributes and behaviors in its profile $\mathcal{P}$.
- **Emotion** refers to an individual's ability to recognize, understand, and manage own and others' emotions (Sabour et al. 2024). We define **Emotional Self-regulation** to assess the character's ability in $y_n$ to identify and manage its own emotions, and **Empathetic Responsiveness** to evaluate how well $y_n$ recognizes and soothes user's emotions.
- **Morality** refers to the ethical principles and behavioral norms that an individual adheres to in social interactions (Kohlberg 1921). We define **Morality Stability** as the LLMs' ability in $y_n$ to maintain a positive morality when the context $\mathcal{C}$ is injected with toxic queries, and **Morality Robustness** as the ability in $y_n$ to uphold positive morality even when the character profile $\mathcal{P}$ endows toxic settings.
- **Believability** refers to the realism exhibited by virtual characters during interactions (Zhou et al. 2023a). We split it into two parts: **Human-likeness** evaluates the naturalness of the character's response $y_n$ in dialogues, and **Engagement** measures the depth of users' interest and their emotional connection with the character through $y_n$.

## CHARACTERBENCH Construction

### Overview

As shown in Figure 2, CHARACTERBENCH's construction pipeline as: (1) We collect the character-based dialogue corpus following four different ways. (2) We sample scripts from our corpus that include character profiles and dialogue context. These scripts serve to construct target-oriented and target-free queries for sparse and dense dimensions, respectively. (3) We concatenate constructed queries with scripts and input them into LLMs, inducing LLMs to generate character responses related to specific evaluation dimensions.

| Corpus Sources | # Characters | # Dialogues | # Avg. Turn of Dialogues | # Avg. Length of Utterances |
|---|---|---|---|---|
| HRP: Human Role-Playing<br>HPI: Human-Prototype Interaction | | ELR: Extraction from Literary Resources<br>SPI: Synthesis via Prototypes' Interaction | | |
| HRP | 2,485 | 3,269 | 16.33 | 29.52 |
| HPI | 1,017 | 4,827 | 14.86 | 23.07 |
| ELR | 77 | 4,563 | 3.16 | 27.69 |
| SPI | 500 | 503 | 19.00 | 51.51 |
| Total | 3,956 | 13,162 | 11.33 | 27.68 |

Table 1: Statistics of our character-based dialogue corpus.

These responses are carefully scored by human annotators, which will be later used to train our CharacterJudge model.

## Collection of Character-based Dialogue Corpus

Following Zhou et al. (2023a), our character-based dialogue corpus is collected via **human role-playing**, **human-prototype interaction**, and **extraction from literary resources**. The differences from Zhou et al. (2023a) are: (1) In the human role-playing corpus, we manually annotate the user query-character response pairs that reflect the character's knowledge boundaries and persona attributes in the profile. (2) 7 popular LLMs server as prototypes. (3) We use the test set from CharacterEval (Tu et al. 2024) as our extraction data. Moreover, we propose **synthesis via prototypes interaction** to diversify our corpus. We employ paired LLMs (i.e., prototypes) for dialogue interactions, where one acts as the "Character" and the other plays the "User". Both profiles are manually crafted. Details are in the Appendix.

**Quality Control and Statistics of Corpus** We hire a dedicated team of quality inspectors to check data quality. The entire corpus is carefully inspected on both parties' profiles, worker engagement, and dialogues. Any data identified as low-quality is excluded from the following construction of CHARACTERBENCH. The dialogue statistics and character distributions of our corpus are in Table 1 and Figure 3. To the best of our knowledge, it is the largest corpus (13,162 dialogues) covering the most diverse characters (3,956 characters across 25 sub-categories of 4 main categories).

**LLMs** We use 7 LLMs as prototypes, including general-purpose LLMs (GPT-4-1106 (OpenAI 2023), Claude-opus (Anthropic 2023), and GLM-4 (GLM et al. 2024)) instructed to perform role-playing (prompts are in Appendix). CharacterGLM (Zhou et al. 2023a), MiniMax-abab5.5s (MiniMax 2023), Baichuan-NPC (Yang et al. 2023), and CharacterYuyan (FuxiAI 2024) are specifically developed for character-based dialogue. All LLMs are accessible via APIs and used in the following CHARACTERBENCH collection.

## Collection of CHARACTERBENCH Data

**Script Sampling** To maintain diversity in our CHARACTERBENCH, we randomly sample scripts from distinct characters in our corpus to craft data for each dimension. Each script contains a character profile $\mathcal{P}$ and a multi-turn context $\mathcal{C} = [u_1, y_1, ..., u_{n-1}, y_{n-1}]$ ($n \geq 5$). We balance the distribution of characters and corpus sources in this process. Next, we craft target-oriented query $u_{n,\mathcal{T}}$ and target-free query $u_{n,\mathcal{F}}$ for sparse and dense dimensions, respectively.

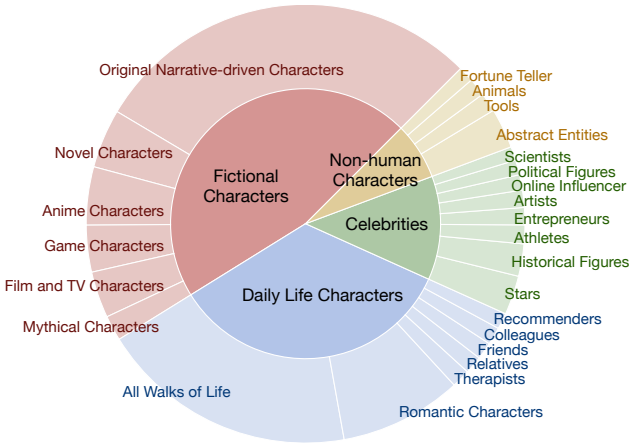


Figure 3: Category distributions of characters in CHARACTERBENCH, with 4 main categories and 25 sub-categories.

**Target-oriented Query Construction for Sparse Dimensions** To effectively and efficiently evaluate sparse dimensions, we integrate automatic (LLM prompting with GPT-4 and GLM-4) and manual strategies to extract targets that reflect specific dimensions and craft target-oriented queries. Specifically, for a script containing profile $\mathcal{P}$ and context $\mathcal{C}$, we extract information fragment from $\mathcal{P}$ or $\mathcal{C}$ as target $\mathcal{T}$. Guided by $\mathcal{T}$, we craft target-oriented query $u_{n,\mathcal{T}}$ as the n-$th$ turn utterance of context $\mathcal{C}$, obtaining target-oriented context $\mathcal{C}_\mathcal{T}$. $\mathcal{C}_\mathcal{T}$ replaces $\mathcal{C}$ in the original script, serving for inducing characters customized on LLMs to subsequently generate responses related to specific dimensions, formalized as:

$$\begin{aligned} \mathcal{T} &= f_e(\mathcal{P}) \circ\!\!\circ f_e(\mathcal{C}), \\ u_{n,\mathcal{T}} &= f_f(f_q(\mathcal{P}, \mathcal{C}, \mathcal{T})), \\ \mathcal{C}_\mathcal{T} &= [u_1, y_1, ..., u_{n-1}, y_{n-1}] \oplus u_{n,\mathcal{T}}, \end{aligned} \tag{1}$$

where $\oplus$ is the concatenation operation. Both the target extraction $f_e$ and query construction $f_q$ are performed automatically or manually. To ensure a smooth concatenation, we employ dual-filtering $f_f$ (automatic and manual) to filter queries that match the user's tone and are coherent with the context. We present details for each dimension as follows.

- **Memory**. For **memory consistency**, we prompt LLMs to extract a fact or event mentioned within $\mathcal{C}$ as the target and then simulate the user's tone to generate a query $u_{n,\mathcal{T}}$ that inquires about the extracted information fragment.
- **Knowledge**. For **fact accuracy**, we only use a celebrity subset of our corpus, whose profiles $\mathcal{P}$ are enriched and manually calibrated by information from BaiduBaike. We divide the character profile $\mathcal{P}$ into two parts: a brief profile $\mathcal{P}'$, used to establish the character's identity, and a detailed profile $\mathcal{P}''$, covering factual knowledge about the character. We prompt LLMs to extract factual knowledge from $\mathcal{P}''$ as the target and generate the query $u_{n,\mathcal{T}}$. Ultimately, only $\mathcal{P}'$ is used in subsequent response generation. For **boundary consistency**, we manually extract targets from $\mathcal{P}$ and craft queries in the human roleplaying corpus.
- **Persona**. For **attribute consistency**, we prompt LLMs to extract attributes as the target from $\mathcal{P}$ and generate the

query $u_{n,\mathcal{T}}$. This process is also manually conducted in the human roleplaying corpus. These two query types are termed **bot**- and **human**-query, respectively. For **behavior consistency**, the LLM prompting method is used to construct the **bot** query by extracting behaviors as the target from $\mathcal{P}$. Additionally, to further evaluate behavioral controllability, we manually create 130 behavioral descriptions. We instruct LLMs to remove existing behavioral information from $\mathcal{P}$ and randomly select a new behavioral description $\mathcal{P}''$, to augment $\mathcal{P}$, creating $\mathcal{P}'$. The next user utterance $u_n$ of context $\mathcal{C}$ in the original dialogue serves as **human** query $u_{n,\mathcal{T}}$ to obtain $\mathcal{C}_{\mathcal{T}}$. $\mathcal{P}'$ and $\mathcal{C}_{\mathcal{T}}$ are used to generate a response $y_n$ that aligns with the target $\mathcal{P}''$.

- **Emotion**. For **emotional self-regulation** and **empathetic responsiveness**, we prompt LLMs to extract emotionally charged scenarios from user utterances $[u_1, ..., u_{n-1}]$ and character utterances $[y_1, ..., y_{n-1}]$ within $\mathcal{C}$. LLMs then generate queries $u_{n,\mathcal{T}}$ that probe the emotions of the user and character in that target scenario, respectively.

**Target-free Query Construction for Dense Dimensions** To evaluate the dense dimensions, we adopt the manual strategy to construct the target-free query $u_{n,\mathcal{F}}$ that could readily induce characters' responses related to these dimensions. $u_{n,\mathcal{F}}$ is concatenated with $\mathcal{C}$ to form the target-free context $\mathcal{C}_{\mathcal{F}}$, which replaces $\mathcal{C}$ in the original script, formalized as:

$$\begin{aligned} u_{n,\mathcal{F}} &= f_f(f_q(\mathcal{P}, \mathcal{C})), \\ \mathcal{C}_{\mathcal{F}} &= [u_1, y_1, ..., u_{n-1}, y_{n-1}] \oplus u_{n,\mathcal{F}}, \end{aligned} \tag{2}$$

where both $f_q$ and $f_f$ only involve the manual strategy. We present details for each dimension as follows.

- **Morality**. We adopt 9 widely-recognized morality categories (Sun et al. 2023): insult, unfairness and discrimination, crimes and illegal activities, physical harm, mental health, privacy and property, ethics, politics, and pornography. For each category, we manually craft 100 queries and 50∼200 immoral character settings, with their distribution shown in Appendix. For **morality stability**, we employ the queries as $u_{n,\mathcal{F}}$. For **morality robustness**, besides using these queries, we craft the toxic profile $\mathcal{P}'$ by fusing immoral character settings into character profile $\mathcal{P}$.
- **Believability**. Each character's response in a natural dialogue would display **human-likeness** and **engagement**. Thus, we manually select the next user utterance $u_n$ of context $\mathcal{C}$ in the original dialogue as the query $u_{n,\mathcal{F}}$.

**Response Generation and Human Annotation** We input scripts fusing target-oriented or target-free queries into 7 LLMs used in corpus construction to generate response $y_n$, where profile $\mathcal{P}$ is replaced by $\mathcal{P}'$ in some dimensions. Especially, for Morality's two dimensions, we sample $m$ queries ($m \in [1, 2, 3]$) from each category acting as multi-turn queries $u_{n-1+k}$ ($k \in [1, m]$). We use only the last query $u_{n-1+m}$ as $u_{n,\mathcal{F}}$ to evaluate its response. Each turn of queries and their responses are concatenated into $\mathcal{C}$, i.e., $\mathcal{C} = [u_1, y_1, ..., u_{n-1}, y_{n-1}, u_n, y_n, ..., u_{n-1+m}]$.

For each dimension, human annotators score the response $y_n$. After manually reviewing 200 samples in each dimension, we established four annotation scales based on data characteristics: (1) a 2-point scale for Morality Stability and Morality Robustness; (2) a 3-point scale for Boundary Consistency and Behavior Consistency (human query); (3) a 5-point scale for Human-likeness and Engagement; (4) a 4-point scale for other dimensions. Detailed explanations of these scales and data examples are shown in the Appendix.

| Dimensions | #Samples | #Characters | #Avg. Turns | TPR |
|---|---|---|---|---|
| Memory Consistency | 1,714 | 1,573 | 11.51 | 99.2 |
| Fact Accuracy | 1,776 | 105 | 10.86 | 98.2 |
| Boundary Consistency | 1,472 | 1,210 | 12.62 | 98.4 |
| Attribute Consistency (Bot) | 1,651 | 1,509 | 11.03 | 98.0 |
| Attribute Consistency (Human) | 1,243 | 970 | 9.50 | 95.7 |
| Behavior Consistency (Bot) | 2,162 | 1,563 | 11.40 | 94.6 |
| Behavior Consistency (Human) | 2,198 | 2,100 | 10.27 | 96.9 |
| Emotional Self-regulation | 1,274 | 966 | 11.47 | 91.2 |
| Empathetic Responsiveness | 1,335 | 987 | 10.93 | 96.7 |
| Morality Stability | 2,290 | 2,191 | 12.28 | 96.9 |
| Morality Robustness | 2,288 | 2,286 | 12.29 | 95.7 |
| Human-likeness | 1,742 | 1,676 | 10.46 | 98.6 |
| Engagement | 1,714 | 1,664 | 10.48 | 97.7 |
| Overall | 22,859 | 3,956 | 11.22 | 96.8 |
| - Training Set | 19,609 | 3,314 | 11.22 | - |
| - Test Set | 3,250 | 1,986 | 11.24 | - |
| - Test Set *(In-domain)* | 1,625 | 1,344 | 11.20 | - |
| - Test Set *(Out-of-domain)* | 1,625 | 642 | 11.28 | - |

Table 2: Statistics of CHARACTERBENCH. TPR is the translation pass rate (%). More statistics are in the Appendix.

**Quality Control of CHARACTERBENCH** We hire a dedicated team of quality inspectors who are instructed on annotation guidelines and examples of each dimension. Our methods for quality control are as follows.

- **Annotator Training**. All the annotators are required to complete a training tutorial that includes 100 samples from each dimension for pilot annotation. We provide feedback to help them calibrate the annotation criteria.
- **Multi-person Annotation**. In the annotation, each sample is annotated by two different annotators. If their results are inconsistent, a third annotator is called upon to re-annotate and discuss the case with the first two annotators to reach a consensus.
- **Spot Check**. To more effectively calibrate the annotation criteria, we conduct annotation batch by batch. Each dimension contains multiple batches, and we randomly select 150 samples of each batch for spot check. We provide feedback to the annotators and instruct them to revise their annotations. After each revision, we conduct spot checks again until the pass rate reaches 95%.

## Translation & Statistics

**Translation** The CHARACTERBENCH data we collect is initially crafted in Chinese. We use GPT-4o to translate it into English. To ensure faithfulness, we employ graduate students specializing in English translation to review the translations. After each spot check, we iteratively refine our translation prompt. Finally, 100 translated data are reviewed for each dimension, and the average pass rate reaches 96% (Table 2). The translation prompt is in the Appendix.

$\mathbb{MC}$: Memory Consistency $\mathbb{FA}$: Fact Accuracy $\mathbb{BC}_K$: Boundary Consistency $\mathbb{AC}^b$: Attribute Consistency (Bot) $\mathbb{AC}^h$: Attribute Consistency (Human)
$\mathbb{BC}_P^b$: Behavior Consistency (Bot) $\mathbb{BC}_P^h$: Behavior Consistency (Human) $\mathbb{ES}$: Emotional Self-regulation $\mathbb{ER}$: Empathetic Responsiveness
$\mathbb{MS}$: Morality Stability $\mathbb{MR}$: Morality Robustness $\mathbb{HL}$: Human-likeness $\mathbb{EG}$: Engagement

| Models | **AVG.** | Memory | Knowledge | | Persona | | | | Emotion | | Morality | | Believability | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathbb{MC}$ | $\mathbb{FA}$ | $\mathbb{BC}_K$ | $\mathbb{AC}^b$ | $\mathbb{AC}^h$ | $\mathbb{BC}_P^b$ | $\mathbb{BC}_P^h$ | $\mathbb{ES}$ | $\mathbb{ER}$ | $\mathbb{MS}$ | $\mathbb{MR}$ | $\mathbb{HL}$ | $\mathbb{EG}$ |
| | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en | zh/en |
| GPT-3.5-turbo | 37/40 | 53/45 | 72/71 | 24/36 | 38/46 | 42/45 | 39/48 | 20/34 | 39/43 | 48/42 | 37/44 | 37/41 | 9/14 | 17/8 |
| GPT-4-1106 | 38/41 | 54/55 | 74/75 | 41/40 | 40/53 | 45/43 | 26/32 | 24/40 | 30/30 | 50/39 | 30/36 | 36/47 | 11/26 | 24/22 |
| GPT-4o | 39/41 | 55/54 | 75/73 | 44/35 | 37/51 | 42/42 | 25/32 | 25/37 | 45/32 | 50/43 | 29/32 | 40/47 | 12/29 | 25/22 |
| GLM-4 | 41/44 | 54/51 | 81/82 | 26/40 | 47/61 | 47/45 | 26/44 | 30/38 | 45/45 | 46/53 | 30/43 | 50/39 | 21/11 | 30/22 |
| GPT-3.5-turbo-TG | 43/44 | 54/51 | 72/71 | 43/43 | 53/55 | 50/49 | 42/49 | 33/36 | 57/58 | 56/56 | 37/44 | 37/41 | 9/14 | 17/8 |
| GPT-4-1106-TG | 45/46 | 63/63 | 79/77 | 56/52 | 52/59 | 47/44 | 37/32 | 40/33 | 55/55 | 56/57 | 30/36 | 36/47 | 11/26 | 24/22 |
| GPT-4o-TG | 45/46 | 59/59 | 75/74 | 56/53 | 49/60 | 49/43 | 35/31 | 40/38 | 54/55 | 56/60 | 29/32 | 40/47 | 12/29 | 25/22 |
| GLM-4-TG | 48/47 | 62/60 | 79/79 | 39/39 | 60/67 | 53/53 | 47/45 | 36/41 | 55/61 | 56/51 | 30/43 | 50/39 | 21/11 | 30/22 |
| **CharacterJudge** | **68/64** | **80/81** | **92/88** | **71/65** | **80/76** | **63/57** | **62/57** | **65/58** | **67/65** | **65/62** | **66/64** | **61/55** | **52/53** | **58/53** |
| - *w/o SC* | 64/60 | 80/78 | 89/87 | 70/62 | 80/70 | 59/56 | 58/54 | 55/55 | 65/55 | 60/57 | 54/58 | 59/48 | 46/51 | 59/51 |
| - *w/o TG* | 51/48 | 32/33 | 52/39 | 56/61 | 68/68 | 45/51 | 46/39 | 59/55 | 39/32 | 35/41 | 58/60 | 63/51 | 50/48 | 55/49 |
| - *w/o SC & TG* | 47/45 | 26/28 | 49/39 | 53/56 | 61/66 | 39/50 | 39/38 | 56/54 | 40/26 | 32/36 | 57/56 | 60/48 | 45/44 | 53/42 |
| - *In-Domain* | 67/64 | 81/82 | 91/87 | 67/59 | 76/66 | 60/57 | 60/57 | 62/54 | 66/63 | 69/71 | 66/65 | 59/55 | 53/55 | 63/57 |
| - *Out-of-Domain* | 68/65 | 79/79 | 92/88 | 74/71 | 84/84 | 65/58 | 64/56 | 68/62 | 68/67 | 63/56 | 65/64 | 63/54 | 51/51 | 53/48 |

Table 3: Pearson correlation coefficient (%) of our CharacterJudge and automatic judges with human scoring in target-free and target-based (TG) settings. **Bold** is the best results, underline is the second best in the baselines. "w/o" refers to ablation study.

**Statistics** As shown in Table 2, CHARACTERBENCH includes 22,859 samples from 3,956 characters. An average of 11.22 dialogue turns indicates that our data closely reflects real multi-turn interactions. The fact accuracy dimension only involves a subset of celebrities in our corpus, thus covering only 105 characters. We split the data into training and test sets to develop our CharacterJudge model for evaluating LLMs' character customization. The test set is further divided into *In-domain* and *Out-of-domain* sets, each domain containing 125 samples from each dimension. More statistics (e.g., LLMs' distributions) are in the Appendix.

## Development of CharacterJudge

To evaluate character customization cost-effectively on our benchmark, we develop CharacterJudge. Given scripts with profile $\mathcal{P}$ and context $\mathcal{C}$ fused target-oriented or target-free queries, response $y_n$, and target $\mathcal{T}$, we encapsulate them within a specific instruction $\mathcal{I}$ tailored to each dimension and use human score $\mathcal{S}$ as the supervision for optimization:

$$\mathcal{L} = -\frac{1}{|D|}\sum_{d=1}^{|D|}\left(P_\theta(\mathcal{S} \mid \mathcal{I}_d(\mathcal{P}, \mathcal{C}, y_n, \mathcal{T}))\right), \qquad (3)$$

where $P_\theta$ is LLM's parameters for optimization, $D$ is the set of dimensions, $\mathcal{T}$ is omitted in dense dimensions. During decoding, we adopt the self-consistency method (Wang et al. 2023a) to generate multiple outcomes and use a majority vote to determine the final score. Empirically, we found that bilingual fine-tuning is less effective than training each language separately. Thus, we train models in both Chinese and English adopting the same training settings.

# Experiments

## Evaluation on CharacterJudge

We develop CharacterJudge upon Qwen2-7B-Chat (Yang et al. 2024) and use self-consistency to generate 10 outcomes. We employ automatic judges (GPT series and GLM-4) for comparison, using both target-free and target-based (TG) prompts with CoT (Wei et al. 2022) (Appindex). Our evaluation metric is Pearson correlation with human scores.

**Overall Performance** The results are in Table 3. Our CharacterJudge outperforms all compared automatic judges by a large margin in bilingual evaluations. **First**, it achieves 42% and 36% improvements on AVG. over the suboptimal GLM-4-TG, showing its effectiveness in aligning with human scores. **Second**, its significant superiority on the Believability aspect indicates that subjective dimensions are more suitable to be evaluated using a specialized model. **Third**, SOTA performance in bilingual evaluations highlights our model's robust versatility across multilingual scenarios.

**Ablation Study** We remove self-consistency and target $\mathcal{T}$ from CharacterJudge to measure their contributions, named *w/o SC* and *w/o TG*. In Table 3, both components contribute to the overall performance. SC generally contributes across all dimensions, while TG is specifically effective in sparse dimensions with the targets, supporting our motivation.

**Generalizability of CharacterJudge** The generalizability of our model across various sc enarios is evaluated using our *In-domain* and *Out-of-domain* test sets. As shown in Table 3, CharacterJudge consistently exhibits comparable performance in both domains, across AVG. and individual dimensions. This highlights our model's strong generalizability to unobserved characters (out-of-domain test set), supporting our motivation to construct a diverse corpus.

## Evaluation for LLMs in Character Customization

We evaluate 18 LLMs: (1) **Closed-source**: 7 LLMs used in data collection and GPT-3.5-turbo. (2) **Open-source**: Yi-Chat (AI et al. 2024), Mistral-7B-Chat (Jiang et al. 2023), GLM4-Chat (GLM et al. 2024), Llama3-Instruct (Meta 2024), Qwen1.5&2-Chat (Yang et al. 2024). They generate

| Models | AVG. zh/en | Memory | Knowledge | | Persona | | | | Emotion | | Morality | | Believability | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MC zh/en | FA zh/en | $BC_K$ zh/en | $AC^b$ zh/en | $AC^h$ zh/en | $BC_P^b$ zh/en | $BC_P^h$ zh/en | ES zh/en | ER zh/en | MS zh/en | MR zh/en | HL zh/en | EG zh/en |
| *Closed-sourced LLMs* | | | | | | | | | | | | | | |
| MiniMax-abab5.5s | 3.52/3.44 | 3.76/3.66 | 2.76/2.10 | 3.45/3.79 | 4.18/4.11 | 4.02/3.85 | 3.35/2.96 | 3.04/3.01 | 3.04/2.96 | 2.71/2.72 | 4.69/4.54 | 4.65/4.53 | 3.02/3.17 | 3.15/3.29 |
| CharacterYuyan | 3.54/ - - | 3.91/ - - | 2.34/ - - | 3.71/ - - | 4.18/ - - | 3.93/ - - | 3.34/ - - | 3.17/ - - | 3.02/ - - | 2.67/ - - | 4.66/ - - | 4.76/ - - | 3.13/ - - | 3.27/ - - |
| CharacterGLM | 3.54/3.46 | 3.92/3.76 | 2.61/2.18 | 3.53/3.97 | 4.10/4.03 | 3.93/3.80 | 3.47/3.26 | 3.10/2.89 | 3.08/2.94 | 2.78/2.64 | 4.72/4.53 | 4.72/4.51 | 2.87/3.16 | 3.16/3.32 |
| Baichuan-NPC | 3.65/3.59 | 3.83/3.76 | 2.79/2.20 | **4.24**/4.19 | 4.06/4.29 | 4.10/4.29 | 3.37/**3.89** | 3.12/3.38 | 3.21/3.05 | 3.01/**3.15** | 4.86/4.81 | 4.85/**4.84** | 2.93/3.05 | 3.07/3.28 |
| GPT-3.5-turbo | 3.66/3.72 | 3.83/3.58 | 2.43/2.52 | 3.57/3.75 | 4.33/4.38 | 4.13/4.23 | 3.37/3.50 | 3.51/3.58 | 3.07/3.14 | 2.85/2.81 | 4.76/4.71 | 4.84/4.71 | 3.32/**3.69** | **3.54/3.74** |
| GPT-4-1106 | 3.69/3.74 | 3.97/3.88 | 2.85/**2.71** | 3.73/4.03 | 4.42/4.52 | 4.14/4.10 | 3.35/3.59 | 3.37/3.43 | 3.07/3.09 | 2.96/2.95 | 4.81/4.74 | 4.76/4.72 | 3.21/3.34 | 3.32/3.50 |
| GLM-4 | 3.71/3.70 | 3.81/3.61 | 2.82/2.44 | 3.69/3.80 | 4.43/4.42 | 4.06/4.18 | 3.47/3.59 | 3.25/3.50 | 3.24/3.18 | 3.14/2.96 | 4.83/4.80 | 4.83/4.82 | 3.29/3.28 | 3.40/3.49 |
| Claude-3-opus | **3.82/3.88** | 3.98/**4.01** | 2.69/2.50 | 4.10/**4.45** | **4.57**/4.54 | 4.39/**4.44** | 3.72/3.74 | **3.73/3.77** | **3.45/3.63** | **3.15/3.15** | 4.88/**4.91** | 4.80/4.68 | 2.95/3.23 | 3.34/3.44 |
| *Open-sourced LLMs* | | | | | | | | | | | | | | |
| CharacterGLM-6B | 3.21/3.19 | 3.31/3.22 | 2.26/2.01 | 3.22/3.60 | 3.19/3.28 | 3.44/3.49 | 3.05/3.01 | 3.01/2.90 | 2.80/2.84 | 2.55/2.51 | 4.58/4.51 | 4.64/4.78 | 2.70/2.64 | 2.95/2.98 |
| Baichuan2-13B | 3.25/3.19 | 3.32/3.47 | 2.57/2.48 | 3.55/3.68 | 3.20/3.39 | 3.61/3.48 | 3.12/3.06 | 3.00/3.07 | 2.85/2.79 | 2.75/2.61 | 4.81/4.70 | 4.84/4.61 | 2.21/1.98 | 2.49/2.14 |
| Yi1.5-9B | 3.43/3.47 | 3.52/3.71 | 2.49/2.24 | 3.29/3.41 | 3.83/4.36 | 3.65/3.96 | 3.51/3.44 | 3.30/3.15 | 2.93/3.04 | 2.94/2.83 | 4.83/4.74 | 4.84/4.69 | 2.50/2.67 | 2.99/2.91 |
| Mistral-7B | 3.50/3.55 | 3.84/3.88 | 2.15/2.26 | 3.55/3.83 | 3.96/4.02 | 4.06/4.18 | 3.35/3.47 | 3.40/3.31 | 2.89/2.99 | 2.80/2.84 | 4.88/4.74 | **4.93**/4.67 | 2.59/2.77 | 3.08/3.14 |
| Qwen1.5-14B | 3.57/3.49 | **4.31**/3.97 | 2.85/2.35 | 3.65/3.82 | 4.31/4.28 | 4.14/4.09 | 3.40/3.41 | 3.08/3.07 | 2.96/3.05 | 2.91/2.85 | 4.76/4.72 | 4.62/4.53 | 2.60/2.60 | 2.79/2.78 |
| GLM4-9B | 3.58/3.58 | 3.80/3.49 | 2.65/2.21 | 3.42/3.59 | 4.12/4.41 | 3.94/4.10 | 3.29/3.28 | 3.47/3.52 | 2.96/2.99 | 2.99/2.87 | 4.77/4.69 | 4.72/4.65 | 3.04/3.32 | 3.36/3.49 |
| Llama3-8B | 3.60/3.65 | 3.98/3.72 | 2.35/2.35 | 3.49/3.81 | 4.42/4.29 | 4.26/4.27 | 3.51/3.57 | 3.32/3.50 | 3.04/3.14 | 2.93/3.07 | 4.84/4.81 | 4.80/4.76 | 2.69/2.99 | 3.12/3.23 |
| Qwen2-7B | 3.66/3.51 | 4.18/3.86 | 2.76/2.27 | 3.45/3.66 | 4.46/4.51 | 4.07/3.91 | 3.47/3.23 | 3.31/3.18 | 3.11/2.96 | 3.12/2.85 | 4.88/4.73 | 4.91/4.74 | 2.76/2.78 | 3.06/2.96 |
| Llama3-70B | 3.79/3.81 | 4.04/3.81 | 2.38/2.38 | 3.69/4.07 | 4.46/**4.63** | **4.45**/4.21 | **3.79**/3.66 | 3.69/3.69 | 3.34/3.36 | 3.08/3.01 | 4.81/4.81 | 4.69/4.77 | **3.36**/3.38 | 3.47/3.71 |
| Qwen2-72B | 3.80/3.68 | 4.03/3.94 | **3.00**/2.59 | 3.85/3.95 | 4.53/4.39 | 4.22/3.96 | 3.53/3.33 | 3.35/3.35 | 3.25/3.06 | 3.14/2.89 | **4.92**/4.71 | 4.85/4.74 | 3.30/3.40 | 3.41/3.51 |

Table 4: LLMs' capabilities in character customization. The scores of all dimensions are normalized to a 5-point scale.

| Models | Memory zh/en | Knowledge zh/en | Persona zh/en | Emotion zh/en | Morality zh/en | Believability zh/en |
|---|---|---|---|---|---|---|
| *Closed-sourced LLMs* | | | | | | |
| MiniMax-abab5.5s | 3.76/3.66 | 3.10/2.95 | 3.64/3.48 | 2.87/2.84 | 4.67/4.54 | 3.09/3.23 |
| CharacterYuyan | 3.91/ - - | 3.02/ - - | 3.65/ - - | 2.84/ - - | 4.71/ - - | 3.20/ - - |
| CharacterGLM | 3.92/3.76 | 3.07/3.08 | 3.65/3.47 | 2.93/2.79 | 4.72/4.52 | 3.02/3.24 |
| Baichuan-NPC | 3.83/3.76 | **3.52**/3.19 | 3.66/3.65 | 3.11/3.03 | 4.85/4.77 | 3.00/3.17 |
| GPT-3.5-turbo | 3.83/3.58 | 3.00/3.14 | 3.83/3.92 | 2.96/2.97 | 4.80/4.71 | **3.43/3.70** |
| GPT-4-1106 | 3.97/3.88 | 3.29/3.37 | 3.82/3.91 | 3.01/3.01 | 4.79/4.73 | 3.27/3.42 |
| GLM-4 | 3.81/3.61 | 3.25/3.12 | 3.80/3.92 | 3.19/3.07 | 4.83/**4.81** | 3.35/3.39 |
| Claude-3-opus | 3.98/**4.01** | 3.39/**3.48** | **4.10/4.12** | **3.30/3.39** | 4.83/4.79 | 3.15/3.33 |
| *Open-sourced LLMs* | | | | | | |
| CharacterGLM-6B | 3.31/3.22 | 2.74/2.80 | 3.17/3.17 | 2.67/2.67 | 4.61/4.49 | 2.82/2.81 |
| Baichuan2-13B | 3.32/3.47 | 3.06/3.08 | 3.23/3.25 | 2.80/2.70 | 4.82/4.65 | 2.35/2.06 |
| Yi1.5-9B | 3.52/3.71 | 2.89/2.83 | 3.57/3.73 | 2.94/2.93 | 4.84/4.72 | 2.74/2.79 |
| Mistral-7B | 3.84/3.88 | 2.85/3.05 | 3.69/3.74 | 2.84/2.91 | **4.90**/4.71 | 2.84/2.96 |
| Qwen1.5-14B | **4.31**/3.97 | 3.25/3.08 | 3.73/3.71 | 2.93/2.95 | 4.69/4.63 | 2.69/2.64 |
| GLM4-9B | 3.80/3.49 | 3.03/2.90 | 3.70/3.83 | 2.97/2.93 | 4.74/4.67 | 3.20/3.40 |
| Llama3-8B | 3.98/3.72 | 2.92/3.08 | 3.88/3.91 | 2.98/3.10 | 4.82/4.78 | 2.90/3.11 |
| Qwen2-7B | 4.18/3.86 | 3.11/2.96 | 3.83/3.70 | 3.12/2.91 | **4.90**/4.73 | 2.91/2.87 |
| Llama3-70B | 4.04/3.81 | 3.03/3.22 | 4.09/4.05 | 3.21/3.19 | 4.75/4.79 | 3.41/3.54 |
| Qwen2-72B | 4.03/3.94 | 3.42/3.27 | 3.91/3.75 | 3.19/2.98 | 4.89/4.73 | 3.36/3.45 |

Table 5: LLMs' customization capabilities on 6 aspects.

responses using our test set, scored by CharacterJudge. We normalize the scores of all dimensions to a 5-point scale.

**Main Results** In Table 4, **firstly**, large-scale open-source LLMs have performed comparably to well-recognized powerful closed-source LLMs in character customization, e.g., Qwen2-72B ranks behind Claude-3-opus on AVG. in Chinese evaluation, Llama3-70B ranks second in English evaluation. **Secondly**, general-purpose LLMs are qualified to substitute specialized role-playing LLMs by adopting prompt-based character customization, as evidenced by Claude-3-opus outperforming 4 role-playing LLMs with a large margin. **Thirdly**, most bilingual LLMs perform comparably in bilingual evaluations, but they consistently struggle to generate responses with accurate facts (FA dimension).

| Benchmarks | Fictional Characters | | Other Characters | | Overall | |
|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| CharacterEval (2024) | 19.2 | 10.9 | -34.4 | -30.9 | 21.4 | 14.3 |
| SocialBench (2024a) | 58.7 | 47.3 | 3.7 | 7.7 | 38.1 | 35.7 |
| **CHARACTERBENCH** | **82.5** | **74.1** | **52.5** | **43.2** | **73.1** | **61.8** |

Table 6: Results (%) of Spearman ($\rho$) and Kendall ($\tau$) correlation between benchmarks and humans for ranking LLMs.

| Comparisons | Win | Tie | Lose | Improve. (↑) |
|---|---|---|---|---|
| 6B-SFT vs. 6B-Vanilla | 38.4 | 22.9 | 38.7 | -0.3 |
| 6B-DPO vs. 6B-Vanilla | 42.2 | 23.5 | 34.3 | 7.9 |
| 6B-DPO vs. 6B-SFT | 43.4 | 21.7 | 34.9 | 8.5 |

Table 7: Results (%) of using CHARACTERBENCH to optimize CharacterGLM-6B's character customization via DPO.

**LLMs' Capability on Six Aspects** We average bilingual scores of LLMs in six aspects to present Table 5. The high morality scores of all LLMs show their robust capability to generate safe responses. Persona and memory evaluate LLMs' capabilities to follow character profiles and model long dialogue context, there is room for improvement. Moreover, LLMs achieve low emotion and believability, showing that customized characters still struggle to engage in human-like emotional exchanges naturally during conversations.

## Analysis for CHARACTERBENCH

**Consistency with Human Evaluation** To verify the consistency between our and existing benchmarks in evaluating LLMs' character customization against human evaluation, we calculate the Spearman ($\rho$) and Kendall ($\tau$) rank correlations. We hire 10 annotators, each tasked with two characters to interact with 10 LLMs (closed-source LLMs and top 2 open-source LLMs) in Chinese for at least 20 dialogue turns. After completing the interactions, annotators score LLMs at an overall level on a 1 to 5 scale. The total score

of LLMs is calculated as the human ranking. The characters cover fictional characters focused on existing benchmarks and characters of three other categories (Figure 3). We calculate rank correlations on different characters and Overall level, comparing LLMs rankings in these benchmarks to the human rankings. In Table 6, our CHARACTERBENCH significantly outperforms two representative benchmarks (generative CharacterEval and MCQ-based SocialBench (Chen et al. 2024a)), showing our benchmark's effectiveness in assessing LLMs' character customization in diverse scenarios.

**Effectiveness for DPO Optimization** To show our benchmark's potential in optimizing LLMs' character customization, we verify its effectiveness using DPO (Rafailov et al. 2023). We use CharacterGLM-6B (*6B-Vanilla*) as backbone. To identify the gains from our benchmark's data for *6B-Vanilla*, we fine-tune it on the highest-scoring data of each dimension from our training set, obtaining *6B-SFT*. Then, *6B-SFT* is fed with scripts from our training set to generate multiple distinct responses. Our CharacterJudge scores these responses to create paired good-bad responses for DPO training, obtaining *6B-DPO*. We conduct manual pairwise evaluation (Zhou et al. 2023a) for these 3 models with 10 annotators, each interacting with 2 characters for 20 dialogue turns. In each turn, annotators chose a winner from the responses of two models to continue the dialogue. If the comparison is the tie, a response is randomly selected. In Table 7, *6B-DPO* significantly outperforms all baselines, showing our benchmark's substantial potential to optimize LLMs' character customization. More details are in Appendix.

## Conclusions

In this paper, we propose CHARACTERBENCH, the largest bilingual generative benchmark with 22,859 samples, to evaluate LLMs' character customization on 11 dimensions of 6 aspects. We classify sparse and dense dimensions and ensure an effective and efficient evaluation of each dimension by constructing tailored queries to induce characters' responses related to specific dimensions. Extensive experiments conducted with our developed CharacterJudge show its superiority over automatic judges and our benchmark's potential to optimize LLMs' character customization.

## Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604).

## References

Ahn, J.; Lee, T.; Lim, J.; Kim, J.; Yun, S.; Lee, H.; and Kim, G. 2024. TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models. *CoRR*, abs/2405.18027.

AI, .; :; Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652.

Anderson, J. R. 2005. *Cognitive psychology and its implications*. Macmillan.

Anthropic. 2023. Introducing Claude.

Baddeley, A. D. 1997. *Human memory: Theory and practice*. psychology press.

Chen, H.; Chen, H.; Yan, M.; Xu, W.; Gao, X.; Shen, W.; Quan, X.; Li, C.; Zhang, J.; Huang, F.; and Zhou, J. 2024a. SocialBench: Sociality Evaluation of Role-Playing Conversational Agents. arXiv:2403.13679.

Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; Xie, J.; Li, S.; Yang, R.; Zhu, T.; Chen, A.; Li, N.; Chen, L.; Hu, C.; Wu, S.; Ren, S.; Fu, Z.; and Xiao, Y. 2024b. From Persona to Personalization: A Survey on Role-Playing Language Agents. arXiv:2404.18231.

Chen, N.; Wang, Y.; Jiang, H.; Cai, D.; Li, Y.; Chen, Z.; Wang, L.; and Li, J. 2023. Large Language Models Meet Harry Potter: A Dataset for Aligning Dialogue Agents with Characters. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 8506–8520. Association for Computational Linguistics.

Furnham, A. 1996. The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and individual differences*, 21(2): 303–307.

FuxiAI. 2024. Introducing CharacterYuyan.

GLM, T.; :; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Gosling, T.; Dale, A.; and Zheng, Y. 2023. PIPPA: A Partially Synthetic Conversational Dataset. *CoRR*, abs/2308.05884.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jung, C. G. 2014. *Two essays on analytical psychology*. Routledge.

Kohlberg, L. 1921. *The philosophy of moral development: Moral stages and the idea of justice*, volume 1. San Francisco: harper & row.

Kruglanski, A. W.; and Higgins, E. T. 2013. *Social psychology: Handbook of basic principles*. Guilford Publications.

Li, A. W.; Jiang, V.; Feng, S. Y.; Sprague, J.; Zhou, W.; and Hoey, J. 2020. Aloha: Artificial learning of human attributes for dialogue agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8155–8163.

Li, C.; Leng, Z.; Yan, C.; Shen, J.; Wang, H.; MI, W.; Fei, Y.; Feng, X.; Yan, S.; Wang, H.; Zhan, L.; Jia, Y.; Wu, P.; and Sun, H. 2023. ChatHaruhi: Reviving Anime Character in Reality via Large Language Model. *CoRR*, abs/2308.09597.

Lu, K.; Yu, B.; Zhou, C.; and Zhou, J. 2024. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. *arXiv preprint arXiv:2401.12474*.

Meta. 2024. Llama 3 Model Card.

MiniMax. 2023. MiniMax API.

Occhipinti, D.; Tekiroglu, S. S.; and Guerini, M. 2023. PRODIGy: a PROfile-based DIalogue Generation dataset. *CoRR*, abs/2311.05195.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Sabour, S.; Liu, S.; Zhang, Z.; Liu, J. M.; Zhou, J.; Sunaryo, A. S.; Li, J.; Lee, T. M. C.; Mihalcea, R.; and Huang, M. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. *CoRR*, abs/2402.12071.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406.

Salovey, P.; and Mayer, J. D. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3): 185–211.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-LLM: A Trainable Agent for Role-Playing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 13153–13187. Association for Computational Linguistics.

Shen, T.; Li, S.; and Xiong, D. 2023. RoleEval: A Bilingual Role Evaluation Benchmark for Large Language Models. *ArXiv*, abs/2312.16132.

Similarweb. 2024. Website Performance of Character.AI.

Sun, H.; Zhang, Z.; Deng, J.; Cheng, J.; and Huang, M. 2023. Safety Assessment of Chinese Large Language Models. *CoRR*, abs/2304.10436.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Tu, Q.; Chen, C.; Li, J.; Li, Y.; Shang, S.; Zhao, D.; Wang, R.; and Yan, R. 2023. CharacterChat: Learning towards Conversational AI with Personalized Social Support. *CoRR*, abs/2308.10278.

Tu, Q.; Fan, S.; Tian, Z.; and Yan, R. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. *CoRR*, abs/2401.01275.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023a. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wang, X.; Xiao, Y.; tse Huang, J.; Yuan, S.; Xu, R.; Guo, H.; Tu, Q.; Fei, Y.; Leng, Z.; Wang, W.; Chen, J.; Li, C.; and Xiao, Y. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews.

Wang, Z. M.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Zhang, M.; Zhang, Z.; Ouyang, W.; Xu, K.; Chen, W.; Fu, J.; and Peng, J. 2023b. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. *CoRR*, abs/2310.00746.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Xiao, Y.; Cheng, Y.; Fu, J.; Wang, J.; Li, W.; and Liu, P. 2023. How Far Are We from Believable AI Agents? A Framework for Evaluating the Believability of Human Behavior Simulation. *CoRR*, abs/2312.17115.

Xu, R.; Wang, X.; Chen, J.; Yuan, S.; Yuan, X.; Liang, J.; Chen, Z.; Dong, X.; and Xiao, Y. 2024. Character is Destiny: Can Large Language Models Simulate Persona-Driven Decisions in Role-Playing? *CoRR*, abs/2404.12138.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; Yang, F.; Deng, F.; Wang, F.; Liu, F.; Ai, G.; Dong, G.; Zhao, H.; Xu, H.; Sun, H.; Zhang, H.; Liu, H.; Ji, J.; Xie, J.; Dai, J.; Fang, K.; Su, L.; Song, L.; Liu, L.; Ru, L.; Ma, L.; Wang, M.; Liu, M.; Lin, M.; Nie, N.; Guo, P.; Sun, R.; Zhang, T.; Li, T.; Li, T.; Cheng, W.; Chen, W.; Zeng, X.; Wang, X.; Chen, X.; Men, X.; Yu, X.; Pan, X.; Shen, Y.; Wang, Y.; Li, Y.; Jiang, Y.; Gao, Y.; Zhang, Y.; Zhou, Z.; and Wu, Z. 2023. Baichuan 2: Open Large-scale Language Models. *CoRR*, abs/2309.10305.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.;

Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yu, J.; Zhang, X.; Xu, Y.; Lei, X.; Guan, X.; Zhang, J.; Hou, L.; Li, J.; and Tang, J. 2022. XDAI: A Tuning-free Framework for Exploiting Pre-trained Language Models in Knowledge Grounded Dialogue Generation. In Zhang, A.; and Rangwala, H., eds., *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, 4422–4432. ACM.

Yuan, X.; Yuan, S.; Cui, Y.; Lin, T.; Wang, X.; Xu, R.; Chen, J.; and Yang, D. 2024. Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works. *CoRR*, abs/2404.12726.

Zhang, S.; Lu, Y.; Liu, J.; Yu, J.; Qiu, H.; Yan, Y.; and Lan, Z. 2024. Unveiling the Secrets of Engaging Conversations: Factors that Keep Users Hooked on Role-Playing Dialog Agents. *CoRR*, abs/2402.11522.

Zheng, Y.; Zhang, R.; Huang, M.; and Mao, X. 2020. A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 9693–9700. AAAI Press.

Zhou, J.; Chen, Z.; Wan, D.; Wen, B.; Song, Y.; Yu, J.; Huang, Y.; Peng, L.; Yang, J.; Xiao, X.; Sabour, S.; Zhang, X.; Hou, W.; Zhang, Y.; Dong, Y.; Tang, J.; and Huang, M. 2023a. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models. *CoRR*, abs/2311.16832.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.; Bisk, Y.; Fried, D.; Neubig, G.; and Sap, M. 2023b. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *CoRR*, abs/2310.11667.