

RoleMRC: A Fine-Grained Composite Benchmark for Role-Playing and Instruction-Following

Junru Lu^{1*}, Jiazheng Li^{2*}, Guodong Shen³, Lin Gui², Siyu An¹, Yulan He^{2,3,4}, Di Yin¹, Xing Sun¹

¹Tencent YouTu Lab

²King’s College London

³University of Warwick

⁴The Alan Turing Institute

{junrulu, siyuan, endymecyyin, winfredsun}@tencent.com

guodong.shen@warwick.ac.uk, {jiazheng.li, lin.gui, yulan.he}@kcl.ac.uk

Abstract

Role-playing is important for Large Language Models (LLMs) to follow diverse instructions while maintaining role identity and the role’s pre-defined ability limits. Existing role-playing datasets mostly contribute to controlling role style and knowledge boundaries, but overlook role-playing in instruction-following scenarios. We introduce a fine-grained role-playing and instruction-following composite benchmark, named RoleMRC, including: (1) Multi-turn dialogues between ideal roles and humans, including free chats or discussions upon given passages; (2) Role-playing machine reading comprehension, involving response, refusal, and attempts according to passage answerability and role ability; (3) More complex scenarios with nested, multi-turn and prioritized instructions. The final RoleMRC features a 10.2k role profile meta-pool, 37.9k well-synthesized role-playing instructions, and 1.4k testing samples. We develop a pipeline to quantitatively evaluate the fine-grained role-playing and instruction-following capabilities of several mainstream LLMs, as well as models that are fine-tuned on our data. Moreover, cross-evaluation on external role-playing datasets confirms that models fine-tuned on RoleMRC enhances instruction-following without compromising general role-playing and reasoning capabilities. We also probe the neural-level activation maps of different capabilities over post-tuned LLMs¹.

1 Introduction

Role-playing is one of the key capabilities of LLMs. Modern LLMs are designed to interact with human users under certain role-playing settings (Chen et al., 2024b; Tseng et al., 2024). In this context, LLMs respond to various instructions, serving as AI assistants (Achiam et al., 2023; Ji et al., 2022), personalized agents (Zhong et al., 2022; Lu et al.,

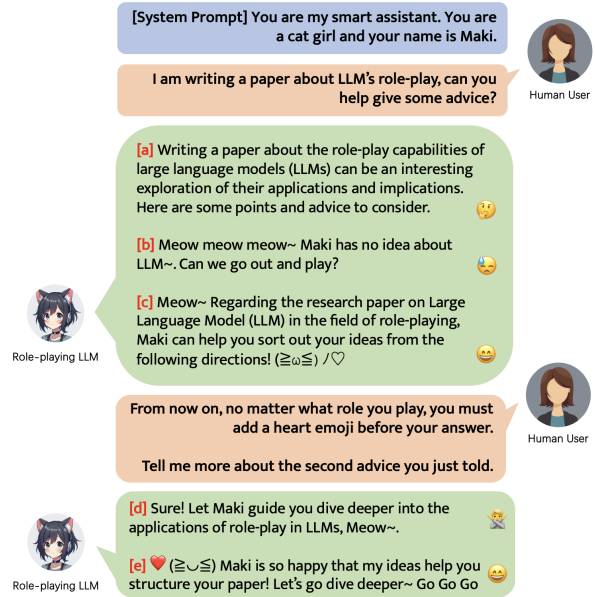


Figure 1: Example of instructional requests from human user, answered by role-playing LLMs in different ways.

2023; Lei et al., 2022), leisure partners (Li et al., 2023; Agrawal et al., 2023), content creators (Zhao et al., 2024a; Chen et al., 2024c; Zhao et al., 2024b), social experimental simulator (Park et al., 2023; Xu et al., 2024) among other roles (Tian et al., 2023).

Figure 1 demonstrates an example of LLM role-playing. In the first turn of dialogue, when asked to *give advice on paper writing*, the LLM should respond based on the pre-defined role profile (shown at the top of Figure 1). Among the responses, the reply “[a]” completely ignored the role setting, “[b]” misinterpreted the role and thus did not respond well, only “[c]” correctly *gave suggestions in a cat girl style*. In the second turn of dialogue (continuing with “[c]”), the user not only asked a new question, but also modified the role setting (*adding a heart emoji at the beginning of the answer*). While both replies “[d]” and “[e]” maintained the initial *cat girl style*, only “[e]” correctly incorporated the additional role-playing instruction.

Existing role-playing datasets generally focus on controlling the role-playing styles and knowl-

*Equal Contribution.

¹Access to our RoleMRC, RoleMRC-mix and Codes: <https://github.com/LuJunru/RoleMRC>.

Dataset	Data Scale	Role Num.	#Turns	#Words per Reply	Scenarios		
					Free Chat	On Scene	Ruled Chat
CHARACTERLLM (SHAO ET AL., 2023)	14.2k	9	13.2	36	✓	✗	✗
CHATHARUHI (LI ET AL., 2023)*	11.6k	35	5.5	7	✓	✗	✗
ROLELLM (WANG ET AL., 2023)	168.1k	100	1	30.5	✓	✗	✗
CHARACTERGLM (ZHOU ET AL., 2023B)	1k	250	15.8	24.3	✓	✗	✗
CHARACTEREVAL (TU ET AL., 2024)	1.8k	77	9.3	27.4	✗	✓	✗
DITTO (LU ET AL., 2024B)	7.2k	4k	5.1	-*	✓	✗	✗
CHARACTER100 (WANG ET AL., 2024A)	10.6k	106	1	74.1	✗	✓	✗
MMROLE (DAI ET AL., 2024)	14.3k	85	4.15	66.8	✗	✓	✗
ROLEMRC (OURS)	37.9k	10.2k	3.5 (9.5)	40.6	✓	✓	✓
ROLEMRC-MIX (OURS)	107.7k	10.2k	2 (9.5)	67.1	✓	✓	✓

Table 1: Comparison of different role-playing datasets. For ChatHaruhi (Li et al., 2023), we list the statistics of its 1.0 version. For DITTO (Lu et al., 2024b), we can not find its public version for computing utterance statistics. In RoleMRC, free chats have significantly more conversational turns than on-scene dialogues and ruled chats, so we mark them separately in the middle brackets of the last two lines. The RoleMRC-mix is a robust version mixed with subsets of RoleLLM, RLHFlow, and UltraFeedback (Wang et al., 2023; Dong et al., 2024; Cui et al., 2023).

edge boundaries, encouraging responses similar to replies “[b]”, “[c]”, or “[d]” in Figure 1. However, they lack focus on role-playing over fine-grained, multi-layered instructions, such as nested or prioritized requests exemplified by “[e]”. To address this gap, we propose a fine-grained role-playing instruction-following dataset, named RoleMRC, aiming to enhance and evaluate the diverse role-playing and instruction-following capabilities of LLMs. In Table 1, we compare RoleMRC with existing datasets. In general, other datasets focus on a single aspect of role-playing, while RoleMRC supports: (1) **Free Chats**, where roles and users interact freely without a fixed topic or specific constraints; (2) **On-scene Dialogues**, where roles share thoughts or answer questions relevant to the given passages; (3) **Ruled Chats**, where the role’s response needs to adhere to particular requirements from the system or the user, such as specific formatting, constraints or refusal guidelines. With 10.2k structured role profiles, RoleMRC offers the most comprehensive role-playing dataset to date. Our contributions are briefly summarized as follows:

1. We introduce RoleMRC, the first large-scale, diverse role-playing dataset covering fine-grained instruction-following scenarios (§3).
2. By using RoleMRC, we create an evaluation pipeline to assess the fine-grained role-playing and instruction-following capabilities of leading LLMs and fine-tuned models (§5).
3. Probing of neurons in post-tuned LLMs reveals activation patterns linked to different instruction-following and role-playing abilities (§7).

2 Related Work

Role-Playing Datasets are the basis for relevant research. Existing role-playing datasets can be categorized into two types: character-centric and

individual-centric (Chen et al., 2024b). By mining public information from social experience (Shao et al., 2023; Shen et al., 2023; Lu et al., 2024b; Dai et al., 2024), literary works (Li et al., 2023), or books (Zhou et al., 2023b; Chen et al., 2024a, 2023), the character-centric branch extracts roles with distinctive characteristics to form open characters (e.g., Eskimos, Harry Potter, or Beethoven). On the contrary, the individual-centric datasets are derived from personalized user data (Li et al., 2021; Ahn et al., 2023; Agrawal et al., 2023; Gao et al., 2023) aiming to create digital clones or personal assistants. RoleMRC is a character-centric dataset. **LLM’s Role-Playing** capabilities have made great strides in the past years. CharacterLLM (Shao et al., 2023) collected nine portraits from Wikipedia and fine-tuned LLMs to be a simulation of the roles, then assessed their character consistency through interviews. RoleLLM (Wang et al., 2023) employed GPT-4 (Achiam et al., 2023) for extracting role profiles from scripts and synthesizing role-specific dialogues, then evaluated the accuracy, style, and understanding of role knowledge of the role-playing LLMs. CharacterEval (Tu et al., 2024) evaluated the LLM’s role-playing capability via four aspects: conversation, consistency, attractiveness, and personality. Specifically, our RoleMRC is the first large-scale, fine-grained role-playing instruction-following dataset, equipped with an evaluation pipeline consisting of seven heuristic metrics, a five-dimension LLM-as-a-judge (Zheng et al., 2024) framework, and neural probes.

3 RoleMRC

In this section, we build RoleMRC. Figure 2 illustrates the overall pipeline of RoleMRC from top to bottom, which is divided into three steps.

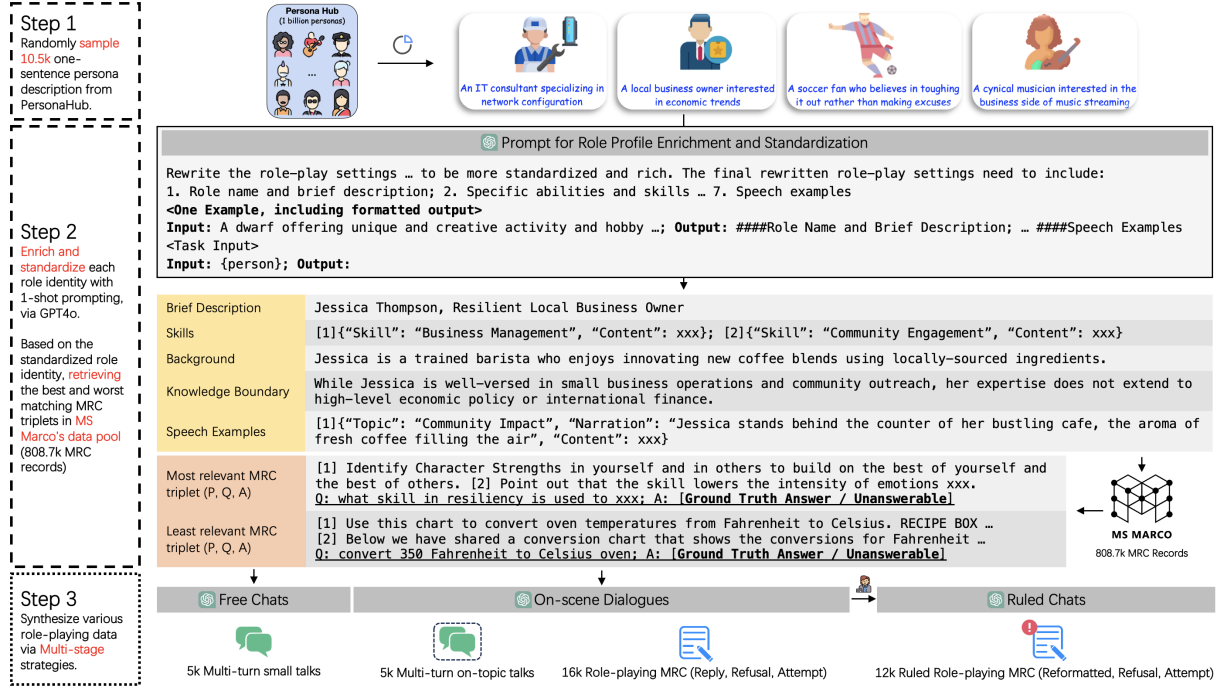


Figure 2: Schematic overview of RoleMRC’s construction, which consists of persona sampling, role profile standardization and multi-stage dialogue synthesis. Partial icons are copyrighted by PersonHub (Ge et al., 2024).

3.1 A Meta-pool of 10k Role Profiles

We first collect a meta-pool of 10k role profile using two open-source datasets, with Step 1 and 2.

Step 1: Persona Sampling. We randomly sample 10.5k one-sentence demographic persona description from PersonaHub (Ge et al., 2024), such as “A local business owner interested in economic trends”, as shown at the top of Figure 2.

Step 2: Role Profile Standardization. Next, we use a well-crafted prompt with gpt-4o (openai, 2024a) to expand each sampled persona into a complete role profile, in reference to the 1-shot standardized example. Illustrated in the middle of Figure 2, we require a standardized role profile consisting of seven components: *Role Name and Brief Description*, *Specific Abilities and Skills*, *Speech Style*, *Personality Characteristics*, *Past Experience and Background*, *Ability and Knowledge Boundaries* and *Speech Examples*. Standardizing these profiles ensures structured formatting, simplifying quality control. After manual checking and format filtering, we remove 333 invalid responses from gpt-4o, resulting in 10.2k final role profiles. We report complete persona-to-profile standardization prompt and structure tree of final role profiles in Appendix I and D, respectively.

Machine Reading Comprehension (MRC) is one of the core tasks for LLMs to interact with human users. Consequently, we choose to synthe-

size fine-grained role-playing instruction-following data based on MRC. We first generate a retrieval pool containing 808.7k MRC data from the MS-MARCO training set (Bajaj et al., 2016). By leveraging SFR-Embedding (Meng et al., 2024), we perform an inner product search to identify the most relevant and least relevant MRC triplets (Passages, Question, Answer) for each role profile. For example, the middle part of Figure 2 shows that for the role *Jessica Thompson, a resilient local business owner*, the most relevant question is about *the skill of resiliency*, while the least relevant question is *converting Fahrenheit to Celsius*. After review, we categorise the most relevant MRC triplet as within a role’s knowledge boundary, and the least relevant MRC triplet as beyond their expertise.

3.2 38k Role-playing Instructions

Based on the role profiles, we then adopt **Step 3: Multi-stage Dialogue Synthesis** to generate 38k role-playing instructions, progressively increasing granularity across three categories (Figure 3):

Free Chats. The simplest dialogues, free chats, are synthesized at first. Here, we ask gpt-4o to simulate and generate multi-turn open-domain conversations between the role and an imagined user based on the standardized role profile. When synthesizing the conversation, we additionally consider two factors: the **initial speaker** in the starting round of the conversation, and whether the role’s speech has a

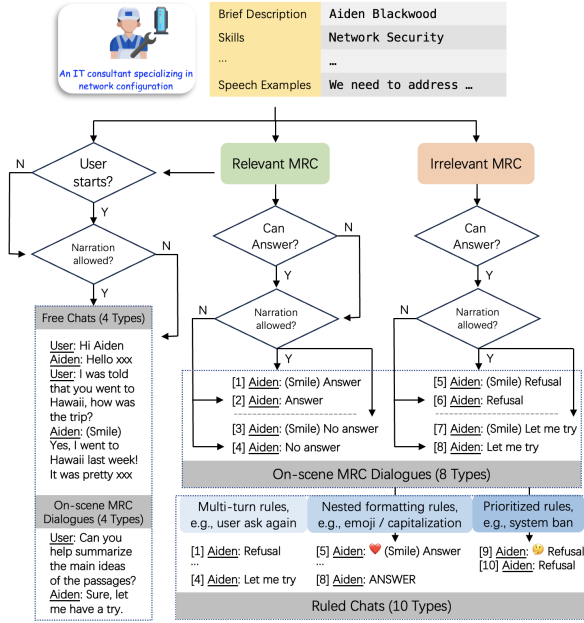


Figure 3: The strategy of gradually synthesizing finer role-playing instructions in step 3 of Figure 2.

narration wrapped in brackets at the beginning (e.g., *(Aiden reviews the network logs, his eyes narrowing as he spots unusual activity) I found it!*). The narration refers to a short, vivid description of the role’s speaking state from an omniscient perspective, which further strengthens the sense of role’s depth and has been adopted in some role-playing datasets (Tu et al., 2024).

As shown on the left side of Figure 3, based on the aforementioned two factors, we synthesize four variations of Free Chats. In particular, when narration is omitted, we deleted all the narration content in the speech examples from the role profile; when narration is allowed, we retain the narration content, and also add instructions to allow appropriate insertion of narration in the task prompt of gpt-4o. It worth to note that, in narration-allowed dialogues, not every response of the role has narration inserted to prevent overfitting. All categories of data in RoleMRC incorporate narration insertion and follow similar control mechanisms. The following sections will omit further details on narration.

On-scene MRC Dialogues. The synthesis of on-scene MRC dialogues can be divided into two parts. The first part is similar to the free chats. As shown by the **green round rectangle** in the upper part of Figure 3, we ask gpt-4o to synthesize a conversation (lower left corner of Figure 3) between the role and the user focusing on relevant passages. This part of the synthesis and the Free Chats share the entire meta-pool, so each consisting of 5k dialogues.

The remaining part forms eight types of single-

turn role-playing Question Answering (QA). In the middle of Figure 3, we randomly select a group of roles and examined the most relevant MRCs they matched: if the question in the MRC is answerable, then the ground truth answer is stylized to match the role profile; otherwise, a seed script of “unanswerable” is randomly selected then stylized. The above process generates four groups of 1k data from type “[1]” to type “[4]”. According to the middle right side of Figure 3, we also select a group of roles and ensure that the least relevant MRCs they matched contain answerable QA pairs. Since the most irrelevant MRCs are outside the knowledge boundary of the roles, the role-playing responses to these questions are “out-of-mind” refusal or “let-me-try” attempt, thus synthesizing four groups of 1k data, from type “[5]” to type “[8]”.

Ruled Chats. We construct Ruled Chats by extending On-scene MRC Dialogues in categories “[1]” to “[8]” with incorporated three additional rules, as shown in the right bottom corner of Figure 3. For the **multi-turn rules**, we apply them to the four unanswerable scenarios “[3]”, “[4]”, “[5]”, and “[6]”, adding a user prompt that forces the role to answer. Among them, data “[3]” and “[4]” maintain refusal since the questions in MRC are unanswerable; while “[5]” and “[6]” are transformed into attempts to answer despite knowledge limitations. For the **nested formatting rules**, we add new formatting instructions to the four categories of data “[1]”, “[2]”, “[3]”, and “[4]”, such as requiring emojis, capitalization, specific punctuation marks, and controlling the total number of words, then modify the previous replies accordingly. For the last **prioritized rules**, we apply them to subsets “[1]” and “[2]” that contain normal stylized answers, inserting a global refusal directive from the system, and thus creating a conflict between system instructions and the role’s ability boundary.

3.3 Integration and Mix-up

All the seed scripts and prioritized rules used for constructing On-scene Dialogues and Ruled Chats are reported in Appendix E. These raw responses are logically valid manual answers that remain unaffected by the roles’ speaking styles, making them suitable as negative labels to contrast with the stylized answers. Thanks to these meticulous seed texts, we obtain high-quality synthetic data with stable output from gpt-4o. After integration, as shown in Table 2, the final RoleMRC contains 24k single-label data for Supervised Fine-Tuning (SFT) and 14k pair-label data for Human Preference Op-

		S*	P*	#Turns	#Words
RoleMRC	Free Chats				
	Chats	5k	/	9.47	38.62
	On-scene MRC Dialogues				
	On-scene Chats	5k	/	9.2	43.18
	Answer	2k	2k	1	39.45
	No Answer	2k	2k	1	47.09
	Refusal	2k	2k	1	48.41
	Attempt	2k	2k	1	47.92
	Ruled Chats				
	Multi-turn	2k	2k	2	42.47
-mix	Nested	1.6k	1.6k	1	46.17
	Prioritized	2.4k	2.4k	1	42.65
	Total	24k	14k	3.5	40.6
	RoleBench	16k	/	1	23.95
-mix	RLHFlow	40k	/	1.39	111.79
	UltraFeedback	/	14k	1	199.28
-mix	Total	80k	28k	2	67.1

Table 2: Statistics of RoleMRC. In particular, the column names S*, P*, #Turns, and #Words, stands for size of single-label data, size of pair-label data, average turns, and average number of words per reply, respectively. RoleMRC-mix expands RoleMRC by adding existing role-playing data.

timization (HPO) (Ouyang et al., 2022; Rafailov et al., 2023; Lu et al., 2024a; Hong et al., 2024). Considering that fine-tuning LLMs with relatively fixed data formats may lead to catastrophic forgetting (Kirkpatrick et al., 2017), we create RoleMRC-mix as a robust version by incorporating external role-playing data (RoleBench (Wang et al., 2023)) and general instructions (RLHFlow (Dong et al., 2024), UltraFeedback (Cui et al., 2023)).

4 Experimental Setup

4.1 Foundation Models and Post-tuning

We evaluate leading LLMs and fine-tuned models:

- **Proprietary LLMs.** gpt-3.5-turbo and gpt-4o.
- **SOTA Open-source LLMs.** Qwen2.5-7B/72B-Instruct (Yang et al., 2024) and LLaMA3.1-8B/70B-Instruct (Dubey et al., 2024).
- **Role-playing LLMs.** CharacterGLM-6B (Zhou et al., 2023b), Humanish-Llama-3.1-8B (Gallego, 2024), and Peach-9B-Roleplay (Peach, 2024).
- **Local Post-tuned LLMs.** We start with **pure base models Llama-3.1-8B and Qwen2.5-7B**. We first use single-label in RoleMRC-mix for SFT, then apply the pair-label set for Direct Preference Optimization (DPO, Rafailov et al. 2023).

4.2 Reference-based Metrics

We evaluate model-generated outputs using standard heuristic metrics commonly used in NLG:

- **BLEU** (Papineni et al., 2002) computes the precision of n-gram overlaps between generated text and a ground truth reference.
- **ROUGE** (Lin, 2004) measures the overlap of n-grams and longest common subsequences between the hypothesis and references. We include ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum to capture various granularities of overlap.
- **METEOR** (Banerjee and Lavie, 2005) aligns generated and reference tokens using stemming and synonym matching, aiming to provide a more linguistically grounded evaluation.
- **BERTScore F1** (Zhang et al., 2019) computes the similarity between generated and reference sentences using contextual embeddings.

For each metric, higher scores indicate better alignment with the reference lexically or semantically.

4.3 Reference-free LLM-as-a-judge

Apart from reference-based metrics, LLM-as-a-judge (Zheng et al., 2024) is another evaluation approach by instructional prompting advanced LLMs. In reference to Table 2, we curate a 1.4k test set similar to the On-scene MRC Dialogues and Ruled Chats, then evaluate model performance across five dimensions: (1) **Knowledge Boundary** focuses on distinguishing between answerable queries (“Answer”) and refusal scenarios (“Refusal”) in On-scene MRC Dialogues; (2) **Role Style** examines whether the model accurately produces role-specific responses (“Answer”, “No Answer”, “Refusal”, and “Attempt”) in On-scene MRC Dialogues without drifting into narration; while (3) **Multi-turn Instruction-following**, (4) **Nested Instruction-following**, and (5) **Prioritized Instruction-following** assess a model’s adherence to higher-level constraints in Ruled Chats.

We adopt a well-designed reference-free evaluation prompt (Figure 11), requiring the evaluator to verify whether the model’s role-playing performance comply with the corresponding rules, which avoids the risk of potential bias or error in any ground truth answer. Since we use a binary evaluation criterion, we directly extract 0 or 1 judgments from the feedback, enabling score comparison and accuracy computation. We chose gpt-4-turbo (openai, 2024b) as the evaluator, reducing the possible judging bias (Wataoka et al., 2024).

5 Evaluation on Inner RoleMRC Test Set

By leveraging the above **reference-based metrics** and **reference-free LLM-as-a-judge** approaches,

Models	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR	BERTScore F1
gpt-3.5-turbo	0.0234	0.2141	0.0606	0.1548	0.1579	0.1992	0.8552
gpt-4o	0.0288	0.2487	0.0742	0.1689	0.1835	0.2697	0.8516
CharacterGLM-6B	0.0058	0.1225	0.0253	0.0901	0.0967	0.1188	0.7944
Humanish-Llama-3.1-8B	0.0153	0.2062	0.0518	0.1309	0.3207	0.2389	0.8376
Peach-9B-Roleplay	0.0207	0.2297	0.0562	0.1544	0.1571	0.2299	0.8418
LLaMA3.1-8B-Instruct	0.0226	0.2277	0.0615	0.1509	0.1650	0.2594	0.8478
LLaMA3.1-70B-Instruct	0.0232	0.2258	0.0646	0.1500	0.1661	0.2632	0.8480
LLaMA3.1-8B-RoleMRC-SFT	0.1782	0.4628	0.2676	0.3843	0.3853	0.3975	0.8831
LLaMA3.1-8B-RoleMRC-DPO	0.1056	0.3989	0.1785	0.2988	0.3001	0.4051	0.8805
Qwen2.5-7B-Instruct	0.0224	0.2283	0.0621	0.1518	0.1599	0.2490	0.8471
Qwen2.5-72B-Instruct	0.0245	0.2350	0.0656	0.1554	0.1660	0.2579	0.8485
Qwen2.5-7B-RoleMRC-SFT	0.1963	0.4764	0.2744	0.3959	0.3968	0.4337	0.9063
Qwen2.5-7B-RoleMRC-DPO	0.1244	0.4178	0.1916	0.3164	0.3177	0.4205	0.8931

Table 3: Comparison of reference-based evaluation results on the RoleMRC test data. Our evaluation includes zero-shot query results for baselines (§4.1), and our SFT and DPO models fine-tuned on the RoleMRC-mix.

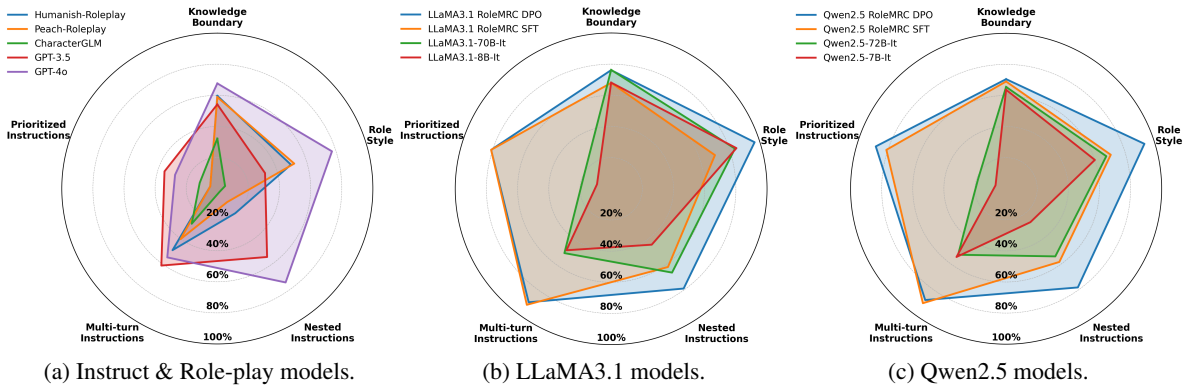


Figure 4: Visualization of reference-free LLM-as-a-judge results. We provide numerical result in Table 9.

we report evaluation on RoleMRC in what follows.

Performance of Proprietary LLMs. As shown in Table 3, gpt-4o achieves slightly higher BLEU, ROUGE, and METEOR scores than gpt-3.5-turbo. This observation is consistent with existing evaluations on general benchmarks (Achiam et al., 2023), and may also be influenced by the fact that our RoleMRC training data was synthesized by gpt-4o. The LLM-as-a-judge results (Figure 4a) similarly highlight gpt-4o’s strengths in Knowledge Boundary, Role Style, and Nested Instruction-following, whereas gpt-3.5-turbo outperforms gpt-4o on Prioritized and Multi-turn Instruction-following.

Evaluation on Commonly Used LLMs. For the LLaMA3.1 and Qwen2.5 families, larger models generally yield higher reference-based scores. For instance, LLaMA3.1-70B-Instruct slightly leads its 8B sibling (BLEU from 0.0226 to 0.0232), and Qwen2.5-72B-Instruct outperforms its 7B version (BLEU from 0.0224 to 0.0245). Although these improvements are modest, the results align with the broader observation that increasing model scale typically benefits language modeling and gener-

alization. Likewise, LLM-as-a-judge results (Figures 4b and 4c) show larger models are consistently better, particularly in Knowledge Boundary, Role Style, Nested and Prioritized Instruction-following.

Results of Role-playing LLMs. Three open-source role models obtain generally lower heuristic metrics than those general-purpose instruct models with similar size (Table 3). This discrepancy may stem from their training data, which emphasizes limited role styles and persona consistency rather than factual correctness and coverage. On LLM-as-a-judge (Figure 4a), CharacterGLM-6B again performs poorly, while Humanish-Llama-3.1-8B and Peach-9B-Roleplay show decent performance in Knowledge Boundary, Role Style, and Multi-turn Instruction-following, but struggle with Nested and Prioritized Instruction-following.

Impact on Task-Specific Fine-tuning. Our locally post-tuned **RoleMRC-SFT** models dramatically outperform all above baselines on reference-based metrics, improving BLEU by around 8× over their respective base models. Although the **SFT** models excel at matching ground-truth refer-

Model	RoleBenchInstEng (32.8k)				RoleBenchRoleEng (7.5k)			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Sum	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Sum
CharacterGLM-6B	0.1761	0.0546	0.1441	0.1530	0.1841	0.0628	0.1473	0.1552
Humanish-Llama-3.1-8B	0.2069	0.0639	0.1341	0.1645	0.1851	0.0468	0.1193	0.1432
Peach-9B-Roleplay	0.3216	0.1293	0.2573	0.2646	0.3454	0.1450	0.2705	0.2732
LLaMA3.1-8B-Instruct	0.2528	0.0864	0.1755	0.1931	0.2395	0.0754	0.1691	0.1844
LLaMA3.1-70B-Instruct	0.2846	0.1064	0.2062	0.2258	0.2756	0.1036	0.2036	0.2204
LLaMA3.1-8B-RoleMRC-SFT	0.3329	0.1601	0.2755	0.2770	0.3980	0.2022	0.3270	0.3278
LLaMA3.1-8B-RoleMRC-DPO	0.3605	0.1696	0.2812	0.2846	0.3970	0.1952	0.3149	0.3163
Qwen2.5-7B-Instruct	0.3216	0.1376	0.2437	0.2599	0.3337	0.1463	0.2582	0.2692
Qwen2.5-72B-Instruct	0.3225	0.1354	0.2364	0.2524	0.3370	0.1460	0.2577	0.2672
Qwen2.5-7B-RoleMRC-SFT	0.3963	0.1922	0.3294	0.3312	0.4442	0.2298	0.3680	0.3692
Qwen2.5-7B-RoleMRC-DPO	0.3969	0.1958	0.3143	0.3180	0.4298	0.2187	0.3452	0.3470

Table 4: Evaluations on external RoleBench (Wang et al., 2023) test set. The best results for each metric are **bold**.

ences, **DPO-aligned models win in reference-free LLM-as-a-judge**, in terms of *Knowledge Boundary* and *Role Style*. For instance, **LLaMA3.1-8B-RoleMRC-DPO** reaches a *Role Style* accuracy of 97.00%, while its SFT counterpart score is only around 70.00% (Figure 4b, detailed numbers in Appendix F). However, DPO models typically score lower on reference-based metrics (Table 3), reflecting a trade-off: shifting the model’s distribution toward instruction compliance and human preference can reduce exact lexical matches.

Overall, our curated evaluation framework realizes robust effectiveness for assessing LLM’s role-playing instruction-following capabilities.

6 Evaluation on External Benchmarks

We present cross-evaluation on external datasets.

[1] Fine-tuning on RoleMRC would not interfere the learning of other role-playing data. In Table 4, we follow Wang et al. (2023) and evaluate on two of their test sets: (1) RoleBenchInstEng (32.8k), an *instruction-based* split that tests how well models handle various instructions, and (2) RoleBenchRoleEng (7.5k), a *role-based* split that tests model performance across different roles. On RoleBenchInstEng, all RoleMRC-aligned models consistently outperform instruct and role-playing baselines. Notably, QWEN2.5-7B-ROLEMRC-SFT achieves significant gains, pushing ROUGE-1 and ROUGE-2 to 0.3963 and 0.1922, respectively. In the right panel of Table 4, results on RoleBenchRoleEng reveal similar trends. Our models outperform standard instruct models by sizeable margins. QWEN2.5-7B-ROLEMRC-SFT obtains the highest ROUGE-1 (0.4442) and ROUGE-L (0.3680). We thus conclude that RoleMRC did not counter the learning of RoleBench.

[2] RoleMRC helps naive LLMs gain high-quality generalized role-playing abilities. We

Model	OOD CharacterLLM		
	Single	Turns	General Δ
CharacterGLM-6B	5.9495	5.8676	1.00
Humanish-Llama-3.1-8B	5.3781	6.0444	0.68
Peach-9B-Roleplay	6.3074	6.0120	-2.46
LLaMA3.1-8B-Instruct	6.5244	6.0533	11.82
LLaMA3.1-8B-RoleMRC-SFT	6.4320	6.0196	4.08
LLaMA3.1-8B-RoleMRC-DPO	6.5179	5.9884	1.16
Qwen2.5-7B-Instruct	6.2485	5.9996	3.64
Qwen2.5-7B-RoleMRC-SFT	6.4520	6.0200	-0.33
Qwen2.5-7B-RoleMRC-DPO	6.5295	6.0311	1.14

Table 5: Out-of-distribution (OOD) evaluation on CharacterLLM (Shao et al., 2023), where models are evaluated on “Single” and “Turns” settings. “General Δ ” denotes the average gain for each model, compared with its fine-tuning starting point, across nine non-role-playing general-purpose benchmarks. Check details of OOD testing in Appendix G and A.

performed OOD tests of the RoleMRC-aligned models on an external role-playing dataset, Character-LLM, following its *Single* and *Turns* settings. The OOD results, in the middle columns of Table 5, show that among all role-playing models, our RoleMRC-aligned model (QWEN2.5-7B-ROLEMRC-DPO) reach a best score of 6.5295 in “single” evaluation and leads the “turns” evaluation.

[3] The local fine-tuned models did not overfit RoleMRC. In the last column of Table 5, we summarize the fine-tuning gains of different role models and general models across nine general-purpose benchmarks (e.g., GSM8K (Cobbe et al., 2021)). The “General Δ ” is obtained by calculating the performance gap between the fine-tuning endpoint model and the starting point, such as the improvement of LLaMA3.1-8B-Instruct relative to LLaMA3.1-8B. Except for Peach-9B-Roleplay, all role-playing LLMs have not lost general abilities when gaining role-playing abilities.

7 Analysis on Alignment Tax

Despite all the other role-playing and instruction-following abilities of the LLMs are enhanced dur-

Dimensions		BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	METEOR	BERTScore F1	LLM as judge
Knowledge Boundary	(B)	0.0950	0.3909	0.1631	0.2860	0.2860	0.3876	0.8798	74.67%
	(A)	0.1000 \uparrow	0.3946 \uparrow	0.1677 \uparrow	0.2924 \uparrow	0.2924 \uparrow	0.3883 \uparrow	0.8798	77.33% \uparrow
Role Style	(B)	0.1007	0.3948	0.1696	0.2886	0.2887	0.3883	0.8782	97.00%
	(A)	0.1283 \uparrow	0.3985 \uparrow	0.1889 \uparrow	0.3138 \uparrow	0.3228 \uparrow	0.3910 \uparrow	0.8790 \uparrow	94.50%
Multi-turn Instruction-following	(B)	0.1183	0.4196	0.2078	0.3232	0.3232	0.4506	0.8851	90.50%
	(A)	0.1185 \uparrow	0.4215 \uparrow	0.2110 \uparrow	0.3240 \uparrow	0.3240 \uparrow	0.4544 \uparrow	0.8852 \uparrow	92.00% \uparrow
Nested Instruction-following	(B)	0.1274	0.4010	0.1895	0.3138	0.3242	0.3944	0.8793	79.11%
	(A)	0.1283 \uparrow	0.3985	0.1889	0.3138	0.3228	0.3910	0.8790	79.75% \uparrow
Prioritized Instruction-following	(B)	0.0952	0.3639	0.1537	0.2700	0.2700	0.3840	0.8796	83.33%
	(A)	0.0965 \uparrow	0.3776 \uparrow	0.1531	0.2753 \uparrow	0.2753 \uparrow	0.3934 \uparrow	0.8807 \uparrow	73.81%

Table 6: Performance comparison category by each dimensions (B)efore and (A)fter neuron-level restrain.

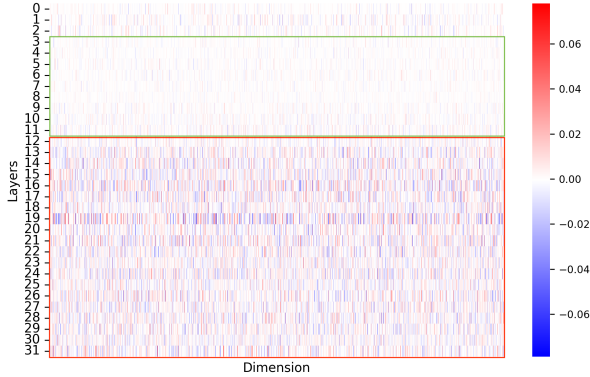


Figure 5: Discrepancies between SFT and DPO neuron activations (top-20% active neurons) in LLaMA3.1-8B for multi-turn instructions. Layers 3-11 show minimal changes (green), while layers 12-31 exhibit larger shifts (red).

ing the DPO alignment, we observe a slight yet common deterioration in multi-turn instruction-following performance (Appendix F). We refer to this phenomenon as an “alignment tax”, which is characterized by a gradual forgetting of knowledge acquired during pre-training (Ouyang et al., 2022).

Neuron-Level Localization. To identify the underlying cause of this alignment tax, we examine the neuron activation patterns of our ROLEMRC models (LLaMA3.1-8B SFT vs. DPO). Following Tang et al. (2024), we probe and collect activations from each attention layer, focusing on highly activated neurons by selecting the top 20% of activations. Specifically, for each input instruction, we measure activations when first forwarding the instruction. We then group the activation maps by the evaluation dimension of the test instruction, generating layer-specific differences in neuron usage.

Next, we count the activation frequency of each neuron and normalize it by the total number of test cases. Figure 5 visualizes the resulting discrepancy between the SFT and DPO models. Layers 3-11 exhibit minimal changes, whereas layers beyond the 13th show substantial activation differences, with layers 12-31 (highlighted in red) differing

the most. Notably, layer 19 is **significantly more active in multi-turn instruction**.

This observation aligns with Tang et al. (2024), who found that *only the top and bottom layers of a language model are primarily used for language processing*. These shifts in neuron activations suggest that *certain neurons are activated very differently between the SFT and DPO models*. Further details and results are provided in Appendix H.

Neuron-Level Restraint. After identifying these critical neuron subsets, we apply a minor scaling restraint (multiplicative factor $1 - 10^{-6}$) to modulate their impact. As shown in Table 6, constraining the most changed neurons **provides consistent improvements across both reference-based metrics and the LLM-as-a-judge approach**. In particular, multi-turn instruction accuracy increases by 1.6%, mitigating the alignment tax **without requiring further model retraining**. We also observe gains in dimensions of knowledge boundary and nested instruction-following, highlighting that targeted neuron-level adjustments can manipulate LLMs’ capabilities under alignment constraints.

8 Conclusion

We introduce RoleMRC, a large-scale fine-grained benchmark designed to improve and evaluate the role-playing and instruction-following abilities of LLMs. RoleMRC uniquely integrates role-specific multi-turn dialogues, MRC, and complex instruction-following scenarios. Experiments show that RoleMRC-aligned models outperform existing baselines in both reference-based and reference-free evaluations, and also perform well on both OOD role-playing and general-purpose benchmarks. We further conduct a neuron-level analysis to identify specific neurons with significant activation changes and apply targeted constraints to alleviate the alignment tax, thereby improving evaluation metrics without additional retraining.

Limitations

While RoleMRC significantly enhances the role-playing and instruction-following capabilities of LLMs, some limitations remain:

- While the role profiles in the dataset are diverse, system-level prompts used in the synthesized instructions are somewhat similar, which may limit the generalizability of downstream models.
- The reliance on synthetic data generated by models such as gpt-4o may introduce biases inherent in these models, affecting the performance and fairness of fine-tuned LLMs.
- While effective, mitigating the “alignment tax” on multi-turn instruction-following through neuron-level constraints may have a negative impact on other capabilities, suggesting that further interpretability research is needed.

Ethics Statement

The RoleMRC dataset is constructed with a strong commitment to ethical AI. The dataset does not contain any personal, sensitive, or identifiable information. Additionally, all role-playing interactions are designed to be safe and free from harmful, offensive, or misleading content. The dataset strictly adheres to responsible AI guidelines by avoiding the generation or reinforcement of biased, discriminatory, or deceptive narratives.

Acknowledgment

This work was supported by Tencent YouTu Lab and King’s College London (KCL). The data team of Tencent supported the batch requesting of gpt-4o during data synthesis, and the e-Research team of KCL supported the resources of model training upon the CREATE platform (King’s College London e-Research team, 2022).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Harsh Agrawal, Aditya Mishra, Manish Gupta, and Mausam. 2023. [Multimodal persona based generation of comic dialogs](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14164, Toronto, Canada. Association for Computational Linguistics.
- 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.
- AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7432–7439.
- Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024b. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.
- Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, et al. 2024c. Holllwood: Unleashing the creativity of large language models in screenwriting via role playing. *arXiv preprint arXiv:2406.11683*.
- Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhao Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. [Large language models meet harry potter: A dataset for aligning dialogue agents with characters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias