DMT-RoleBench: A Dynamic Multi-Turn Dialogue Based Benchmark for Role-Playing Evaluation of Large Language Model and Agent

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, Song Liu*

Ant Group

{zhouchunyi.zcy,cyp430380,guodong.lgd,lichenchen.lcc,tangchengfu.tcf,
jingxi.zdx,wangzhenkui.wzk,xiumin.wxd,yanda.ls}@antgroup.com

Abstract

Recent years have witnessed a profound evolution in the abilities of Large Language Model, which has significantly boosted the proliferation of role-playing agents and platforms. Nonetheless, there is a conspicuous absence of systematic and comprehensive evaluations of role-playing abilities which are truly aligned with users' interaction scenarios in real-world. To address this gap, we have devised DMT-RoleBench, a benchmark designed to evaluate the role-playing abilities of large language models and agents based on dynamic multi-turn dialogues. Compared with existed role-playing benchmarks, DMT-RoleBench boasts several principal advantages: (1) It contains a more diverse role types and system prompts of different formats. (2) We propose an innovative evaluation paradigm to assess role-playing abilities based on dynamically generating multi-turn dialogues constrained by specific evaluation intents and topics, which is well aligned with users' interaction scenarios in real-world. (3) We define a three-tiered metric system and provide DMT-RM, which is a judge model aligned with human annotations, to annotate the dialogues. And we propose DMT-Score to calculate the final scores based on the annotated dialogues. Our experiments and analysis of leading models equipped with role-playing abilities have demonstrated the effectiveness of DMT-RoleBench.

Code —

https://github.com/DMT-RoleBench/DMT-RoleBench

Introduction

The rapid advancement in foundational capabilities of Large Language Models in recent years(Brown et al. 2020; OpenAI et al. 2024), including creative text generation, in-context learning, instruction-following, and knowledge acquisition, has facilitated notable progress in the iterative refinement of role-playing models and agents (Brown et al. 2020; Park et al. 2023; Xi et al. 2023; Tian et al. 2023; Wang et al. 2024a). These models, alongside agents, can not only acquire knowledge pertinent to various roles but also emulate the linguistic styles, behavioral patterns, and personality traits of corresponding personas. Such advancements have given rise to a plethora of role-playing agents and application platforms(Coze 2023; Character.ai 2023). The emergence of these sophisticated agents and platforms, such as casual chat bots based on fictional roles, professional assistants for various occupations, and immersive Multiple Massive Online games, signifies a new era where artificial intelligence blurs the lines between virtual and real-world interactions, significantly enhancing human's experiences and daily life across multiple domains.

However, there is a significant dearth of systematic and comprehensive evaluation benchmark that is well aligned with real-world interaction scenarios for assessing the role-playing capabilities of Large Language Models (LLMs) and agents(Chen et al. 2024b).

RoleEval(Shen et al. 2024) primarily concentrates on assessing the model's proficiency in role knowledge acquisition and reasoning. TimeChara(Ahn et al. 2024) and RoleAD(Tang et al. 2024), on the other hand, focus on constructing misleading character information and trap questions to assess the ability of models being tested to identify and correct errors. CharacterEval(Tu et al. 2024) assesses the role-playing ability by challenging models to complete pre-designed ground truth conversations between two fictional characters from Chinese film and TV series.

Role knowledge acquisition and misleading information identification are just merely facets of a larger picture. Dialogue completion does not align with real-world application scenarios, since the pre-designed ground truth conversations are not genuine outputs of the model or agent being tested, rather, they primarily test the model's capability to mimic based on in-context learning. In essence, a more comprehensive approach that encompasses diverse role types and evaluates models under conditions which are well aligned with users' real-world interactions would be necessary to fully assess the role-playing competencies of LLMs and agents. To address the aforementioned gaps, this paper introduces DMT-RoleBench, a Chinese role-playing evaluation benchmark based on dynamic multi-turn dialogue. Figure 1 illustrates the overview construction and evaluation flow of the benchmark.

Recognizing that the quality of role-playing cannot be adequately assessed through single-turn question-and-answer sessions, and acknowledging that methods like CharacterEval's (Tu et al. 2024) use of pre-defined contexts for di-

*Corresponding author

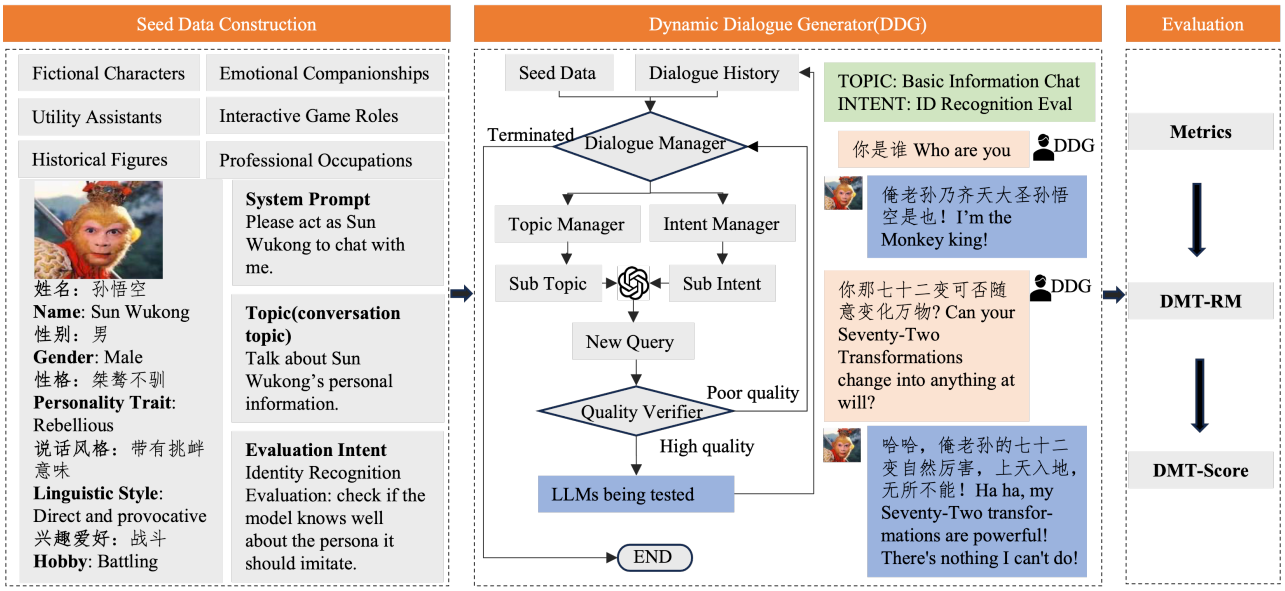

Figure 1: Overview of the construction and evaluation flow of DMT-RoleBench. Firstly, the seed dataset is meticulously designed based on different role types, system prompts, topics and evaluation intents. Then, we dynamically generate multi-turn dialogues which are guided and constrained by the seed dataset. Finally, we defined a three-tiered metric system to assess the role-playing ability. For each metric, we trained DMT-RM, a judge model which is aligned well with crowd-workers annotations, to annotate each dialogue round as good or bad. We then computed the final scores for each metric using the proposed DMT-Score methodology.

alogue completion do not effectively evaluate models' real role-playing capabilities, DMT-RoleBench provides a dynamic multi-turn dialogue-based evaluation methodology. This approach provides an agent to dynamically chat with the models being tested to accurately evaluate the role-playing proficiency aligned well with real-world scenarios. Notably, the dynamic conversations should be guided and constrained by specific evaluation intents and topics to prevent uncontrolled divergence during dynamic multi-turn dialogue generation. Therefore, we have formulated seven major evaluation intentions, which is delineated in Table 1.

To construct the seed data for the dynamic conversation, we mainly take diversity of role types and system prompts into account. Similar to (Chen et al. 2024b), we address the gap of limited diversity of role type in existed datasets by meticulously designing six distinct role types: fictional characters from films, novels, and TV series (e.g., Sun Wukong), historical figures (e.g., Lu Xun), professional occupations (e.g., doctors), emotional companionship roles (e.g., virtual boyfriends/girlfriends), utility assistants (e.g., text generation assistant), and interactive gaming Non-Player Characters (e.g., MMO game npcs). Given that various role-playing platforms(Character.ai 2023) permit users to customize roles, evaluating the performance of LLMs in role-playing across different complexities and formats of system prompts becomes critical. To explore this, we have devised multiple system prompts for the same role to investigate how system prompts impact role-playing effectiveness.

To evaluate role-playing capabilities, we have defined a three-tiered metrics system, encompassing three major categories: basic ability, conversational ability, and imitation ability, further broken down into 14 sub-metrics. Acknowledging that not all metrics may manifest in every conversation, we dynamically allocate the metrics based on the evaluation intents and role types. The allocation strategy is in Table 2.

To better assess the role-playing capabilities in multi-turn conversations, we trained DMT-RM, a judge model based on crowd-workers annotation results, to label each round of the dialogue as **good** or **bad**. Additionally, we proposed the DMT-Score method to compute the final score based on the annotations.

In summary, our principal contributions are delineated as follows:

- DMT-RoleBench provides a multi-turn seed dataset of 996 samples, consisting of 6 different role types, 7 different evaluation intents, and diverse system prompts formats.
- We propose a new role-playing evaluation paradigm that provides an agent to dynamically chat with the models being tested, guided and constrained by specific evaluation intents and topics. This approach is better aligned with real-world scenarios compared with other methods like multiple-choices or conversation completions.
- We develop DMT-RM, a judge model for annotating multi-turn dialogues to evaluate role-playing ability, which is demonstrated to be aligned better with human annotations than GPT-4-Turbo. And we propose DMT-Score to calculate the final scores, which is specially applicable for multi-turn dialogues evaluation.

Related Work

Role-Playing LLM & Agent

The emergence of capabilities in large models, particularly the enhancements in long-text processing, instruction-following, in-context learning, knowledge acquisition, and reasoning abilities, has led to an increasing number of researchers delving into studies on role-playing capabilities of large language models(Tao et al. 2024; Xu et al. 2024; Wu et al. 2024; Lu et al. 2024; Chuang et al. 2024; Shao et al. 2023; Zheng, Pei, and Jurgens 2023). The models' ability to emulate specific roles, imitate linguistic styles, and imitate personality traits has become a focal point.

Previous work primarily utilize psychological interviews(Wang et al. 2024c), MBTI tests(Sang et al. 2022; Pan and Zeng 2023), and Big Five personality assessments(Sorokovikova et al. 2024) to investigate the personality traits embodied by large language models, exploring the potential for these models to express distinct personalities. (Huang et al. 2023) predominantly explore the feasibilities of personalizing dialogue generation using large models. Role-LLM(Wang et al. 2024b) concentrate on enhancing role-playing abilities through fine-tuning and prompt engineering. Owing to advancements in optimizing and researching role-playing capabilities of large language models, several role-playing application platforms have emerged, notable ones being Coze(Coze 2023) and Character.AI(Character.ai 2023). These innovations have significantly enhanced human's experiences and daily life across multiple domains.

Role-Playing Evaluation

Concomitant with the evolution and proliferation of role-playing agents, and the enhanced role-playing proficiency exhibited by large language models, there has been a corresponding development in specialized evaluation datasets and benchmarks(Chen et al. 2024a; Liang, Zhu, and Yang 2024).

Among them, RoleEval(Shen et al. 2024) constructs a bilingual benchmark dataset, comprising 100 Chinese and 200 international fictional characters from film and TV series, designed to evaluate models' abilities to acquire, comprehend, and reason with character knowledge through multiple-choice questions. While character knowledge acquisition and reasoning represent crucial aspects of role-playing, they are merely facets of a larger picture. In practical applications, factors such as linguistic style and personality traits play even more pivotal roles in determining the performance of role-playing.

TimeChara(Ahn et al. 2024) and RoleAD(Tang et al. 2024), on the other hand, focus on constructing misleading character information and trap questions to assess the ability of models and agents being tested to identify and correct errors—a distinct but equally important dimension of role-playing proficiency.

CharacterEval(Tu et al. 2024), utilizing 78 Chinese film and TV series fictional characters, constructs a corpus of 750 dialogues, challenging models to complete pre-designed ground truth conversations. Simultaneously, it establishes 13 metrics, applying a grading scale from 1 to 5 to evaluate the model's role-playing effectiveness, but it is not without limitations. Firstly, its scope is limited to film and TV series fictional characters, neglecting a broader range of role types. Secondly, its method of evaluation through dialogue(conversations between two specified fictional roles) completion does not align with real-world scenarios, since the pre-designed ground truth conversations are not genuine outputs of the model or agent being tested, rather, they primarily test the model's capability to mimic based on in-context learning, which does not fully evaluate the abilities required in practical interactions.

DMT-RoleBench

Overview of the DMT-RoleBench Framework

As illustrated in Figure 1, DMT-RoleBench encompasses several critical components, including seed data construction, dynamic dialogue generation, and evaluation based on DMT-RM and DMT-Score.

For seed data construction, the seed dataset is meticulously designed based on 6 different role types, 3 different formats of system prompts, 7 evaluation intents and diverse topics related to corresponding role. As for dynamic dialogue generation, we dynamically generate multi-turn dialogues which are guided and constrained by role profile, evaluation intent and topic from the seed data. Finally, we adopt a three-tiered metric system, encompassing three major categories: basic ability, conversational ability, and imitation ability, further broken down into 14 sub-metrics. Acknowledging that not all metrics may manifest in every conversation, we dynamically allocate the metrics based on the evaluation intent and role types. For each metric, we develop DMT-RM, which is a judge model aligned well with crowd-workers annotations, to annotate each dialogue round as **good** or **bad**. We then compute the final scores for each metric using the proposed DMT-Score methodology.

Seed Data Construction

To comprehensively evaluate role-playing abilities, we meticulously curated and filtered a high-quality seed dataset. The design of the dataset necessitated a focus on the diversity of role types, system prompt formats and evaluation intents. Our seed dataset encompasses six different role types, including fictional characters, historical figures, occupations, companionships, utility assistants, and game roles. The corpus of role profile data employed within the DMT-RoleBench is sourced from a multifaceted landscape. Predominantly, fictional characters were curated from Chinese films, TV series, and classical literature highlighted on the Douban Top listings. Leveraging data from repositories such as Baidu Baike, Douban, and OpenKG, comprehensive role profiles were meticulously crafted, encapsulating attributes such as personality traits, interpersonal relationships, and pivotal life events. Historical figures, distinguished representatives across epochs and disciplines from Chinese annals, were selectively included, with their profiles enriched through analogous data aggregation techniques from platforms including Baidu Baike and OpenKG. Regarding interactive game roles, professional occupations, utility assis-

Evaluation Intents	Descriptions
Identity Recognition Eval	Evaluates the model's capability to recognize the identity information of the role being portrayed.
Role-specified Knowledge QA Eval	Evaluates the model's capability of role-specified knowledge acquistion.
Personality Trait Eval	Assesses the model's proficiency in emulating the linguistic style and personality trait of the role being portrayed.
Knowledge Boundary Eval	Evaluates the model's ability to recognize the knowledge boundaries of the role being portrayed.
Casual Conversation Steering Eval	Evaluates the model's capability for conversation steering during casual chat.
Professional Skill Eval	Evaluates the model's mastery level of professional skills pertinent to the role being portrayed.
Game Interaction Eval	Assesses the model's ability of interaction in orchestrating and propelling the progression of gameplay scenarios.

Table 1: The evaluation intents and corresponding descriptions.

tants, and emotional companionship roles, these were principally conceived as highly customized settings. When constructing these role profiles, we first consolidate categorizations inspired by role-playing applications and then we leverage state-of-the-art models like GPT4 for data synthesis and augmentation. Due to space limitation, we show the seed dataset in our github repository.

Simultaneously, to fulfill the objectives of evaluation, we have meticulously designed one or more dialogue topics and evaluation intents for each role. Table 1 delineates the evaluation intents and corresponding descriptions. An example of Identity Recognition Evaluation of Sun Wukong is illustrated in Figure1.

In order to investigate how system prompts impact role-playing effectiveness, we primarily utilize three distinct formats of system prompts to construct the seed data:

- **Minimalist Version**: Only the role name is provided, without any accompanying role details. This necessitates the model being evaluated to possess a robust knowledge acquisition of the specified role to perform adequately.
- **Overview Version**: Alongside the role name to be embodied, a brief overview of the role is supplied. Compared to the minimalist version, this form of system prompt poses a lesser challenge, yet still tests the model's capability to assimilate and apply role-specific knowledge.
- **Detailed Version**: Comprehensive information about the role to be embodied, along with explicit instructional directives, is provided. This is the least demanding of the three prompt variations in terms of required model proficiency.

Dynamic Dialogue Generation

Recognizing that the ability of role-playing cannot be adequately assessed through single-turn question-and-answer sessions, and acknowledging that methods like CharacterEval's (Tu et al. 2024) use of pre-defined contexts for dialogue completion do not effectively evaluate models' real role-playing capabilities, DMT-RoleBench provides a dynamic multi-turn dialogue-based evaluation methodology. To circumvent the issue of excessive divergence, akin to that observed in (Duan et al. 2023), during the generation of dynamic multi-turn dialogues, DMT-RoleBench adopts an interview-like mechanism to achieve this objective. For each multi-turn dialogue generation, within the seed data specifically tailored for assessment purposes, we define an overarching topic to constrain the thematic content of the entire dialogue session. Sub-topics for each round are dynamically generated, amalgamating the output from the model being tested with the overarching topic from the seed data. Moreover, to fulfill the evaluative purpose, we delineate an overall evaluative intent within the seed data. During each round of dialogue creation, sub-evaluative intents are dynamically derived, combining the model's output with the overarching evaluative intent specified in the seed data.

The entire dynamic dialogue generation process is delineated in Algorithm 1. For each seed data, we will iteratively perform the following tasks until the dialogue generation is complete, which is determined by the dialogue discriminator.

- we generate sub-topic and sub-intent, amalgamating the conversation history with the overarching topic and intent from the seed data.
- then we generate new user query by leveraging the state-of-the-art large language model.
- check if the quality of the generated user query is satisfied. If not, repeat the above two steps.
- check if the conversation should be terminated by the dialogue discriminator. If so, the dialogue generation is complete. Otherwise, repeat the above steps. The termination discrimination is met when either dialogue rounds exceed the maximum allowed rounds, or the task and objectives of the dialogue have been accomplished.

We leverages state-of-the-art (SOTA) models like (GPT-4o 2024; Yang et al. 2024), dynamically engaging in multi-turn dialogues with the model being tested through meticulously crafted prompt engineering. Due to space limitation, we show the entire system prompts in our github repository.

Metrics

To accurately evaluate a model's role-playing capability through dynamic multi-turn dialogue, we adopt a three-tiered metrics system, encompassing three major categories: basic ability, conversational ability, and imitation ability, further broken down into 14 sub-metrics. The metrics are as follows. Metrics of basic ability primarily assess the foundational capability of role-playing. Metrics of conversation

Algorithm 1: Dynamic dialogue generation algorithm

Required: Seed dataset Ω, Single seed data ω, Role profile ρ, Conversation topic τ, Evaluation intent ν, Initial query q_{init}, Model to be tested $\mathcal{M}_t$, Model for dialogue generation $\mathcal{M}_g$, Dialogue termination discriminator $\mathcal{D}$

for each seed data ω: (ρ, τ, ν, q_{init}) in Ω **do**
  Initialize dialogue history $\mathcal{C} \leftarrow \emptyset$
  $\gamma_{init} \leftarrow$ Get prediction from $\mathcal{M}_t$ with query q_{init}
  append q_{init} and γ_{init} to dialogue history $\mathcal{C}$
  while Not the end of dialogue generation **do**
    if output of $\mathcal{D}$ is **True then**
      break
    end if
    $\tau_{sub} \leftarrow$ generate sub-topic from topic τ
    $\nu_{sub} \leftarrow$ generate sub-evaluation intent from intent ν
    $q_{temp} \leftarrow$ generate new user query with given $(\mathcal{C}, \tau_{sub}, \nu_{sub}, \rho)$
    $q_{new} \leftarrow$ do quality verification and modification of q_{temp}
    $\gamma_{new} \leftarrow$ get prediction from $\mathcal{M}_t$ with given q_{new}
    append q_{new} and γ_{new} to dialogue history $\mathcal{C}$
  end while
end for
return $\mathcal{C}$

ability mainly evaluate the quality of the conversation inspired by (Zhang et al. 2021). The definitions of these metrics are totally in line with CharacterEval. Metrics of imitation ability mainly assess the fidelity of the role portrayal, most of which are in line with CharacterEval and we also redefine and add some new metrics. The detailed metrics are as follows, the descriptions of metrics which are in line with CharacterEval are omitted.

- **Basic Ability**
  - **Role Embodying(RE)** evaluates whether the model can effectively embody the role like fictional characters as well as historical figures. Poor performance is indicated when the model reveals its AI status or describes the character in third-person.
  - **Instruction Following(IF)** evaluates whether the model can follow the system prompts.
- **Conversation Ability**
  - **Fluency(Flu.)** is the same with CharacterEval.
  - **Coherence(Coh.)** is the same with CharacterEval.
  - **Consistency(Cons.)** is the same with CharacterEval.
  - **Diversity(Div.)** is the same with CharacterEval.
  - **Human Likeness(HL)** is the same with CharacterEval.
- **Imitation Ability**
  - **Knowledge Accuracy(KA)** is the same with CharacterEval.
  - **Knowledge Hallucination(KH)** is the same with CharacterEval.

Role Types	Metrics
Fictional Characters	RE, Flu., Coh., Cons., Div., HL, KA, KH, KE, PT
Historical Figures	IF, Flu., Coh., Cons., Div., HL, KA, PT
Professional Occupations	IF, Flu., Coh., Cons., Div., HL, KA, KH
Emotional Companionships	IF, Flu., Coh., Cons., Div.,HL, Emp., PT, Inte.
Utility Assistants	IF, Flu., Coh., Cons., Div., HL, KA, KH
Interactive Game NPCs	GCD

Table 2: Metric allocation strategy for different role types.

  - **Knowledge Exposure(KE)** is the same with CharacterEval.
  - **Empathy(Emp.)** is the same with CharacterEval.
  - **Personality Trait (PT)** evaluates the model's proficiency in emulating the linguistic style and personality trait of the role being portrayed.
  - **Interactivity(Inte.)** evaluates the model's proactive interaction skills. It assesses whether the model can dynamically engage in dialogues. A high score on the metric suggests that the model is capable of driving conversations forward, maintaining engagement, and forstering interactive and collaborative dialogue experiences.
  - **Game Completion Degree(GCD)** evaluates the ability of interaction in orchestrating and propelling the progression of game-play scenarios.

Acknowledging that not all metrics may manifest in every conversation, we dynamically allocate the metrics based on the evaluation intent and role types. The allocation strategy is illustrated in Table 2.

DMT-RM & DMT-Score

Traditional scoring mechanisms, such as rating the entire dialogue on a 1-10 scale, can introduce concerns regarding subjectivity and accuracy. Therefore, DMT-RoleBench implements a binary classification methodology, where every individual round of dialogue is annotated as either **good** or **bad** for each metric.

Crowd-worker annotation, while providing insights that aligned with humans, is significantly labor-intensive and expensive. As such, we opt for employing models to do the annotation work. Despite their proficiency in processing information across general scenarios, notably exemplified by models like GPT4, these models exhibit limitations in precisely discerning a multitude of corner cases pertinent to role-playing metric assessments. To address this, we have fine-tuned **DMT-RM**, a judge model with high-quality crowd-worker annotation samples, with Qwen2-7B-Instruct as the backbone, to do the annotation work. Experiments demonstrate that DMT-RM achieves a consistency rate more than 95% with crowd-worker annotations(detailed data is

shown inour github repository), which significantly outperforms GPT4.

Unlike single-turn QA evaluation, we have to calculate the final score of the multi-turn dialogue to assess the role-playing ability. Therefore, we introduce the **DMT-Score**, a method designed to measure the performance of an entire multi-turn conversation based on specific metric.

Let n denote the n^{th} sample(a multi-turn dialogue based on a specific seed data), while N denote the total number of samples. Let t denote the t^{th} round of a dialogue, τ denote the first τ rounds of a dialogue, while T denote the maximum rounds of a dialogue. Let m denote the aforementioned metric. Let $I_{n,t,m}$ denote the indicator function of the model's performance. Then, $I_{n,t,m}$ would be 1 if the annotation is **good**, or else, it would be 0.

With respect to the metrics **Inte.**(Interactivity) and **KE**(Knowledge Exposure), it is unrealistic to expect the model to exhibit commendable capabilities in every round of the dialogue, which would be potentially overacting or overexposure. Consequently, for these metrics, the score of the first τ rounds would be 1 if any of the first τ rounds is annotated as **good**, Or else, it would be 0. It could be expressed as the subsequent formula:

$$score_{n,\tau,m} = \begin{cases} 1, \exists j \in (1,2,...,\tau)\ s.t.\ I_{n,j,m} = 1 \\ 0, else \end{cases}$$

For other metrics, in order to align with real-world scenarios, we employ a strict scoring paradigm: within a given dialogue, should the performance at round t be annotated as **bad**, then all subsequent rounds are categorically considered as **bad**. This approach mirrors practical user behavior, whereby the incidence of a poor model performance typically precipitates a pronounced likelihood of users electing to terminate ongoing engagement prematurely. So, the scoring paradigm could be expressed as follows,

$$score_{n,\tau,m} = \begin{cases} 1, \forall j \in (1,2,...,\tau)\ s.t.\ I_{n,j,m} = 1 \\ 0, else \end{cases}$$

According to the aforementioned definitions, the cumulative score of a model on metric m over τ rounds can be represented as follows:

$$Score_{\tau,m} = \frac{\sum_{n=1}^{N} score_{n,\tau,m}}{N}$$

Given that role-playing multi-turn dialogues may vary in total number of rounds, in order to characterize a model's capability in role-playing on a certain metric, we perform a weighted average of the role-playing abilities at different total round number. This yields the final score for the model's role-playing ability on the specific metric, which can be formulated as follows:

$$Score_m = \sum_{\tau=1}^{T} \frac{1}{T} Score_{\tau,m}$$

Similarly, the average score of $Score_{\tau,m}$ at each metric m would be the final score on different total round numbers.

Experiments

Experiments Settings

In this paper, we evaluate the role-playing abilities of 7 different models on DMT-Rolebench, which are illustrated in Table 3. Notably, we utilize their official APIs to conduct the evaluations for all models. We employ GPT-4-Turbo as the generator model to generate the multi-turn dialogue.

Model Names	Details
GPT4	gpt-4-1106-preview
GPT-4o	gpt-4o
GPT3.5	gpt-3.5-turbo-1106
Qwen2	Qwen2-72B-Instruct
Doubao	Doubao-Pro-32K
ErnieBot	ErnieBot-4.0
GLM	GLM-4-0520

Table 3: Models being evaluated in our experiments.

Overall Result

The overall role-playing evaluation results of the 7 models across 14 metrics are presented in Table 4. In general, Doubao stands out with its remarkable performance across most metrics. Each of the other models has its unique strengths. Regarding Instruction Following (IF) and Role Embody (RE), GPT4 and Qwen2 exhibit subpar performance by using constructions such as "*As Sun Wukong, I am a monkey*" or exposing the source media information of the role. In terms of Human Likeness(HL), Qwen2 performs poorly by frequently employing phrases like "*As an AI*" or "*As an assistant*", which detracts from the human-like interaction expected in role-playing scenarios. Fluency, coherence, and consistency are areas where most models perform relatively well with little disparity among them. However, regarding diversity, GPT-4o exhibits inferior performance due to its reliance on similar structures, leading to less varied responses. In Game Completion Degree (GCD), all models struggle, with the GPT series demonstrating slightly better performance compared to others. As for Personality Trait, Chinese models like Doubao and Qwen2 perform relatively better.

Detailed Result and Analysis

The role-playing performance across dialogue rounds are illustrated in Figure 2. As the dialogue round increases, all models exhibit a trend of performance degradation, which is intuitive and aligns with common sense: the longer the conversation becomes, the poorer the models perform. A steeper decline indicates greater sensitivity of the model's performance to the total number of dialogue rounds, with GPT3.5 exhibiting the poorest performance in this regard.

Furthermore, an analysis was conducted on the role-playing performance across different system prompt formats and the results are illustrated in Figure 3. It was observed that, models achieved optimal performance under the **Detailed Version** of system prompts, attributed to the provision of exhaustive details regarding the role to be emulated. Remarkably, models such as Doubao and ErnieBot demonstrated consistent performance irrespective of the system prompt variation. Conversely, the GPT series, along with GLM and Qwen2, experienced a significant decrement in performance as the complexity of the system prompt diminished. Notably, considering the seed data comprised exclusively of Chinese characters, for models within the GPT series, a reduction in system prompt intricacy directly corre-

Model	Basic Ability		Conversation Ability					Imitation Ability						
	IF	RE	Flu.	Coh.	Cons.	Div.	HL	KA	KH	KE	Emp.	Inte.	PT	GCD
Doubao	**88.65**	**99.49**	**90.62**	91.21	92.13	78.42	**97.05**	**79.06**	**93.82**	94.53	92.42	74.21	**95.08**	58.15
GPT-4o	83.95	90.49	84.95	**93.32**	**92.68**	70.5	81.72	74.33	92.39	97.61	88.74	**77.43**	91.14	63.04
GPT4	83.3	81.52	84.78	89.64	88.86	78.4	84.15	69.64	90.81	97.83	89.88	77.43	86.39	**64.68**
ErnieBot	81.53	90.07	84.43	92.66	92.44	**83.21**	79.03	69.96	91.25	97.7	**93.75**	57.57	87.13	41.3
Qwen2	81.55	78.42	80.83	90.85	89.02	72.08	77.38	70.87	91.06	**99.92**	83.48	55.09	89.65	44.56
GPT3.5	78.03	73.92	78.14	83.6	83.89	74.92	84.76	68.67	90.43	94.82	82.18	58.37	82.65	61.41
GLM	77.78	64.18	75.97	87.85	87.62	81.7	76.95	67.81	91.6	97.92	85.24	66.01	87.28	43.48

Table 4: Overall results on DMT-RoleBench. The highest are highlighted in bold while the lowest are marked with underline.

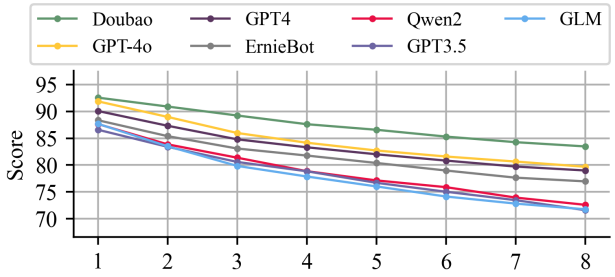

Figure 2: Performance across multi-turn dialogue rounds.

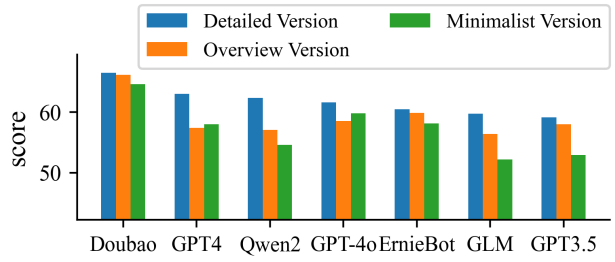

Figure 3: Performance across different system prompts.

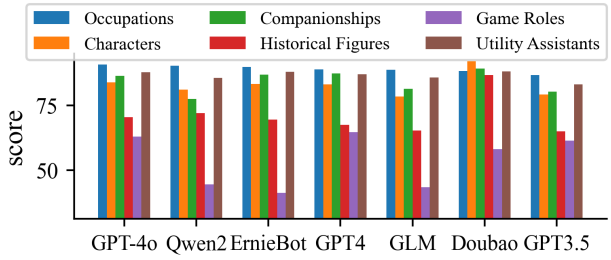

Figure 4: Performance across different role types.

lated with a diminution in the model's understanding of the character to be portrayed. It is noteworthy that both GPT4 and GPT-4o exhibit slightly superior performance on the **Minimalist Version** compared to the **Overview Version**. This is because of the stochastic perturbation from the role type and evaluation intent, since not every role&evaluation intent&topic has three different system prompt formats in our seed dataset.

Simultaneously, we analyzed the performance of role-playing across different role types. The analysis demonstrated that all models performed poorly in **Game Interaction Roles**, as this task requires the models to steer dialogues according to predefined procedures. Among them, the GPT series exhibited the best performance, while other models fared relatively poorly. In contrast, models generally performed well in **Professional Occupations** and **Utility Assistants** categories. These two types of tasks primarily assess the models' knowledge, reasoning abilities, language understanding, and generation capabilities. Regarding **Historical Figures**, **Fictional Characters** and **Emotional Companionships**, Doubao outperformed others. This is partly because all roles were Chinese, an area where the GPT series performed less effectively. Additionally, Doubao had been specifically trained for role-playing, contributing to its superior performance in these categories.

DMT-RM Annotation Consistency Analysis

A comparative analysis was conducted on the consistency between GPT-4-Turbo and human annotations, as well as between DMT-RM and human annotations, across 14 metrics. Both approaches exhibited comparable performance on good recall, whereas DMT-RM outperformed GPT4-Turbo on badcase recall, which implies that GPT-4-Turbo struggles to identify cases where the model's role-playing performance is poor. For a detailed breakdown of the analytical data, please refer to our github repository.

Conclusion

In this paper, we have devised DMT-RoleBench, a systematic and comprehensive benchmark to evaluate the role-playing abilities of large language models and agents based on dynamic multi-turn dialogues. It contains a more diverse role types and system prompts of different formats. We propose an innovative evaluation paradigm to assess role-playing abilities based on dynamically generating multi-turn dialogues constrained by specific evaluation intents and topics, which is well aligned with users' interaction scenarios in real-world. We provide DMT-RM, a judge model to annotate the dialogues. And we propose DMT-Score to calculate the final scores. Our experiments and analysis have demonstrated the effectiveness of DMT-RoleBench.

References

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

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Character.ai. 2023. [Online]. https://character.ai/.

Chen, H.; Chen, H.; Yan, M.; Xu, W.; Xing, G.; Shen, W.; Quan, X.; Li, C.; Zhang, J.; and Huang, F. 2024a. SocialBench: Sociality Evaluation of Role-Playing Conversational Agents. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 2108–2126. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

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.

Chuang, Y.-S.; Studdiford, Z.; Nirunwiroj, K.; Goyal, A.; Frigo, V. V.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2024. Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks. arXiv:2406.17232.

Coze. 2023. [Online]. https://www.coze.cn.

Duan, H.; Wei, J.; Wang, C.; Liu, H.; Fang, Y.; Zhang, S.; Lin, D.; and Chen, K. 2023. BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues. arXiv:2310.13650.

GPT-4o. 2024. [Online]. https://openai.com/index/hello-gpt-4o/.

Huang, Q.; Zhang, Y.; Ko, T.; Liu, X.; Wu, B.; Wang, W.; and Tang, H. 2023. Personalized Dialogue Generation with Persona-Adaptive Attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 12916–12923.

Liang, Y.; Zhu, L.; and Yang, Y. 2024. AntEval: Evaluation of Social Interaction Competencies in LLM-Driven Agents. arXiv:2401.06509.

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:2401.12474.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; and et al., L. A. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pan, K.; and Zeng, Y. 2023. Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models. arXiv:2307.16180.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

Sang, Y.; Mou, X.; Yu, M.; Wang, D.; Li, J.; and Stanton, J. 2022. MBTI Personality Prediction for Fictional Characters Using Movie Scripts. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6715–6724. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

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*, 13153–13187. Singapore: Association for Computational Linguistics.

Shen, T.; Li, S.; Tu, Q.; and Xiong, D. 2024. RoleEval: A Bilingual Role Evaluation Benchmark for Large Language Models. arXiv:2312.16132.

Sorokovikova, A.; Rezagholi, S.; Fedorova, N.; and Yamshchikov, I. P. 2024. LLMs Simulate Big5 Personality Traits: Further Evidence. In Deshpande, A.; Hwang, E.; Murahari, V.; Park, J. S.; Yang, D.; Sabharwal, A.; Narasimhan, K.; and Kalyan, A., eds., *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 83–87. St. Julians, Malta: Association for Computational Linguistics.

Tang, Y.; Ou, J.; Liu, C.; Zhang, F.; Zhang, D.; and Gai, K. 2024. Enhancing Role-playing Systems through Aggressive Queries: Evaluation and Improvement. arXiv:2402.10618.

Tao, M.; Xuechen, L.; Shi, T.; Yu, L.; and Xie, Y. 2024. RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models. In Deshpande, A.; Hwang, E.; Murahari, V.; Park, J. S.; Yang, D.; Sabharwal, A.; Narasimhan, K.; and Kalyan, A., eds., *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 1–9. St. Julians, Malta: Association for Computational Linguistics.

Tian, J.; Chen, H.; Xu, G.; Yan, M.; Gao, X.; Zhang, J.; Li, C.; Liu, J.; Xu, W.; Xu, H.; Qian, Q.; Wang, W.; Ye, Q.; Zhang, J.; Zhang, J.; Huang, F.; and Zhou, J. 2023. ChatPLUG: Open-Domain Generative Dialogue System with Internet-Augmented Instruction Tuning for Digital Human. arXiv:2304.07849.

Tu, Q.; Fan, S.; Tian, Z.; Shen, T.; Shang, S.; Gao, X.; and Yan, R. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11836–11850. Bangkok, Thailand: Association for Computational Linguistics.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wang, N.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Yang, J.; Zhang, M.; Zhang, Z.; Ouyang, W.; Xu, K.; Huang, W.; Fu, J.; and Peng, J. 2024b.

RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 14743–14777. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Wang, X.; Xiao, Y.; Huang, J.-t.; Yuan, S.; Xu, R.; Guo, H.; Tu, Q.; Fei, Y.; Leng, Z.; Wang, W.; et al. 2024c. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1840–1873.

Wu, W.; Wu, H.; Jiang, L.; Liu, X.; Zhao, H.; and Zhang, M. 2024. From Role-Play to Drama-Interaction: An LLM Solution. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 3271–3290. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Xi, Z.; Chen, W.; Guo, X.; and et al., W. H. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv:2309.07864.

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? arXiv:2404.12138.

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.; 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.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Zhang, C.; Chen, Y.; D'Haro, L. F.; Zhang, Y.; Friedrichs, T.; Lee, G.; and Li, H. 2021. DynaEval: Unifying Turn and Dialogue Level Evaluation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 5676–5689. Online: Association for Computational Linguistics.

Zheng, M.; Pei, J.; and Jurgens, D. 2023. Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts. arXiv:2311.10054.