RoleEval: A Bilingual Role Evaluation Benchmark for Large Language Models

Anonymous ACL submission

Abstract

The rapid evolution of large language models necessitates effective benchmarks for evaluating their role knowledge, which is essential for establishing connections with the real world and providing more immersive interactions. This paper introduces RoleEval, a bilingual benchmark designed to assess the memorization, utilization, and reasoning capabilities of role knowledge. RoleEval comprises *RoleEval-Global* (including internationally recognized characters) and *RoleEval-Chinese* (including characters popular in China), with 6,000 Chinese-English parallel multiple-choice questions focusing on 300 influential people and fictional characters drawn from a variety of domains including celebrities, anime, comics, movies, TV series, games, and fictions. These questions cover basic knowledge and multi-hop reasoning abilities, aiming to systematically probe various aspects such as personal information, relationships, abilities, and experiences of the characters. To maintain high standards, we perform a hybrid quality check process combining both automatic and human verification, ensuring that the questions are diverse, challenging, and discriminative.

Our extensive evaluations with RoleEval across various open-source and proprietary large language models, under both the zero- and few-shot settings, reveal insightful findings. Notably, while GPT-4 outperforms other models on *RoleEval-Global*, Chinese large language models excel on *RoleEval-Chinese*, highlighting significant knowledge distribution differences. We expect that RoleEval would highlight the significance of assessing role knowledge for large language models across various languages and cultural settings.

1 Introduction

Recent years have witnessed the huge success of large language models (LLMs), and agents based on these models present immense potential for reshaping our engagement with machines with their expansive knowledge and remarkable predictive capabilities (Wang et al., 2023a; Xi et al., 2023). The cornerstone of this transformation lies in the development of LLM agents with a keen perception of the real world, offering users a more immersive experience than ever before and laying a solid foundation for the emergence of applications such as Character AI[1] and AI Dungeon.[2] This necessitates the role-playing capabilities of these models, which aim to establish connections with real-world people or characters created by LLMs.

Previous studies show that much of LLM knowledge is acquired during the pretraining phase (Zhou et al., 2023a; Lin et al., 2023). A comprehensive pretraining allows for better role-playing capabilities by covering a wide range of knowledge. This is crucial not just for representing existing characters but also for creating new, believable personas without prior history. Just as actors need a solid understanding of real-world facts to convincingly portray fictional characters, learning from real-world personas enables a model to authentically and consistently represent both real and imaginary characters.

However, there is a shortage of systematic evaluation of role knowledge for these LLMs. Traditional persona-based evaluation benchmarks, such as PersonaChat (Zhang et al., 2018) and PersonalDialog (Zheng et al., 2019), often rely on artificially constructed personas or occupations abstracted from a group of people, which lack the complexity and real-world connection of genuine personas. Thus it is hard for using such conventional benchmarks to assess the models' capability in handling intricate character knowledge. Recent role-playing benchmarks like RoleBench (Wang et al., 2023b) evaluate consistency in language style and knowledge extracted from scripts. However, the knowledge extracted from scripts is fragmented and lacks systematicness, making it challenging to

[1] https://beta.character.ai
[2] https://aidungeon.com

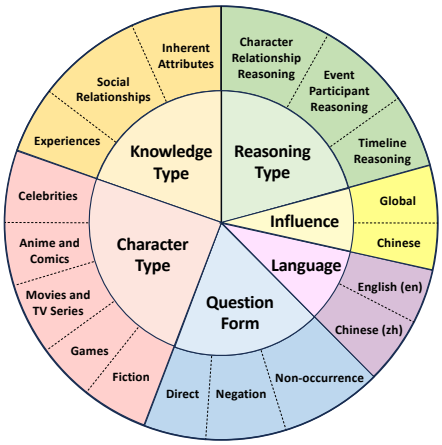


Figure 1: RoleEval includes different types of characters, knowledge, reasoning, question forms, influence, and languages.

accurately and comprehensively assess the breadth of knowledge captured by LLMs.

In light of these issues, we introduce RoleEval, a bilingual benchmark designed to assess the capturing and understanding of role knowledge related to both real-world and fictional characters for LLMs. RoleEval is comprised of two sub-benchmarks: *RoleEval-Global* and *RoleEval-Chinese*, with 300 characters and 6,000 Chinese-English parallel questions in total. Characters in RoleEval are from five categories: 1) celebrities, 2) anime and comics, 3) movies and TV series, 4) games, and 5) fictions. Each character is associated with 20 questions (17 focusing on basic knowledge and 3 on multi-hop reasoning), which evaluate the knowledge and capacity of LLMs in understanding personal information, relationships, abilities, and experiences, as illustrated in Figure 1. These questions are initially crafted in Chinese, then translated into English to construct bilingual parallel data.

Specifically, *RoleEval-Global* is structured around 200 internationally recognized characters with 4,000 questions, which is suitable for evaluating both English and Chinese LLMs. *RoleEval-Chinese* is targeting another set of 100 influential characters in China with 2,000 questions, following the same question construction framework as *RoleEval-Global*, to specifically evaluate Chinese LLMs. The benchmark incorporates both automatic and manual quality controls to maintain high-quality and challenging questions.

To the best of our knowledge, RoleEval is the first benchmark to systematically evaluate the role knowledge for LLMs. It aims to benchmark the current state of LLMs in capturing, understanding, utilizing, and reasoning over knowledge of a wide range of both public figures and fictional characters. By delving into the depths of how LLMs perceive and portray an extensive array of characters, we aim to uncover new avenues for their application, making them not only repositories of information but also active, context-aware participants in our digital narratives.

In a nutshell, our contributions are as follows:

- We propose RoleEval, a bilingual role evaluation benchmark with 6,000 Chinese-English parallel questions covering 300 diverse characters, to systematically examine the ability of capturing, understanding, and reasoning over role knowledge for LLMs, which is an important prerequisite for successful role-playing.
- To ensure quality and boost efficiency, we propose a hybrid quality check process with the combination of both automatic and manual verification to ensure appropriate difficulty and discrimination ability control for questions.
- We conduct extensive evaluations using RoleEval on a variety of large language models under both zero- and few-shot settings, encompassing models with varying parameter sizes and those mainly designed for English and Chinese, as well as both open-source and closed-source proprietary models.

2 Related Work

Role-playing agents (Shuster et al., 2022; Li et al., 2023a; Shao et al., 2023; Wang et al., 2023b), which can be even traced back to ELIZA (Weizenbaum, 1966), the first automated dialogue agent that conducts psychological consulting, have recently attracted growing interest in both industry and academia. However, their evaluation is predominantly conducted on models after supervised fine-tuning. This approach does not incorporate direct feedback from pretrained base models, which can offer critical insights into their intrinsic role-playing capabilities and limitations. On the other hand, existing evaluations largely rely on outputs from the ChatGPT (Ouyang et al., 2022) or humans. However, ChatGPT is not an infallible evaluator, and human evaluation lacks reproducibility. This leads to a lack of objective, accurate, and systematic knowledge assessments. In this case, our benchmark can serve as a robust standard for evaluating current role-playing agents, assessing

whether the models possess sufficient role knowledge.

Previous research in role-playing evaluation has largely focused on abstract personas (Zhang et al., 2018; Zheng et al., 2019; Xu et al., 2022; Wei et al., 2023a; Ahn et al., 2023; Zhou et al., 2023b) or specific professions like psychology consultants, chemists, or software engineers (Liu et al., 2021; Wang et al., 2023a). However, these methods often oversimplify real-world personas, failing to capture the complex nature of real-world human personalities and behaviors in role-playing scenarios. Despite the value of these approaches, they fall short in exploring individual character knowledge, which is critical for authentic role-playing experiences. Therefore, there is a clear need for more sophisticated and realistic persona models in role-playing research. There are also some closely related works for character-based evaluation, intending to inspect the ability to mimic real-world characters (Shao et al., 2023; Wang et al., 2023b; Zhou et al., 2023b). However, they do not have a detailed and systematic framework to evaluate role knowledge, leading to fragmented and incomplete knowledge assessment. Moreover, their evaluation relies on humans or other powerful large language models such as ChatGPT. However, as we have stated above, this kind of evaluation suffers from the reproducibility and accuracy of judges. In contrast, RoleEval examines the knowledge required for role play in a detailed, objective and systematic approach for LLMs.

3 RoleEval

We develop RoleEval to assess how well role knowledge is captured, utilized, and reasoned with. To that end, we gather a diverse range of characters and manually create questions that systematically evaluate various aspects of role knowledge. Such created questions usually demand a thorough understanding and adaptable use of role knowledge, as well as multi-level reasoning over role knowledge.

3.1 Character Collection

RoleEval compiles its character set from diverse sources, including Wikipedia[3], Baidu Baike[4], Fandom[5], and Moegirlpedia.[6] While Wikipedia and Baidu Baike provide a wide range of information, Fandom and Moegirlpedia offer detailed insights into anime, comics, and games. For *RoleEval-Global*, we gather 200 characters known internationally across five categories: celebrities, anime/comics, movies/TV series, games, and fiction, ensuring diversity by selecting an equal number of characters from each category and different works for each fictional character. *RoleEval-Chinese* extends this collection with an additional 100 characters popular primarily in China, totaling 300 characters for RoleEval. These characters are detailed in the appendix A.5.

	RoleEval-Global	RoleEval-Chinese	Total
# Characters	200	100	300
# Questions	4,000	2,000	6,000
Languages	zh, en	zh, en	zh, en
Avg. Q. Tokens (zh)	24.6	26.1	25.1
Avg. Q. Tokens (en)	14.5	16.0	15.0

Table 1: Statistics of RoleEval. Here we use zh and en to indicate Chinese and English respectively.

Before manually creating questions related to collected characters, we rigorously verify each character's encyclopedia information to ensure comprehensiveness. We prefer characters with abundant, collaboratively-contributed information for its reliability and objectivity, ensuring our dataset's diversity, accuracy, and true representation of characters. Additionally, we assess character popularity on social media (via follower counts and engagement) and search engine presence (by search result volume). During the manual question creation phase, we refer to multiple online encyclopedias, selecting only consistent knowledge points to avoid conflicts.

3.2 Question Design

Our focus is primarily on factual knowledge and comprehensively assessing the ability of LLMs in capturing and understanding various character roles and contexts. In addition to basic knowledge that is directly stated in encyclopedias, we also design multi-hop questions to examine the ability of LLMs to dynamically combine and reason with existing knowledge.

Specifically, after collecting the 300 characters from online encyclopedias, we build RoleEval in the form of multiple-choice questions, with four options for each question, which is a common practice adopted by many existing benchmarks such as MMLU (Hendrycks et al., 2021) and C-Eval (Huang et al., 2023). Each character is associated with 17 and 3 unique questions for basic knowledge

[3] https://www.wikipedia.org
[4] https://baike.baidu.com
[5] https://www.fandom.com
[6] https://zh.moegirl.org.cn

and multi-hop reasoning respectively, thus culminating in a total of 6,000 questions. The statistics of RoleEval are shown in Table 1. We also illustrate the length distribution of created questions in Figure 5.

In RoleEval, we consider three types of fundamental knowledge required to depict a character:

1. **Inherent Attributes**: This type of knowledge includes the fundamental characteristics intrinsic to the character, such as gender, race, personality, skills, and abilities. These attributes are typically presented in a tabular format, or directly described in online encyclopedias.
2. **Social Relationships**: This type of knowledge pertains to the relationships of the character with other individuals, which could include parents, disciples, and other significant personal or professional relationships.
3. **Experiences**: This type of knowledge details the experiences or events that the character has undergone. For real-world individuals, this usually includes significant life events or experiences in which they were direct participants. For fictional characters, this involves extracting key plot points or story arcs described in online encyclopedias.

To enhance the comprehensive assessment and diversify the scope of multiple-choice questions, our approach extends beyond merely querying knowledge from online encyclopedias. As a supplement to direct questions, we incorporate two additional question types, which are designed to be combined with the previously identified four types of knowledge for a more dynamic and comprehensive evaluation:

1. **Negation Question Type:** These questions, usually formatted as "Which of the following is NOT...", require a comprehensive understanding of a specific knowledge point.
2. **Non-occurrence Scenario Question Type:** These questions test for non-occurrences, with correct answers often framed in the negative (e.g., "Did not happen..."). This format examines whether the model generates illusions or false assumptions.

We further add three reasoning questions for each character that need multi-hop reasoning over these types of fundamental knowledge. According to the knowledge required in the intermediate reasoning steps, we classify these reasoning questions into three categories, and assign one question for each available (role, reasoning category) pair:

1. **Character Relationship Reasoning:** When answering this type of questions, models need to reason about the relationship between characters.
2. **Event Participant Reasoning:** To answer a question in this category, models need to reason out the participants of an event, and then combine it with other information in the question to locate the answer.
3. **Timeline Reasoning:** Answering questions in this group requires models to understand the sequence of events, infer the time of occurrence of the event in question stems and options, and then select the correct option based on the timeline.

These reasoning questions go beyond simply memorizing one-hop knowledge in text, intending to connect multiple related characters, events, and storylines. It requires the compositionality capability of models to intrinsically and dynamically combine multiple knowledge points and answer the given questions, thus making our RoleEval a challenging benchmark for LLMs. For more detailed information, we illustrate the examples of negation, non-occurrence and different types of reasoning in Appendix A.1 and A.2.

3.3 Quality Check

To ensure quality and boost efficiency during benchmark construction, we propose a hybrid quality check process with the combination of both automatic and human verification. In the automatic checking stage, we use GPT-4 and GPT-3.5 API[7] to control the degree of difficulty and discrimination ability. Assume the accuracy of GPT-4 and GPT-3.5 API for answering questions of character c is x_c and y_c respectively. We apply the below criteria as the preliminary and instant feedback for human annotators:

$$x_l \leq x_c \leq x_u \quad (1)$$

$$y_c \leq y_u \quad (2)$$

$$x_c - y_c \geq d \quad (3)$$

where x_l and x_u indicate the predefined lower threshold and upper threshold for x_c, and y_u indicates the predefined upper threshold for y_c. These

[7]Unless otherwise specified, the term "GPT-4" and "GPT-3.5" shall henceforth be used to represent `gpt-4-0613` and `gpt-3.5-turbo-0613` respectively.

以下是关于名人的单项选择题，请选出其中的正确答案。(The following are multiple choice questions about celebrities, please select the correct answer.)

以下哪件事发生在唐纳德·特朗普投资橄榄球队“新泽西将军”之前？(Which of the following events happened before Donald Trump invested in the football team “New Jersey Generals”?)
A. 出让特朗普航空公司 (Selling Trump Shuttle)
B. 买下纽约广场饭店 (Buying the New York Plaza Hotel)
C. 建成了凯悦酒店 (Completion of the Hyatt Hotel)
D. 赢得美国总统大选 (Winning the US presidential election)
答案：C (Answer: C)

(Other few-shot examples with the same format…)

达芬奇为什么没有将自己的手稿和作品交给弗朗西斯科·梅尔兹？(Why didn’t Da Vinci give his manuscripts and works to Francesco Melzi?)
A. 没有来得及 (Didn't have enough time)
B. 交给了弗朗西斯科·梅尔兹 (Gave to Francesco Melzi)
C. 不喜欢弗朗西斯科·梅尔兹 (Didn't like Francesco Melzi)
D. 想交给自己的孩子 (Wanted to give to his own children)
答案：(Answer:)

Figure 2: Few-shot prompt examples for RoleEval. English translations are placed after the corresponding Chinese text for clarity.

three hyperparameters control the overall degree of difficulty, making sure that questions are not too easy or too hard for LLMs, in terms of the answering accuracy of GPT-4/3.5. And d is the lower threshold of difference between x_c and y_c. Since GPT-4 is a significantly more powerful model than GPT-3.5 in practice, we use this hyperparameter to ensure the discrimination ability of this benchmark for various models. Only questions that satisfy the conditions in Eq. (1)-(3) are kept. In our preliminary study, we find $x_l = 0.3$, $x_u = 0.9$, $y_u = 0.8$, and $d = 0.15$ achieve appropriate difficulty and discrimination ability control for questions.

After this automatic scrutinization, we manually check the kept questions and options to ensure the quality of our benchmark. To easily check the factual correctness and prevent questions from overemphasis on peripheral aspects, we ask annotators to also provide links to the referenced text in the online encyclopedias along with each question to achieve effective oversight.

3.4 Translation

To evaluate LLMs in different languages, we translate Chinese RoleEval questions to English using GPT-4.[8] We have found that GPT-4 produce more adaptable and precise character-specific translations than traditional machine translation engines like Google Translate, due to its customizable prompts. To address potential ambiguities in entity translations, we hire human translators to review and adjust them based on the original Chinese content. Given RoleEval’s focus on factual accuracy over stylistic elements, combining machine translation with human post-editing effectively minimizes information loss.

[8]We use RoleEval (**en**) and RoleEval (**zh**) to indicate RoleEval in English and Chinese respectively.

Models	Open Source?	Model Size	Primary Language
BLOOM	Yes	1.1B, 1.7B, 3B, 7.1B	English
Pythia	Yes	1.4B, 2.8B, 6.9B, 12B	English
LLaMA	Yes	7B, 13B, 30B, 65B	English
LLaMA-2	Yes	7B, 13B, 70B	English
Falcon	Yes	7B, 40B	English
Mistral	Yes	7B	English
GPT-4	No	undisclosed	English
GPT-3.5	No	undisclosed	English
ChatGLM3	Yes	6B	Chinese
Baichuan2	Yes	7B, 13B	Chinese
Qwen	Yes	1.8B, 7B, 14B, 72B	Chinese
Chinese-LLaMA-2	Yes	7B, 13B	Chinese
Skywork	Yes	13B	Chinese
Yi	Yes	6B, 34B	Chinese
MiniMax	No	undisclosed	Chinese

Table 2: Details of models evaluated in our experiments. For simplicity, we only list the sizes of evaluated models in the “Model Size” column.

4 Experiments

We evaluated a wide range of English and Chinese LLMs on RoleEval, aiming to analyze the capturing, utilizing, and reasoning capabilities of role knowledge for these LLMs.

4.1 Setup

We implemented an evaluation pipeline for RoleEval with the lm-evaluation-harness framework (Gao et al., 2023) under both zero- and five-shot setting. Considering RoleEval and MMLU (Hendrycks et al., 2021) share the same multiple-choice format, we adopted a similar setup with MMLU for RoleEval. Specifically, for open-source models, we calculated the probability of subsequent tokens following the initial prompt. Among the options

Model	RoleEval-Global (4,000 questions)						RoleEval-Chinese (2,000 questions)					
	CE	AC	MT	GA	FI	Avg.	CE	AC	MT	GA	FI	Avg.
GPT-3.5-0613	46.62	48.38	51.75	49.50	47.38	48.73	42.25	43.50	39.75	43.75	39.00	41.65
GPT-3.5-1106	48.75	51.88	51.25	49.88	48.38	50.02	47.50	46.75	41.75	44.75	38.75	43.90
GPT-4-0613	73.38	72.12	74.25	72.25	69.62	72.32	57.75	60.25	57.75	60.00	58.00	58.75
GPT-4-1106	**74.75**	**73.62**	**74.38**	**72.50**	71.62	**73.38**	62.50	**63.25**	63.00	**62.00**	63.00	62.75
Falcon-7B	23.88	28.12	24.50	28.00	28.12	26.52	24.75	30.50	31.50	29.75	25.25	28.35
Falcon-40B	39.62	32.25	32.38	30.00	45.00	35.85	28.25	33.00	30.25	29.25	38.50	31.85
LLaMA-7B	25.50	31.87	25.87	26.00	28.88	27.62	28.50	24.75	20.50	27.75	29.00	26.10
LLaMA-13B	28.50	28.50	28.25	26.50	27.75	27.90	27.25	29.75	27.25	26.00	29.00	27.85
LLaMA-30B	24.88	31.13	30.25	27.75	28.62	28.52	30.00	28.75	26.00	31.75	28.00	28.90
LLaMA-65B	32.12	31.87	32.75	31.00	34.88	32.52	30.00	32.25	29.00	35.50	29.00	31.15
LLaMA-2-7B	37.00	29.88	28.75	34.50	38.25	33.67	25.75	28.00	33.75	29.75	34.50	30.35
LLaMA-2-13B	36.50	34.00	33.00	31.87	31.75	33.42	28.75	30.50	25.25	29.75	28.25	28.50
LLaMA-2-70B	53.50	43.25	39.25	40.25	47.25	44.70	36.00	38.00	36.25	36.25	34.75	36.25
Mistral-7B	36.12	33.50	32.00	30.25	35.00	33.38	32.50	37.50	26.25	33.25	31.50	32.20
MiniMax	51.75	54.50	62.62	56.75	52.75	55.67	54.00	55.00	52.75	57.50	54.00	54.65
Baichuan2-7B	56.00	49.62	45.50	40.50	52.38	48.80	52.25	43.75	49.00	47.25	55.00	49.45
Baichuan2-13B	60.25	52.38	51.00	46.88	60.75	54.25	54.75	47.75	54.00	47.50	60.00	52.80
ChatGLM3-6B	56.50	47.62	48.38	41.88	54.50	49.78	50.00	44.50	48.00	44.25	58.00	48.95
Chinese-LLaMA-2-7B	35.62	36.75	35.62	35.38	34.38	35.55	34.50	29.00	33.00	30.25	36.25	32.60
Chinese-LLaMA-2-13B	45.38	38.25	39.88	31.87	42.12	39.50	36.50	36.50	34.00	34.00	40.50	36.30
Qwen-7B	54.75	44.38	44.62	42.75	53.00	47.90	49.00	42.00	47.50	44.75	51.25	46.90
Qwen-14B	62.50	52.38	55.00	45.50	58.00	54.67	56.25	45.50	54.75	51.50	56.75	52.95
Qwen-72B	72.88	63.88	70.38	56.75	**73.50**	67.47	**70.00**	59.75	66.00	61.25	74.00	**66.20**
Skywork-13B	59.13	51.75	51.88	44.50	58.75	53.20	55.25	45.75	56.00	48.50	57.50	52.60
Yi-6B	61.88	51.38	52.38	45.38	60.75	54.35	59.25	46.00	61.50	47.75	62.00	55.30
Yi-34B	72.38	60.62	69.75	53.25	73.12	65.83	65.50	54.50	**70.00**	56.00	**77.00**	64.60

Table 3: Five-shot results on RoleEval (**zh**) in five categories: celebrities (**CE**), anime and comics (**AC**), movie and TV series (**MT**), games (**GA**), and fiction (**FI**).

"A", "B", "C", and "D", we chose the one with the highest probability as the preferred choice of the model. For closed-source models such as GPT-4 (OpenAI, 2023), we followed Huang et al. (2023) and Li et al. (2023b) to use regular expressions to extract the preferred choice of the model. All experiments in this paper were conducted using two NVIDIA A800 80G GPUs.

4.2 Prompt

Figure 2 illustrates the prompt we used for evaluation. For each question, we added "以下是关于[category]的单项选择题，请选出其中的正确答案。" ("The following multiple-choice questions are about [category]. Please choose the correct answer.") before the question stem, and "答案：" ("Answer: ") after four options, where "[category]" was chosen from 名人 ("celebrities"), 动漫角色 ("anime and comics"), 影视角色 ("movies and TV series"), 游戏角色 ("games") and 小说人物 ("fictions"). For the five-shot setting, we added in-context examples before the actual question to answer. These examples shared the same format as the actual question, except that the ground truth option was provided for each in-context example.

4.3 Models

We selected a wide range of publicly available LLMs for evaluation. Due to the limit of computational resources, we only evaluated models with the number of parameters ranging from 1B to 72B parameters since they can produce meaningful results and be loaded in two 80GB Nvidia A800 GPUs with bf16 format.

Chinese LLMs For Chinese questions in RoleEval, we evaluated popular Chinese open-source LLMs, such as ChatGLM (Du et al., 2022; Zeng et al., 2022), Baichuan (Yang et al., 2023), Qwen (Bai et al., 2023), Yi[9], Skywork (Wei et al., 2023b), Chinese-LLaMA-2 (Cui et al., 2023), along with close-sourced LLMs like Minimax.[10]

English LLMs For English questions in RoleEval, we evaluated open-source LLMs including BLOOM (Workshop et al., 2022), Pythia (Biderman et al., 2023), LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b), Falcon (Almazrouei et al., 2023), Mistral-7B (Jiang et al., 2023), and two closed-source LLMs: ChatGPT (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023).

Since Baichuan, Qwen, Yi, Chinese-LLaMA-2, BLOOM, Pythia, LLaMA, LLaMA2, and Falcon have multiple sizes of LLMs that satisfy our model restriction, we evaluated various sizes of LLMs for each LLM family as shown in Table 2.

[9] https://github.com/01-ai/Yi
[10] https://api.minimax.chat

Model	RoleEval-Global (4,000 questions)						RoleEval-Chinese (2,000 questions)					
	CE	AC	MT	GA	FI	Avg.	CE	AC	MT	GA	FI	Avg.
GPT-3.5-0613	57.38	59.62	58.13	59.50	57.50	58.43	42.00	47.75	42.50	42.25	45.50	44.00
GPT-3.5-1106	58.75	56.62	55.75	58.00	55.00	56.82	38.25	45.50	44.00	44.50	46.00	43.65
GPT-4-0613	**77.62**	**79.50**	73.12	74.88	**75.00**	**76.02**	54.25	61.75	**63.00**	**63.00**	**63.00**	**61.00**
GPT-4-1106	75.12	78.75	**75.00**	**76.12**	**75.00**	76.00	**57.50**	**63.50**	60.00	62.50	58.00	60.30
Falcon-7B	26.25	27.75	28.50	29.38	31.00	28.58	27.25	27.75	27.75	29.75	28.50	28.20
Falcon-40B	47.38	45.00	49.62	43.12	50.00	47.02	34.00	38.25	30.75	38.75	35.25	35.40
LLaMA-7B	29.38	30.50	29.25	33.50	28.50	30.23	24.00	27.50	29.75	33.00	29.25	28.70
LLaMA-13B	39.38	40.25	39.88	40.62	43.00	40.63	32.75	31.75	30.75	38.50	32.00	33.15
LLaMA-30B	51.62	46.88	48.62	43.12	52.62	48.57	34.75	35.75	30.75	40.00	35.00	35.25
LLaMA-65B	58.13	50.50	54.37	47.62	54.50	53.02	41.50	38.50	33.50	43.25	37.50	38.85
LLaMA-2-7B	38.88	37.00	37.50	41.62	42.38	39.48	28.75	29.25	32.75	37.50	32.25	32.10
LLaMA-2-13B	49.38	43.50	46.50	44.25	48.25	46.38	30.50	36.50	33.25	36.50	33.25	34.00
LLaMA-2-70B	63.25	57.38	59.00	50.00	63.25	58.58	43.25	41.50	40.25	47.50	43.50	43.20
Mistral-7B	54.87	46.75	49.62	44.25	52.25	49.55	35.75	42.00	30.00	41.75	31.50	36.20
MiniMax	54.87	56.38	53.50	54.12	51.38	54.05	34.00	39.50	40.75	38.25	39.00	38.30
Baichuan2-7B	51.00	45.12	49.00	42.12	50.00	47.45	37.25	35.75	33.00	40.25	37.00	36.65
Baichuan2-13B	56.12	47.50	51.50	45.62	54.00	50.95	35.50	36.50	31.25	42.25	34.75	36.05
ChatGLM3-6B	55.12	46.62	49.25	43.25	52.62	49.37	36.25	36.25	35.25	42.25	43.50	38.70
Chinese-LLaMA-2-13B	47.75	46.00	46.88	45.00	48.38	46.80	34.00	38.50	27.75	37.50	34.00	34.35
Chinese-LLaMA-2-7B	36.50	30.75	31.75	36.25	39.50	34.95	30.50	27.75	33.00	30.50	27.75	29.90
Qwen-7B	53.87	46.12	48.12	40.00	51.12	47.85	36.25	36.00	36.25	42.25	40.00	38.15
Qwen-14B	61.12	49.00	53.87	45.38	56.12	53.10	41.00	38.75	38.25	43.25	41.00	40.45
Qwen-72B	70.12	62.00	69.00	55.75	69.50	65.27	52.75	47.50	46.50	54.25	50.50	50.30
Skywork-13B	56.25	46.75	51.62	44.38	53.62	50.52	39.25	34.50	38.25	41.75	38.50	38.45
Yi-6B	59.25	52.00	54.12	47.50	56.25	53.82	42.25	38.50	41.50	44.25	45.00	42.30
Yi-34B	73.12	61.75	67.88	57.12	67.25	65.42	56.00	52.00	47.50	55.00	57.00	53.50

Table 4: Five-shot results on RoleEval (**en**) in five categories: celebrities (**CE**), anime and comics (**AC**), movie and TV series (**MT**), games (**GA**), and fiction (**FI**).

4.4 Results and Analysis

Overall Performance Table 3 and 4 show the few-shot experimental results. Since the zero-shot results are generally lower than the few-shot experimental results, we provide them in Appendix A.3. We find that GPT-4 maintains a lead on *RoleEval-Global*, with its latest version (`gpt-4-1106`) outperforming the earlier (`gpt-4-0613`). We believe this superiority is attributed to the more recent knowledge cutoff in `gpt-4-1106`, enhancing its performance in domains with rapidly evolving information. Nevertheless, the overall accuracy still indicates large room for improvement even for the state-of-the-art LLMs. On the *RoleEval-Chinese* (**zh**) dataset, some Chinese LLMs, such as Qwen-72B and Yi-34B, show superior performance to GPT-4. This is likely due to their higher proportion of Chinese training data and abundant high-quality discussions about characters in *RoleEval-Chinese* on Chinese online platforms. Notably, GPT-4 retains its edge in anime, comics, and games, where many characters are also popular in the English-speaking world. These results highlight the importance of choosing balanced training data and evaluating role knowledge across various languages and cultural settings. We also provide results and analysis grouped by knowledge and reasoning questions in Appendix A.4.

Differences in Languages We observe a significant improvement in GPT-4 and 3.5's performance on RoleEval (**en**) dataset compared to RoleEval (**zh**), even though these two parts of the dataset have identical semantic content. This observation is also evident in predominantly English-language open-source models, especially for models with little Chinese training data, such as LLaMA, Mistral, and Falcon. This suggests that even the most powerful LLMs still lack effective cross-lingual knowledge transfer, indicating that these models fail to build complete bi-directional mappings between entities in different languages. Conversely, Chinese models generally underperform on the English RoleEval datasets, highlighting a similar language-specific limitation, though the gap is narrower since Chinese LLMs still use a large amount of English pre-training data. Among all open-source models, Yi-34B achieves almost the same performance in both Chinese and English on *RoleEval-Global*, indicating its balanced training for global influential characters in both languages.

Comparative Analysis of Open-Source Models LLaMA2-70B has emerged as the best open-source model primarily trained in English, closely matching GPT-3.5. While for Chinese LLMs, Qwen-72B and Yi-34B not only surpass GPT-3.5 but also exceed GPT-4 on the *RoleEval-Chinese* (**zh**) dataset. However, these Chinese models still show notice-

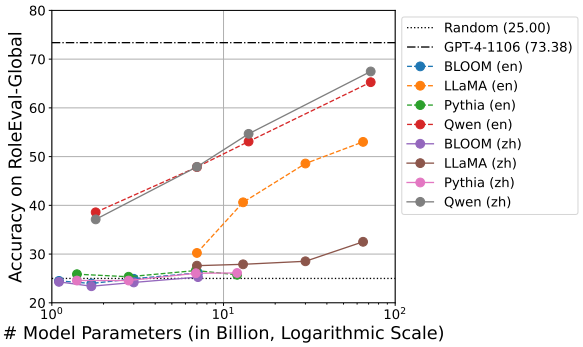


Figure 3: Parameter Scaling Laws on *RoleEval-Global*.

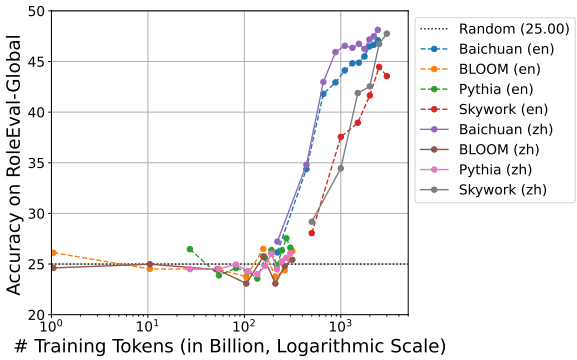


Figure 4: Token Scaling Laws on *RoleEval-Global*.

able gaps compared to GPT-4 in other scenarios.

Parameter Scaling Laws We also analyzed the correlation between the accuracy of role knowledge and the size and the number of training tokens of LLMs on *RoleEval-Global*.[11] As shown in Figure 3, accuracy generally improves with model size for LLaMA and Qwen, and the trend is consistent with previously established knowledge transfer patterns: For Chinese LLMs, the rate of improvement is greater for RoleEval (**zh**) than for RoleEval (**en**). In contrast, LLMs primarily trained on English corpora show opposite trends. Notably, BLOOM and Pythia do not show performance improvements across various settings. We speculate that this is due to their relatively lower token training volume (366B for BLOOM and 300B for Pythia), while most models with great performance have already been trained on more than 1TB tokens.

Token Scaling Laws To further explore the scaling law on the number of tokens, we conducted experiments on publicly available intermediate checkpoints of BLOOM-7B1, Pythia-6.9B, Baichuan-7B, and Skywork-13B. For BLOOM, Baichuan, and Skywork, we selected all available intermediate checkpoints, resulting in 8, 11, and 6 checkpoints respectively. For Pythia, we chose a checkpoint every 13,000 steps and obtained 12 intermediate checkpoints for evaluation.

Results from Figure 4 indicate that while the performance of Baichuan and Skywork with their first checkpoint (trained with 220B and 500B respectively) is near random, which is similar to BLOOM and Pythia, their subsequent checkpoints show steady improvement after 500B tokens, which indicates the importance of sufficient training tokens with fixed model size. However, the bottleneck of cross-lingual knowledge transfer can still be observed with increasing training tokens, which means that simply increasing the number of parameters and adding training tokens may not be the best way to break down the barriers between languages. In future research, we intend to examine other LLMs with intermediate checkpoints trained on larger-scale datasets primarily in English, aiming to substantiate our hypotheses with more robust evidence.

[11]To obtain meaningful results, we only chose model families with more than three models in this experiment, and we required the amount of training data and training tokens to be the same for models with different parameters within the same model family.

5 Conclusion

In this paper, we have presented RoleEval, a large-scale bilingual role evaluation benchmark, featuring 6,000 Chinese-English parallel questions across 300 diverse characters (200 for *RoleEval-Global* and 100 for *RoleEval-Chinese*) from five different domains. RoleEval is specifically designed to scrutinize the capabilities of LLMs in capturing, understanding, and reasoning over role knowledge. Our hybrid quality check process, blending automatic and human verification, guarantees the questions' difficulty and discrimination level, setting a new standard in benchmark design. Extensive evaluations of RoleEval on various LLMs, including both zero- and few-shot settings, highlight significant differences in knowledge distribution, as evidenced by GPT-4's superior performance in *RoleEval-Global* and the notable excellence of Chinese LLMs in *RoleEval-Chinese*. These findings not only demonstrate the disparities in LLMs' knowledge proficiency but also illuminate the path for future enhancements in bilingual and culture-specific LLMs. Through RoleEval, we aim to provide a robust framework for future advancements in LLM evaluation, particularly in role-playing scenarios, thereby enriching the landscape of language understanding and reasoning benchmarks.

Limitations

In evaluating the effectiveness of real-world role evaluation benchmarks, there are two existing limitations. Firstly, the aspect of timeliness is crucial; the knowledge regarding real-world characters may change over time, making the benchmark outdated or irrelevant. To address this, we plan to explore methods for the automatic updating of benchmarks, ensuring that they remain current and reflective of ongoing changes. Secondly, the current format of benchmarks often restricts questions to having only one correct answer. This approach fails to adequately test scenarios where multiple answers could be correct, thus limiting the benchmark's ability to evaluate complex decision-making skills. A potential solution could be to incorporate a more dynamic question format that allows for the identification and acceptance of multiple correct answers, thereby enriching the assessment process by acknowledging the multi-faceted nature of real-world problems.

Ethics Statement

Our benchmark is designed to enhance the model's understanding of role knowledge, which is crucial for improving persona consistency and factual accuracy while reducing hallucination. To achieve this, we have selected encyclopedic content that has been edited by multiple individuals. This approach helps to minimize factual errors and biases, particularly in comparison to other texts sourced from the Internet. Furthermore, all data collected for this project originates from publicly available materials, ensuring no concerns regarding privacy infringement. We confirm that all materials employed were utilized for non-commercial purposes, in adherence to copyright regulations and privacy policies.

Moreover, our goal is to promote a comprehensive and accurate understanding of role knowledge. Therefore, although most of our questions and options are positive, we have not entirely excluded potential negative aspects of the selected characters from our benchmark. Users should be aware of this when using this benchmark. Our commitment is to provide a comprehensive and balanced understanding, but users should remain critical and mindful of the context in which this information is used.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards Multimodal Persona-Grounded Conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *ArXiv preprint*, abs/2309.16609.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *ArXiv preprint*, abs/2304.01373.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *ArXiv preprint*, abs/2304.08177.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *ArXiv preprint*, abs/2305.08322.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *ArXiv preprint*, abs/2310.06825.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023a. ChatHaruhi: Reviving Anime Character in Reality via Large Language Model. *ArXiv preprint*, abs/2308.09597.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. CMMLU: Measuring massive multitask language understanding in Chinese. *ArXiv preprint*, abs/2306.09212.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. *ArXiv preprint*, abs/2312.01552.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. *ArXiv preprint*, abs/2310.10158.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv preprint*, abs/2307.09288.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023a. A Survey on Large Language Model based Autonomous Agents. *ArXiv preprint*, abs/2308.11432.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023b. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. *ArXiv preprint*, abs/2310.00746.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023a. Multi-Party Chat: Conversational Agents in Group Settings with Humans and Models. *ArXiv preprint*, abs/2304.13835.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023b. Skywork: A More Open Bilingual Foundation Model. *ArXiv preprint*, abs/2310.19341.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi,

Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv preprint*, abs/2211.05100.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *ArXiv preprint*, abs/2309.07864.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open Large-scale Language Models. *ArXiv preprint*, abs/2309.10305.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: An Open Bilingual Pre-trained Model. *ArXiv preprint*, abs/2210.02414.

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

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized Dialogue Generation with Diversified Traits. *ArXiv preprint*, abs/1901.09672.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less Is More for Alignment. *ArXiv preprint*, abs/2305.11206.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023b. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models. *ArXiv preprint*, abs/2311.16832.

A Appendix

A.1 Examples of Negation and Non-occurrence Scenario Questions

Negation Question Type: For example, "在《火影忍者》中，以下哪个不是漩涡鸣人的忍术？ A. 万象天引 B. 影分身之术 C. 色诱术 D. 螺旋丸" ("In *Naruto*, Which of the following is not a ninjutsu of Naruto Uzumaki? A. Universal Pull B. Shadow Clone Jutsu C. Sexy Jutsu D. Rasengan"). The correct answer is A because it is the only ninjutsu that Naruto Uzumaki cannot use among the four given options.

Non-occurrence Scenario Question Type: For example, "在《火影忍者》中，漩涡鸣人是什么时候成为中忍的？ A. 第四次忍界大战时 B. 没有成为中忍 C. 佩恩入侵时 D. 第七班完成任务后" ("In *Naruto*, When did Naruto Uzumaki become a Chūnin? A. During the Fourth Great Ninja War B. Never became a Chūnin C. During Pain's invasion D. After Team 7's mission completion."). Actually, Naruto never became a Chūnin since he never passed his Chūnin selection exams, even when he became the Seventh Hokage, making the correct choice B.

A.2 Examples of Different Types of Reasoning Questions

Character Relationship Reasoning: For example, "海莉·比伯的丈夫在《黑衣人3》中客串的什么角色？ A. 外星人 B. 警官 C. 医生 D. 教师" ("What role did Hailey Bieber's husband make a cameo in *Men in Black 3*? A. Alien B. Police Officer C. Doctor D. Teacher"). Models need to first reason out that Hailey Bieber's husband is Justin Bieber, and then find out that Justin Bieber made a cameo appearance as an alien in *Men in Black 3*, so the correct answer is A.

Event Participant Reasoning: For example, "下面哪个人既征服了波斯，又攻下了埃及？ A. 亚历山大大帝 B. 腓力二世 C. 冈比西斯二世 D. 阿明塔斯三世" ("Which of the following people conquered both Persia and Egypt? A. Alexander the Great B. Philip II C. Cambyses II D. Amyntas III"). Models need to know that although more than one person conquered Persia or Egypt (e.g., Cambyses II in option C used to conquer Egypt in 525 BC), Alexander the Great is the only one who conquered both Persia and Egypt among the four options.

Timeline Reasoning: For example, "在《巴黎圣母院》中，哪件事在卡西莫多受鞭笞之刑前发生？ A. 埃斯梅拉达被判绞刑 B. 卡西莫多看见埃斯梅拉达被绞死 C. 埃斯梅拉达给卡西莫多喝水 D. 卡西莫多把埃斯梅拉达从绞刑架上救下" ("In *The Hunchback of Notre Dame*, which event occurs before Quasimodo is whipped? A. Esmeralda was sentenced to hanging B. Quasimodo saw Esmeralda being hanged C. Esmeralda gave Quasimodo water D. Quasimodo took Esmeralda saved from the gallows"). Among these options, only option C happened before whipping, while A, B, and D all happened after that.

A.3 Zero-shot Experimental Results

We list zero-shot results by five categories in Table 5 and 6. Overall, the zero-shot performance of most models is slightly lower than (or similar to) five-shot results. We find that although GPT-4 and GPT-3.5 API already have great instruction-following abilities, providing five examples containing characters that are not related to who we aim to assess can still improve both utilization and reasoning of role knowledge. We conjecture that for powerful LLMs like GPT-4 and GPT-3.5, these examples act as cues that activate role knowledge stored in the model. By presenting scenarios where specific roles are depicted, the model can more effectively access and utilize its internal role knowledge, even if these roles are not related to the character that the question actually asks.

A.4 Results by Knowledge and Reasoning

We provide results of knowledge and reasoning questions for RoleEval in Table 7 and 8. We observe a clear positive correlation between the breadth of knowledge and the correctness of reasoning. Among these LLMs, the GPT-4 model shows both extensive knowledge and robust reasoning capability. However, it is noteworthy that in specific versions of GPT-4 API, the `gpt-4-1106` model demonstrates a more comprehensive knowledge base, while the `gpt-4-0613` model is better at reasoning. This delineation suggests that while there is a general trend of knowledge and reasoning abilities advancing in tandem, individual models may specialize or excel in one aspect over the other.

Further, the experimental findings in LLMs trained for different languages reveal intriguing insights. In English LLMs, Mistral-7B has shown remarkable reasoning capabilities, closely rivaling

Model	RoleEval-Global (4,000 questions)						RoleEval-Chinese (2,000 questions)					
	CE	**AC**	**MT**	**GA**	**FI**	**Avg.**	**CE**	**AC**	**MT**	**GA**	**FI**	**Avg.**
GPT-3.5-0613	45.25	45.88	46.38	47.12	45.50	46.03	41.00	41.00	34.75	33.50	38.50	37.75
GPT-3.5-1106	45.62	45.50	45.50	44.50	43.25	44.87	40.75	41.25	33.25	36.50	39.50	38.25
GPT-4-0613	70.75	71.50	73.12	69.38	70.50	71.05	62.00	59.75	57.25	55.00	58.75	58.55
GPT-4-1106	**75.12**	**73.75**	**75.25**	**71.50**	70.75	**73.27**	**63.00**	61.25	56.25	62.00	60.00	60.50
Falcon-7B	23.88	25.25	25.87	28.00	25.12	25.62	26.00	24.50	25.50	25.00	26.25	25.45
Falcon-40B	40.50	35.25	32.12	35.50	45.75	37.82	32.25	30.00	36.25	29.75	31.00	31.85
LLaMA-7B	27.12	29.50	28.38	30.88	25.12	28.20	31.50	26.25	26.50	30.50	28.50	28.65
LLaMA-13B	28.38	28.38	23.25	26.62	30.00	27.33	22.50	23.50	25.75	32.00	26.50	26.05
LLaMA-30B	27.12	28.25	28.25	27.38	23.62	26.92	25.00	30.50	23.00	22.75	32.00	26.65
LLaMA-65B	31.87	29.38	30.50	29.62	30.63	30.40	27.25	29.00	28.25	29.00	29.00	28.50
LLaMA-2-7B	28.38	30.75	29.25	29.88	28.50	29.35	27.50	26.00	31.25	27.50	26.75	27.80
LLaMA-2-13B	30.38	30.75	32.38	29.38	32.50	31.08	27.50	27.75	26.50	26.50	29.50	27.55
LLaMA-2-70B	46.12	40.25	37.62	40.12	42.50	41.32	36.50	37.00	34.50	31.00	34.00	34.60
Mistral-7B	37.00	31.37	29.00	32.12	30.63	32.02	32.25	27.50	29.00	26.50	32.00	29.45
MiniMax	53.12	56.62	60.00	56.12	55.00	56.17	56.00	52.25	53.00	53.50	54.25	53.80
Baichuan2-7B	54.37	47.25	42.50	38.38	56.12	47.72	40.25	53.25	53.50	49.25	41.25	47.50
Baichuan2-13B	60.50	51.50	46.88	45.25	57.12	52.25	45.75	51.25	55.25	54.50	44.25	50.20
ChatGLM3-6B	55.38	47.62	49.12	42.00	56.88	50.20	45.00	51.25	55.50	49.50	44.00	49.05
Chinese-LLaMA-2-7B	29.12	27.25	23.25	26.75	30.00	27.27	22.50	27.00	27.25	31.25	24.00	26.40
Chinese-LLaMA-2-13B	35.62	34.50	33.62	29.25	36.88	33.97	31.00	31.75	34.50	31.00	33.00	32.25
Qwen-7B	56.38	43.62	46.50	38.88	53.87	47.85	40.25	49.50	50.75	46.25	43.25	46.00
Qwen-14B	59.50	54.12	54.25	43.88	58.63	54.08	43.25	51.00	58.75	50.75	50.75	50.90
Qwen-72B	70.62	63.75	68.38	59.25	**72.00**	66.80	58.75	66.50	**72.25**	**64.00**	**62.75**	**64.85**
Skywork-13B	58.13	50.00	47.88	43.75	56.25	51.20	44.50	53.00	57.50	52.50	46.00	50.70
Yi-6B	62.75	54.25	53.00	45.12	58.13	54.65	46.00	57.50	62.00	58.50	45.25	53.85
Yi-34B	67.38	59.00	62.75	51.62	67.88	61.73	54.75	**67.25**	68.75	59.75	56.00	61.30

Table 5: Zero-shot results on RoleEval (**zh**) in five categories: celebrities (**CE**), anime and comics (**AC**), movie and TV series (**MT**), games (**GA**), and fiction (**FI**).

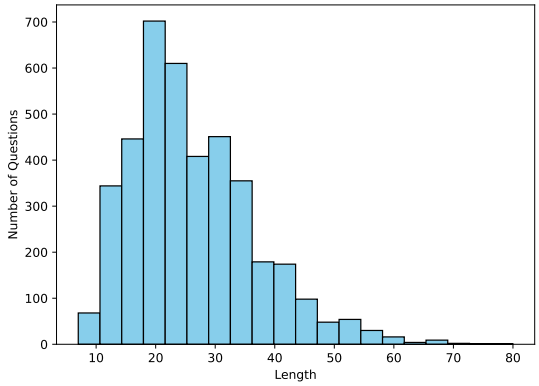
those of the LLaMA-65B model. This performance indicates a significant advancement in reasoning, as smaller models approach the upper echelons of reasoning skills previously dominated by more advanced models. In the realm of Chinese LLMs, Qwen-72B and Yi-34B also stand out with their exceptional reasoning abilities, positioning themselves between the GPT-3.5 and GPT-4 models. This not only underscores the progress in developing LLMs in different languages but also indicates the potential for LLMs to improve both knowledge and reasoning by considering a wider range of languages when constructing training corpora.

A.5 Detailed Character List

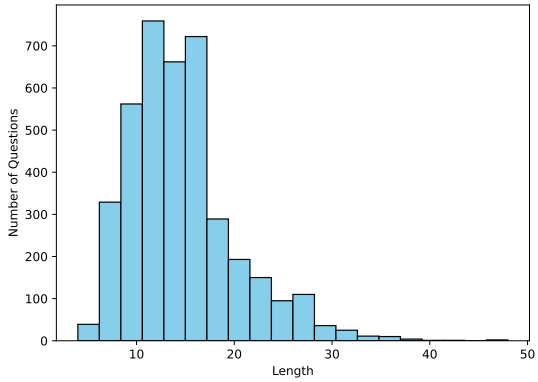
We list all the collected 300 characters in Table 9, 10, 11, 12, 13 and 14, with both the name and source of each character in Chinese and English.

Model	RoleEval-Global (4,000 questions)						RoleEval-Chinese (2,000 questions)					
	CE	AC	MT	GA	FI	Avg.	CE	AC	MT	GA	FI	Avg.
GPT-3.5-0613	57.00	59.00	56.12	59.13	57.50	57.75	37.75	44.50	43.75	43.25	45.00	42.85
GPT-3.5-1106	55.50	53.62	53.50	56.00	53.50	54.42	39.25	40.50	42.00	39.00	42.00	40.55
GPT-4-0613	74.38	74.12	**73.88**	72.75	71.25	73.28	48.50	56.00	**60.50**	58.75	55.25	55.80
GPT-4-1106	**74.38**	**78.62**	72.38	**74.38**	**73.00**	**74.55**	51.00	**56.75**	59.75	**60.25**	**57.75**	**57.10**
Falcon-7B	28.38	26.75	23.12	26.12	31.37	27.15	26.50	24.50	30.75	28.75	24.00	26.90
Falcon-40B	47.12	43.88	47.12	42.38	48.88	45.88	37.50	36.75	30.25	38.25	36.75	35.90
LLaMA-7B	32.00	31.50	29.38	31.00	32.88	31.35	23.00	24.75	29.25	30.25	29.25	27.30
LLaMA-13B	35.88	36.50	37.00	32.25	42.25	36.78	33.50	28.25	29.25	34.75	30.75	31.30
LLaMA-30B	45.62	41.62	47.12	41.88	51.00	45.45	35.75	31.75	35.00	38.00	36.00	35.30
LLaMA-65B	53.62	49.75	50.88	41.75	54.75	50.15	38.00	35.25	35.00	40.00	37.25	37.10
LLaMA-2-7B	33.62	32.88	31.87	33.88	37.50	33.95	25.25	27.00	30.50	29.25	31.75	28.75
LLaMA-2-13B	48.88	45.62	48.75	42.50	47.50	46.65	33.25	38.25	35.50	39.75	29.75	35.30
LLaMA-2-70B	58.50	54.37	57.75	47.88	61.00	55.90	37.75	41.50	39.25	43.50	42.50	40.90
Mistral-7B	53.87	48.38	48.50	44.75	50.12	49.12	37.50	36.00	33.25	38.50	36.00	36.25
MiniMax	54.50	55.88	52.50	54.12	53.00	54.00	36.75	40.50	41.00	40.25	40.50	39.80
Baichuan2-7B	52.12	43.75	44.88	39.62	49.00	45.87	35.25	34.00	30.00	38.50	34.25	34.40
Baichuan2-13B	55.00	47.38	48.12	42.62	54.25	49.47	39.00	38.50	30.75	40.75	35.50	36.90
ChatGLM3-6B	52.12	46.75	51.62	43.75	52.62	49.37	34.75	37.50	35.00	41.00	39.75	37.60
Chinese-LLaMA-2-7B	29.00	26.62	22.38	25.50	30.25	26.75	23.75	22.75	31.75	26.75	26.75	26.35
Chinese-LLaMA-2-13B	41.00	43.38	43.75	40.25	45.38	42.75	31.50	33.00	33.00	35.00	33.25	33.15
Qwen-7B	53.87	43.88	49.00	40.50	50.88	47.63	35.25	31.75	36.25	39.75	35.75	35.75
Qwen-14B	59.62	50.88	54.37	44.62	55.88	53.07	42.25	42.25	37.50	44.50	41.75	41.65
Qwen-72B	69.00	62.25	67.88	57.12	68.12	64.87	48.00	47.00	46.25	51.25	55.00	49.50
Skywork-13B	53.62	46.25	50.25	41.38	53.62	49.02	36.75	36.50	33.25	42.00	37.50	37.20
Yi-6B	57.38	51.50	52.88	43.62	53.25	51.73	39.75	39.50	37.25	42.00	42.00	40.10
Yi-34B	67.00	61.62	63.00	54.00	62.50	61.62	**52.25**	50.00	39.50	51.75	49.25	48.55

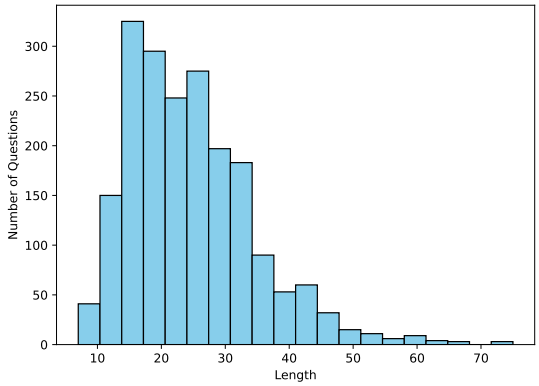
Table 6: Zero-shot results on RoleEval (**en**) in five categories: celebrities (**CE**), anime and comics (**AC**), movie and TV series (**MT**), games (**GA**), and fiction (**FI**).



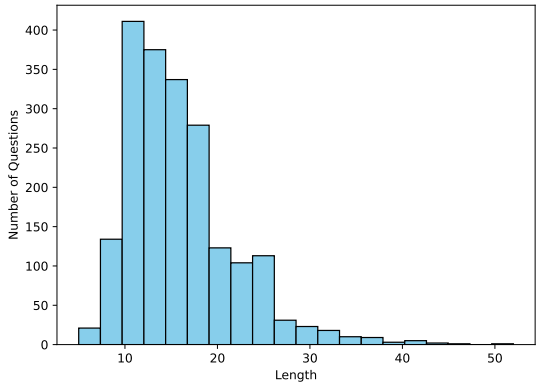
(a) *RoleEval-Global* (**zh**)



(b) *RoleEval-Global* (**en**)



(c) *RoleEval-Chinese* (**zh**)



(d) *RoleEval-Chinese* (**en**)

Figure 5: Visualization of the length distribution of questions in RoleEval.

Model	RoleEval-Global (4,000 questions)				RoleEval-Chinese (2,000 questions)			
	Knowledge (3,400 questions)		Reasoning (600 questions)		Knowledge (1,700 questions)		Reasoning (300 questions)	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
GPT-3.5-0613	46.06	48.47	45.83	50.17	39.06	42.47	30.33	37.00
GPT-3.5-1106	44.03	49.53	49.67	52.83	38.65	45.12	36.00	37.00
GPT-4-0613	71.09	72.74	**70.83**	**70.00**	58.94	59.65	56.33	53.67
GPT-4-1106	**73.76**	**73.97**	70.50	**70.00**	61.00	63.59	57.67	58.00
Falcon-7B	25.91	27.06	24.00	23.50	25.88	28.53	23.00	27.33
Falcon-40B	37.32	35.65	40.67	37.00	32.47	32.29	28.33	29.33
LLaMA-7B	28.74	28.12	25.17	24.83	29.12	26.53	26.00	23.67
LLaMA-13B	26.29	28.03	33.17	27.17	25.76	28.41	27.67	24.67
LLaMA-30B	27.38	28.24	24.33	30.17	26.53	28.82	27.33	29.33
LLaMA-65B	29.94	32.62	33.00	32.00	27.82	31.76	32.33	27.67
LLaMA-2-7B	29.59	33.38	28.00	35.33	28.06	30.47	26.33	29.67
LLaMA-2-13B	31.29	33.41	29.83	33.50	27.59	28.59	27.33	28.00
LLaMA-2-70B	41.18	44.47	42.17	46.00	35.18	36.65	31.33	34.00
Mistral-7B	32.32	33.15	30.33	34.67	30.06	32.35	26.00	31.33
MiniMax	54.82	54.29	63.83	63.50	52.94	54.00	58.67	58.33
Baichuan2-7B	46.76	48.32	53.17	51.50	48.18	49.59	43.67	48.67
Baichuan2-13B	52.35	54.03	51.67	55.50	51.24	53.41	44.33	49.33
ChatGLM3-6B	49.82	48.79	52.33	55.33	49.18	48.65	48.33	50.67
Chinese-LLaMA-2-7B	26.62	34.94	31.00	39.00	26.12	33.53	28.00	27.33
Chinese-LLaMA-2-13B	33.12	39.06	38.83	42.00	31.88	35.71	34.33	39.67
Qwen-7B	47.47	47.76	50.00	48.67	45.88	46.41	46.67	49.67
Qwen-14B	53.79	54.68	55.67	54.67	51.47	53.18	47.67	51.67
Qwen-72B	67.18	67.59	64.67	66.83	**65.76**	**67.00**	**59.67**	61.67
Skywork-13B	50.03	52.44	57.83	57.50	51.00	52.76	49.00	51.67
Yi-6B	55.18	54.09	51.67	55.83	55.53	55.88	44.33	52.00
Yi-34B	62.79	65.38	55.67	68.33	62.29	64.71	55.67	**64.00**

Table 7: Results by knowledge and reasoning questions in RoleEval (**zh**).

Model	RoleEval-Global (4,000 questions)				RoleEval-Chinese (2,000 questions)			
	Knowledge (3,400 questions)		Reasoning (600 questions)		Knowledge (1,700 questions)		Reasoning (300 questions)	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
GPT-3.5-0613	57.03	57.88	61.83	61.50	42.71	43.47	43.67	47.00
GPT-3.5-1106	53.15	55.82	61.67	62.50	39.65	43.18	45.67	46.33
GPT-4-0613	72.94	75.79	**75.17**	**77.33**	55.65	60.65	**56.67**	**63.00**
GPT-4-1106	**74.62**	**76.29**	74.17	74.33	**57.82**	**60.71**	53.00	58.00
Falcon-7B	26.50	28.68	30.83	28.00	27.00	28.18	26.33	28.33
Falcon-40B	45.53	46.71	47.83	48.83	36.65	36.41	31.67	29.67
LLaMA-7B	30.74	29.41	34.83	34.83	27.06	28.06	28.67	32.33
LLaMA-13B	35.47	39.35	44.17	47.83	30.88	32.47	33.67	37.00
LLaMA-30B	44.47	47.97	51.00	52.00	34.88	35.18	37.67	35.67
LLaMA-65B	49.41	52.62	54.33	55.33	37.00	38.94	37.67	38.33
LLaMA-2-7B	33.32	38.59	37.50	44.50	28.76	31.82	28.67	33.67
LLaMA-2-13B	46.44	46.35	47.83	46.50	35.82	34.00	32.33	34.00
LLaMA-2-70B	55.41	58.41	58.67	59.50	41.76	43.65	36.00	40.67
Mistral-7B	48.26	48.50	54.00	55.50	36.65	36.12	34.00	36.67
MiniMax	52.44	52.79	62.83	61.17	38.76	38.06	45.67	39.67
Baichuan2-7B	44.79	46.71	52.00	51.67	33.71	35.88	38.33	41.00
Baichuan2-13B	48.97	50.21	52.33	55.17	36.94	36.53	36.67	33.33
ChatGLM3-6B	48.38	48.50	55.00	54.33	37.35	38.59	39.00	39.33
Chinese-LLaMA-2-7B	25.82	34.59	32.00	37.00	26.00	29.41	28.33	32.67
Chinese-LLaMA-2-13B	41.62	46.21	49.17	50.17	33.12	34.12	33.33	35.67
Qwen-7B	47.32	47.35	49.33	50.67	35.53	38.24	37.00	37.67
Qwen-14B	52.44	52.97	56.67	53.83	41.35	40.41	43.33	40.67
Qwen-72B	63.91	64.79	70.33	68.00	49.47	50.82	49.67	47.33
Skywork-13B	47.26	49.44	59.00	56.67	36.94	38.59	38.67	37.67
Yi-6B	51.03	53.35	55.67	56.50	40.00	42.82	40.67	39.33
Yi-34B	61.24	64.94	63.83	68.17	48.53	53.82	48.67	51.67

Table 8: Results by knowledge and reasoning questions in RoleEval (**en**).

Name	Name	Name	Name
爱因斯坦 (Albert Einstein)	安徒生 (Hans Christian Andersen)	李相赫 (Lee Sang-hyeok, known as Faker)	苏格拉底 (Socrates)
亚历山大大帝 (Alexander the Great)	村上春树 (Haruki Murakami)	达芬奇 (Leonardo da Vinci)	史蒂夫·乔布斯 (Steve Jobs)
亚里士多德 (Aristotle)	宫崎骏 (Hayao Miyazaki)	梅西 (Lionel Messi)	西尔维斯特·史泰龙 (Sylvester Stallone)
奥黛丽·赫本 (Audrey Hepburn)	成龙 (Jackie Chan)	鲁迅 (Lu Xun)	泰勒·斯威夫特 (Taylor Swift)
比尔·盖茨 (Bill Gates)	贾斯汀·比伯 (Justin Bieber)	甘地 (Mahatma Gandhi)	托马斯·爱迪生 (Thomas Edison)
李小龙 (Bruce Lee)	马克思 (Karl Marx)	玛丽莲·梦露 (Marilyn Monroe)	汤姆·汉克斯 (Tom Hanks)
达尔文 (Charles Darwin)	金·卡戴珊 (Kim Kardashian)	迈克尔·杰克逊 (Michael Jackson)	梵高 (Vincent van Gogh)
马斯克 (Elon Musk)	孔子 (Kongzi, known as Confucius)	拿破仑 (Napoleon)	莎士比亚 (William Shakespeare)
埃尔维斯·普雷斯利 (Elvis Presley)	权志龙 (Kwon Ji-yong, known as G-Dragon)	曼德拉 (Nelson Mandela)	姚明 (Yao Ming)
成吉思汗 (Genghis Khan)	郎朗 (Lang Lang)	伊丽莎白二世 (Queen Elizabeth II)	羽生结弦 (Yuzuru Hanyu)

Table 9: Characters from real-world celebrities in *RoleEval-Global*.

Name	Source	Name	Source
阿姆罗·雷 (Amuro Ray)	机动战士高达 (*Mobile Suit Gundam*)	蒙奇·D·路飞 (Monkey D. Luffy)	海贼王 (*One Piece*)
花花 (Blossom)	飞天小女警 (*Powerpuff Girls*)	墨菲斯 (Morpheus)	睡魔 (*The Sandman*)
樱桃子 (Chibi Maruko-chan)	樱桃小丸子 (*Chibi Maruko-chan*)	草薙素子 (Motoko Kusanagi)	攻壳机动队 (*Ghost in the Shell*)
江户川柯南 (Conan Edogawa)	名侦探柯南 (*Detective Conan*)	漩涡鸣人 (Naruto Uzumaki)	火影忍者 (*Naruto*)
唐老鸭 (Donald Duck)	米老鼠和唐老鸭 (*Mickey Mouse and Donald Duck*)	绫波丽 (Rei Ayanami)	新世纪福音战士 (*Neon Genesis Evangelion*)
哆啦A梦 (Doraemon)	哆啦A梦 (*Doraemon*)	瑞克 (Rick Sanchez)	瑞克和莫蒂 (*Rick and Morty*)
艾伦·耶格尔 (Eren Yeager)	进击的巨人 (*Attack on Titan*)	越前龙马 (Ryoma Echizen)	网球王子 (*The Prince of Tennis*)
菲·瓦伦坦 (Faye Valentine)	星际牛仔 (*Cowboy Bebop*)	木之本樱 (Sakura Kinomoto)	魔卡少女樱 (*Cardcaptor Sakura*)
加菲猫 (Garfield)	加菲猫 (*Garfield*)	五条悟 (Satoru Gojo)	咒术回战 (*Jujutsu Kaisen*)
坂田银时 (Gintoki Sakata)	银魂 (*Gintama*)	野原新之助 (Shinnosuke Nohara)	蜡笔小新 (*Crayon Shin-chan*)
乔鲁诺·乔巴纳 (Giorno Giovanna)	JOJO的奇妙冒险 (*JoJo's Bizarre Adventure*)	史努比 (Snoopy)	花生漫画 (*Peanuts*)
樱木花道 (Hanamichi Sakuragi)	灌篮高手 (*Slam Dunk*)	白雪公主 (Snow White)	白雪公主和七个小矮人 (*Snow White and the Seven Dwarfs*)
地狱男爵 (Hellboy)	地狱男爵系列 (*Hellboy Series*)	海绵宝宝 (SpongeBob SquarePants)	海绵宝宝 (*SpongeBob SquarePants*)
黑崎一护 (Ichigo Kurosaki)	BLEACH (*Bleach*)	超人 (Superman)	超人 (*Superman*)
杰瑞 (Jerry)	猫和老鼠 (*Tom and Jerry*)	夏目贵志 (Takashi Natsume)	夏目友人帐 (*Natsume's Book of Friends*)
桔梗 (Kikyo)	犬夜叉 (*Inuyasha*)	天津饭 (Tien Shinhan)	龙珠 (*Dragon Ball*)
露西 (Lucy)	赛博朋克：边缘行者 (*Cyberpunk: Edgerunners*)	丁丁 (Tintin)	丁丁历险记 (*The Adventures of Tintin*)
玛奇玛 (Makima)	链锯人 (*Chainsaw Man*)	薇尔莉特·伊芙加登 (Violet Evergarden)	紫罗兰永恒花园 (*Violet Evergarden*)
御坂美琴 (Mikoto Misaka)	某科学的超电磁炮 (*A Certain Scientific Railgun*)	维尼 (Winnie the Pooh)	小熊维尼 (*Winnie the Pooh*)
宫水三叶 (Mitsuha Miyamizu)	你的名字 (*Your Name*)	金刚狼 (Wolverine)	金刚狼系列 (*Wolverine Series*)

Table 10: Characters from anime and comics in *RoleEval-Global*.

Name	Source	Name	Source
亚伦·霍奇纳 (Aaron Hotchner)	犯罪心理 (*Criminal Minds*)	莱昂 (Léon)	这个杀手不太冷 (*Léon: The Professional*)
阿蕾莎 (Alessa Gillespie)	寂静岭 (*Silent Hill*)	万磁王 (Magneto)	X战警 (*X-Men*)
巴斯光年 (Buzz Lightyear)	玩具总动员 (*Toy Story*)	威震天 (Megatron)	变形金刚 (*Transformers*)
美国队长 (Captain America)	复仇者联盟 (*The First Avenger*)	迈克尔·斯科菲尔德 (Michael Scofield)	越狱 (*Prison Break*)
卡莉·马西森 (Carrie Mathison)	国土安全 (*Homeland*)	米格尔·里韦拉斯 (Miguel Rivera)	寻梦环游记 (*Coco*)
克莱尔·邓菲 (Claire Dunphy)	摩登家庭 (*Modern Family*)	三澄美琴 (Mikoto Misumi)	非自然死亡 (*Unnatural*)
丹妮莉丝·坦格利安 (Daenerys Targaryen)	权力的游戏 (*Game of Thrones*)	永尾完治 (Nagao Kanji)	东京爱情故事 (*Tokyo Love Story*)
达里尔·迪克森 (Daryl Dixon)	行尸走肉 (*The Walking Dead*)	尼奥 (Neo)	黑客帝国 (*The Matrix*)
戴安娜·普林斯 (Diana Prince)	神奇女侠 (*Wonder Woman*)	艾莎公主 (Princess Elsa)	冰雪奇缘 (*Frozen*)
伊丽莎白·斯旺 (Elizabeth Swann)	加勒比海盗 (*Pirates of the Caribbean*)	瑞秋·格林 (Rachel Green)	老友记 (*Friends*)
艾伦·雷普莉 (Ellen Ripley)	异形系列 (*Alien series*)	拉斐尔（忍者神龟） (Raphael)	忍者神龟 (*Teenage Mutant Ninja Turtles*)
阿甘 (Forrest Gump)	阿甘正传 (*Forrest Gump*)	谢尔顿·库珀 (Sheldon Cooper)	生活大爆炸 (*The Big Bang Theory*)
弗兰西斯·安德伍德 (Francis Underwood)	纸牌屋 (*House of Cards*)	辛巴 (Simba)	狮子王 (*The Lion King*)
格鲁 (Gru)	神偷奶爸 (*Despicable Me*)	托尼·史塔克 (Tony Stark)	钢铁侠 (*Iron Man*)
哈利·波特 (Harry Potter)	哈利波特 (*Harry Potter*)	迪迦奥特曼 (Ultraman Tiga)	迪迦奥特曼 (*Ultraman Tiga*)
印第安纳·琼斯 (Indiana Jones)	夺宝奇兵 (*Indiana Jones*)	毒液 (Venom)	毒液系列 (*Venom series*)
杰克·萨利 (Jake Sully)	阿凡达 (*Avatar*)	维托·柯里昂 (Vito Corleone)	教父 (*The Godfather*)
詹姆斯·邦德 (James Bond)	007系列 (*James Bond series*)	瓦力 (WALL-E)	机器人总动员 (*WALL-E*)
约翰·麦克莱恩 (John McClane)	龙胆虎威 (*Die Hard*)	沃尔特·怀特 (Walter White)	绝命毒师 (*Breaking Bad*)
卢克·天行者 (Luke Skywalker)	星球大战 (*Star Wars*)	甄嬛 (Zhen Huan)	甄嬛传 (*Empresses in the Palace*)

Table 11: Characters from movies and TV series in *RoleEval-Global*.

Name	Source	Name	Source
亚特鲁·克里斯汀 (Adol Christin)	伊苏系列 (*Ys Series*)	桐生一马 (Kazuma Kiryu)	如龙系列 (*Yakuza Series*)
杀手47 (Agent 47)	杀手系列 (*Hitman Series*)	奎托斯 (Kratos)	战神 (*God of War*)
艾丹 (Aidan)	暗黑破坏神 (*Diablo*)	约翰-117 (Master Chief Petty Officer John-117)	光环 (*Halo*)
艾登·皮尔斯 (Aiden Pearce)	看门狗 (*Watch Dogs*)	洛克人 (Mega Man)	洛克人系列 (*Mega Man Series*)
艾伦·韦克 (Alan Wake)	心灵杀手 (*Alan Wake*)	米亚·卡恩斯坦 (Mia Karnstein)	噬血代码 (*Code Vein*)
埃洛伊 (Aloy)	地平线系列 (*Horizon Series*)	蒙葛特 (Morgott)	艾尔登法环 (*Elden Ring*)
阿尔萨斯·米奈希尔 (Arthas Menethil)	魔兽 (*Warcraft*)	纳西妲 (Nahida)	原神 (*Genshin Impact*)
亚瑟·摩根 (Arthur Morgan)	荒野大镖客 (*Red Dead Redemption*)	内森·德雷克 (Nathan Drake)	神秘海域系列 (*Uncharted Series*)
贝优妮塔 (Bayonetta)	猎天使魔女 (*Bayonetta*)	山姆·费舍尔 (Sam Fisher)	细胞分裂 (*Splinter Cell*)
春丽 (Chun-Li)	街头霸王 (*Street Fighter*)	萨姆斯·阿兰 (Samus Aran)	银河战士系列 (*Metroid Series*)
薛帕德 (Commander Shepard)	质量效应系列 (*Mass Effect Series*)	马力欧 (Mario)	超级马力欧兄弟 (*Super Mario Bros*)
但丁 (Dante)	鬼泣 (*Devil May Cry*)	希侬·埃梅利斯 (Shionne Imeris)	破晓传说 (*Tales of Arise*)
龙裔 (Dragonborn)	上古卷轴5 (*The Elder Scrolls V: Skyrim*)	索利德·斯内克 (Solid Snake)	合金装备系列 (*Metal Gear Series*)
艾黛尔贾特·冯·弗雷斯贝尔古 (Edelgard von Hresvelg)	火焰之纹章：风花雪月 (*Fire Emblem: Three Houses*)	索拉 (Sora)	王国之心 (*Kingdom Hearts*)
爱莉希雅 (Elysia)	崩坏3 (*Honkai Impact 3rd*)	猎空 (Tracer)	守望先锋 (*Overwatch*)
艾吉奥 (Ezio Auditore da Firenze)	刺客信条2 (*Assassin's Creed II*)	崔佛·菲利普 (Trevor Philips)	GTA5 (*Grand Theft Auto V*)
杰洛特 (Geralt of Rivia)	巫师 (*The Witcher*)	瓦尔基里 (Valkyrie)	Apex英雄 (*Apex Legends*)
GLaDOS (GLaDOS)	传送门 (*Portal*)	威廉·亚当斯 (William Adams)	仁王系列 (*Nioh Series*)
苇名一心 (Isshin Ashina)	只狼：影逝二度 (*Sekiro: Shadows Die Twice*)	尤尔哈2B (YoRHa No.2 Type B (2B))	尼尔：机械纪元 (*NieR: Automata*)
詹姆斯·波特 (James Porter)	彩虹六号：围攻 (*Tom Clancy's Rainbow Six Siege*)	塞尔达 (Zelda)	塞尔达传说 (*The Legend of Zelda*)

Table 12: Characters from games in *RoleEval-Global*.

Name	Source	Name	Source
爱丽丝 (Alice)	爱丽丝梦游仙境 (*Alice's Adventures in Wonderland*)	利奥波德·布卢姆 (Leopold Bloom)	尤利西斯 (*Ulysses*)
阿迪克斯·芬奇 (Atticus Finch)	杀死一只知更鸟 (*To Kill a Mockingbird*)	林黛玉 (Lin Daiyu)	红楼梦 (*Dream of the Red Chamber*)
曹操 (Cao Cao)	三国演义 (*Romance of the Three Kingdoms*)	达西 (Mr. Darcy)	傲慢与偏见 (*Pride and Prejudice*)
亚哈船长 (Captain Ahab)	白鲸 (*Moby-Dick*)	奥利弗·崔斯特 (Oliver Twist)	雾都孤儿 (*Oliver Twist*)
尼摩船长 (Captain Nemo)	海底两万里 (*Twenty Thousand Leagues Under the Seas*)	保罗·厄崔迪 (Paul Atreides)	沙丘 (*Dune*)
爱德蒙·邓蒂斯 (Edmond Dantès)	基督山伯爵 (*The Count of Monte Cristo*)	保尔·柯察金 (Paul Korchagin)	钢铁是怎样炼成的 (*How the Steel Was Tempered*)
爱德华·卡伦 (Edward Cullen)	暮光之城 (*Twilight*)	波西·杰克逊 (Percy Jackson)	波西杰克逊系列 (*Percy Jackson & the Olympians*)
葛朗台 (Eugénie Grandet)	欧也妮·葛朗台 (*Eugénie Grandet*)	皮埃尔·别祖霍夫 (Pierre Bezukhov)	战争与和平 (*War and Peace*)
芬恩·麦克库尔 (Finn McCool)	芬尼亚传奇 (*The Fenian Cycle*)	卡西莫多 (Quasimodo)	巴黎圣母院 (*The Hunchback of Notre-Dame*)
甘道夫 (Gandalf)	霍比特人 (*The Hobbit*)	鲁滨逊·克鲁索 (Robinson Kreutznaer)	鲁宾逊漂流记 (*Robinson Crusoe*)
格列佛 (Gulliver)	格列佛游记 (*Gulliver's Travels*)	桐原亮司 (Ryōji Kirihara)	白夜行 (*Journey Under the Midnight Sun*)
希斯克利夫 (Heathcliff)	呼啸山庄 (*Wuthering Heights*)	索伦 (Sauron)	魔戒 (*The Lord of the Rings*)
赫尔克里·波洛 (Hercule Poirot)	东方列车谋杀案 (*Murder on the Orient Express*)	斯嘉丽·奥哈拉 (Scarlett O'Hara)	飘 (*Gone with the Wind*)
霍尔顿·考尔菲德 (Holden Caulfield)	麦田里的守望者 (*The Catcher in the Rye*)	夏洛克·福尔摩斯 (Sherlock Holmes)	福尔摩斯探案集 (*The Adventures of Sherlock Holmes*)
简·爱 (Jane Eyre)	简·爱 (*Jane Eyre*)	宋江 (Song Jiang)	水浒传 (*Water Margin*)
盖茨比 (Jay Gatsby)	了不起的盖茨比 (*The Great Gatsby*)	孙悟空 (Sun Wukong)	西游记 (*Journey to the West*)
冉·阿让 (Jean Valjean)	悲惨世界 (*Les Misérables*)	汤姆·索亚 (Tom Sawyer)	汤姆索亚历险记 (*The Adventures of Tom Sawyer*)
于连 (Julien Sorel)	红与黑 (*The Red and the Black*)	维克多·弗兰肯斯坦 (Victor Frankenstein)	弗兰肯斯坦 (*Frankenstein*)
凯特尼斯·伊夫狄恩 (Katniss Everdeen)	饥饿游戏 (*The Hunger Games*)	温斯顿·史密斯 (Winston Smith)	1984 (*1984*)
加贺恭一郎 (Kyōichirō Kaga)	加贺恭一郎系列 (*Police Detective Kaga series*)	叶文洁 (Ye Wenjie)	三体 (*The Three-Body Problem*)

Table 13: Characters from fiction in *RoleEval-Global*.

Name	Source	Category	Name	Source	Category
何炅 (He Jiong)	-	celebrities	路明非 (Lu Mingfei)	龙族 (*Dragon Raja*)	fiction
周杰伦 (Jay Chou)	-	celebrities	任盈盈 (Ren Yingying)	笑傲江湖 (*The Smiling, Proud Wanderer*)	fiction
李敏镐 (Lee Min-ho)	-	celebrities	顷莲 (Song Lian)	妻妾成群 (*Wives and Concubines*)	fiction
雷军 (Lei Jun)	-	celebrities	孙少平 (Sun Shaoping)	平凡的世界 (*Ordinary World*)	fiction
李白 (Li Bai)	-	celebrities	唐三 (Tang San)	斗罗大陆 (*Douluo Dalu*)	fiction
李荣浩 (Li Ronghao)	-	celebrities	祥子 (Xiangzi)	骆驼祥子 (*Rickshaw Boy*)	fiction
李宇春 (Li Yuchun)	-	celebrities	徐凤年 (Xu Fengnian)	雪中悍刀行 (*Sword Snow Stride*)	fiction
李子柒 (Li Ziqi)	-	celebrities	张起灵 (Zhang Qiling)	盗墓笔记 (*The Lost Tomb*)	fiction
刘翔 (Liu Xiang)	-	celebrities	张无忌 (Zhang Wuji)	倚天屠龙记 (*The Heaven Sword and Dragon Saber*)	fiction
刘亦菲 (Liu Yifei)	-	celebrities	张小凡 (Zhang Xiaofan)	诛仙 (*Zhu Xian*)	fiction
马伊 (Ma Yili)	-	celebrities	奥古斯特·奥特姆 (Augustus Autumn)	辐射3 (*Fallout 3*)	games
秦始皇 (Qin Shi Huang)	-	celebrities	克劳德 (Cloud Strife)	最终幻想7 (*Final Fantasy VII*)	games
孙燕姿 (Stefanie Sun)	-	celebrities	爱梅斯 (Emes)	公主连结Re:Dive (*Princess Connect! Re:Dive*)	games
木村拓哉 (Takuya Kimura)	-	celebrities	艾丝蒂尔·布莱特 (Estelle Bright)	英雄传说轨迹系列 (*The Legend of Heroes: Trails Series*)	games
王鹤棣 (Wang Hedi)	-	celebrities	能天使 (Exusiai)	明日方舟 (*Arknights*)	games
王俊凯 (Wang Junkai)	-	celebrities	盖伦 (Garen)	英雄联盟 (*League of Legends*)	games
杨振宁 (Yang Chen-Ning)	-	celebrities	季沧海 (Ji Canghai)	永劫无间 (*Naraka: Bladepoint*)	games
岳云鹏 (Yue Yunpeng)	-	celebrities	约翰·普莱斯 (John Price)	使命召唤 (*Call of Duty*)	games
张艺谋 (Zhang Yimou)	-	celebrities	三岛一八 (Kazuya Mishima)	铁拳系列 (*Tekken series*)	games
张艺兴 (Zhang Yixing)	-	celebrities	李逍遥 (Li Xiaoyao)	仙剑奇侠传 (*Chinese Paladin*)	games
阿尼亚·福杰 (Anya Forger)	间谍过家家 (*Spy x Family*)	anime and comics	马可斯·菲尼克斯 (Marcus Fenix)	战争机器系列 (*Gears of War series*)	games
小智 (Ash Ketchum)	宝可梦 (*Pokémon*)	anime and comics	米拉波雷亚斯 (Milla Borealis)	怪物猎人 (*Monster Hunter*)	games
荆天明 (Jing Tianming)	秦时明月 (*Qin's Moon*)	anime and comics	蕾米莉亚·斯卡蕾特 (Remilia Scarlet)	东方project (*Touhou Project*)	games
四宫辉夜 (Kaguya Shinomiya)	辉夜大小姐想让我告白 (*Kaguya-sama: Love Is War*)	anime and comics	赛丽亚·克鲁敏 (Seria Kirmin)	地下城与勇士 (*Dungeon Fighter Online*)	games
奇·接救客 (Killua Zoldyck)	全职猎人 (*Hunter × Hunter*)	anime and comics	史特列洛克 (Strelok)	潜行者系列 (*S.T.A.L.K.E.R. series*)	games
马克·霍夫曼 (Marc)	灵笼 (*Ling Cage: Incarnation*)	anime and comics	塔里昂 (Talion)	中土世界：战争之影 (*Middle-earth: Shadow of War*)	games
中野三玖 (Miku Nakano)	五等分的新娘 (*5-toubun no Hanayome*)	anime and comics	狸克 (Tom Nook)	动物森友会 (*Animal Crossing*)	games
早濑未沙 (Misa Hayase)	超时空要塞 (*Super Dimension Fortress Macross*)	anime and comics	谢衣 (Xie Yi)	古剑奇谭2 (*Gujian2*)	games
桃园奈奈生 (Nanami Momozono)	元气少女缘结神 (*Kamisama Kiss*)	anime and comics	杨宁 (Yang Ning)	剑侠情缘网络版叁 (*Swordsman Online*)	games
喜羊羊 (Pleasant Goat)	喜羊羊与灰太狼 (*Pleasant Goat and Big Big Wolf*)	anime and comics	尤里 (Yuri)	红色警戒 (*Command & Conquer: Red Alert*)	games
红细胞 (Red Blood Cell)	工作细胞 (*Cells at Work!*)	anime and comics	都敏俊 (Do Min-joon)	来自星星的你 (*My Love from the Star*)	movies and TV series
成步堂龙一 (Ryūichi Naruhodō)	逆转裁判 (*Ace Attorney*)	anime and comics	樊胜美 (Fan Shengmei)	欢乐颂 (*Ode to Joy*)	movies and TV series
玉 (Saitama)	一拳超人 (*One Punch Man*)	anime and comics	东方青苍 (Dongfang Qingcang)	苍兰诀 (*Love Between Fairy and Devil*)	movies and TV series
菜月昴 (Subaru Natsuki)	Re:从零开始的异世界生活 (*Re:Zero - Starting Life in Another World*)	anime and comics	范闲 (Fan Xian)	庆余年 (*Joy of Life*)	movies and TV series
炭门炭治郎 (Tanjiro Kamado)	鬼灭之刃 (*Demon Slayer: Kimetsu no Yaiba*)	anime and comics	顾千帆 (Gu Qianfan)	梦华录 (*A Dream of Splendor*)	movies and TV series
安艺伦也 (Tomoya Aki)	路人女主的养成方法 (*Saekano: How to Raise a Boring Girlfriend*)	anime and comics	好彩妹 (Hao Cai Mei)	搜神传 (*Legend of the Demigods*)	movies and TV series
尤尼 (Tsunayoshi "Tsuna" Sawada)	家庭教师 (*Katekyo Hitman Reborn!*)	anime and comics	洪世贤 (Hong Shixian)	回家的诱惑 (*The Temptation to Go Home*)	movies and TV series
涂山苏苏 (Tushan Susu)	妖妖小红娘 (*Fox Spirit Matchmaker*)	anime and comics	胡一菲 (Hu Yifei)	爱情公寓 (*iPartment*)	movies and TV series
月野兔 (Usagi Tsukino)	美少女战士 (*Sailor Moon*)	anime and comics	花千骨 (Hua Qiangu)	花千骨 (*The Journey of Flower*)	movies and TV series
袁天罡 (Yuan Tiangang)	画江湖之不良人 (*Hua Jiang Hu Zhi Bu Liang Ren*)	anime and comics	高逸 (Ko Chun)	赌神系列 (*God of Gamblers series*)	movies and TV series
白素贞 (Bai Suzhen)	白蛇传 (*Legend of the White Snake*)	fiction	洛晴川 (Luo Qingchuan)	宫锁心玉 (*Palace*)	movies and TV series
陈平安 (Chen Ping'an)	剑来 (*Sword of Coming*)	fiction	平泽直树 (Naoki Hanzawa)	半泽直树 (*Naoki Hanzawa*)	movies and TV series
楚留香 (Chu Liuxiang)	楚留香传奇 (*The Legend of Chu Liuxiang*)	fiction	盛明兰 (Sheng Minglan)	知否知否应是绿肥红瘦 (*The Story of MingLan*)	movies and TV series
崔莺莺 (Cui Yingying)	崔莺莺待月西厢记 (*Romance of the Western Chamber*)	fiction	武媚娘 (Wu Meiniang)	武媚娘传奇 (*The Empress of China*)	movies and TV series
戴凤莲 (Dai Fenglian)	红高粱家族 (*Red Sorghum Family*)	fiction	小燕子 (Xiao Yanzi)	还珠格格 (*My Fair Princess*)	movies and TV series
封三娘 (Feng Sanniang)	聊斋志异 (*Strange Stories from a Chinese Studio*)	fiction	姚金铃 (Yao Jinling)	宫心计 (*Beyond the Realm of Conscience*)	movies and TV series
何以琛 (He Yichen)	何以笙箫默 (*Silent Separation*)	fiction	余则成 (Yu Zecheng)	潜伏 (*Lurk*)	movies and TV series
胡八一 (Hu Bayi)	鬼吹灯 (*Ghost Blows Out the Light*)	fiction	余罪 (Yu Zui)	余罪 (*Yu Zui*)	movies and TV series
花城 (Hua Cheng)	天官赐福 (*Heaven Official's Blessing*)	fiction	张麻子 (Zhang Mazi)	让子弹飞 (*Let the Bullets Fly*)	movies and TV series
孔乙己 (Kong Yiji)	孔乙己 (*Kong Yiji*)	fiction	紫霞仙子 (Zixia Fairy)	大话西游 (*A Chinese Odyssey*)	movies and TV series

Table 14: Characters in *RoleEval-Chinese*.