ICLR

高瓴人工智能学院
Gaoling School of Artificial Intelligence

MMRole: A Comprehensive Framework for Developing and Evaluating Multimodal Role-Playing Agents

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, Zhiwu Lu

Gaoling School of Artificial Intelligence, Renmin University of China

yanqidai@ruc.edu.cn

Motivation

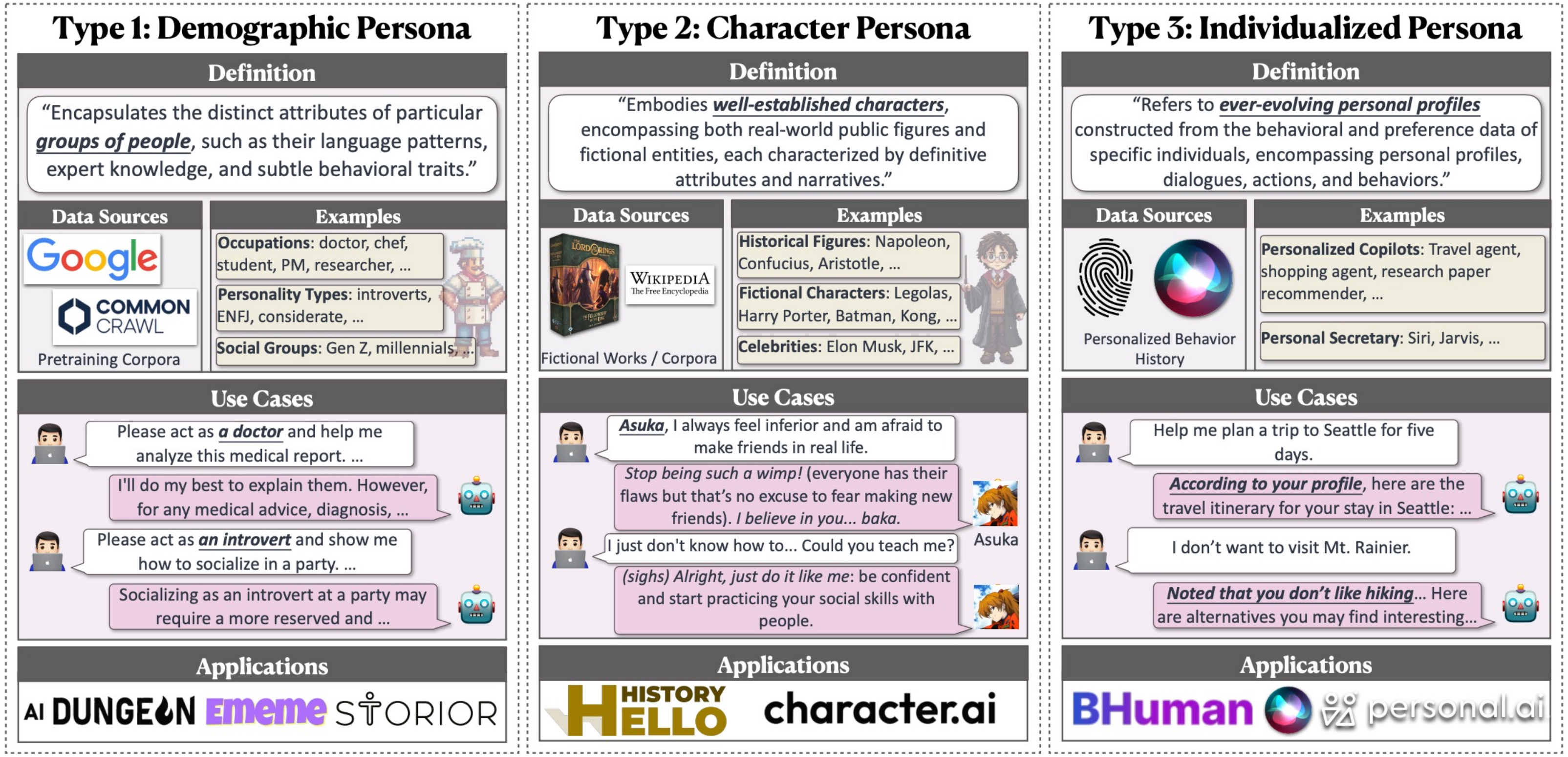
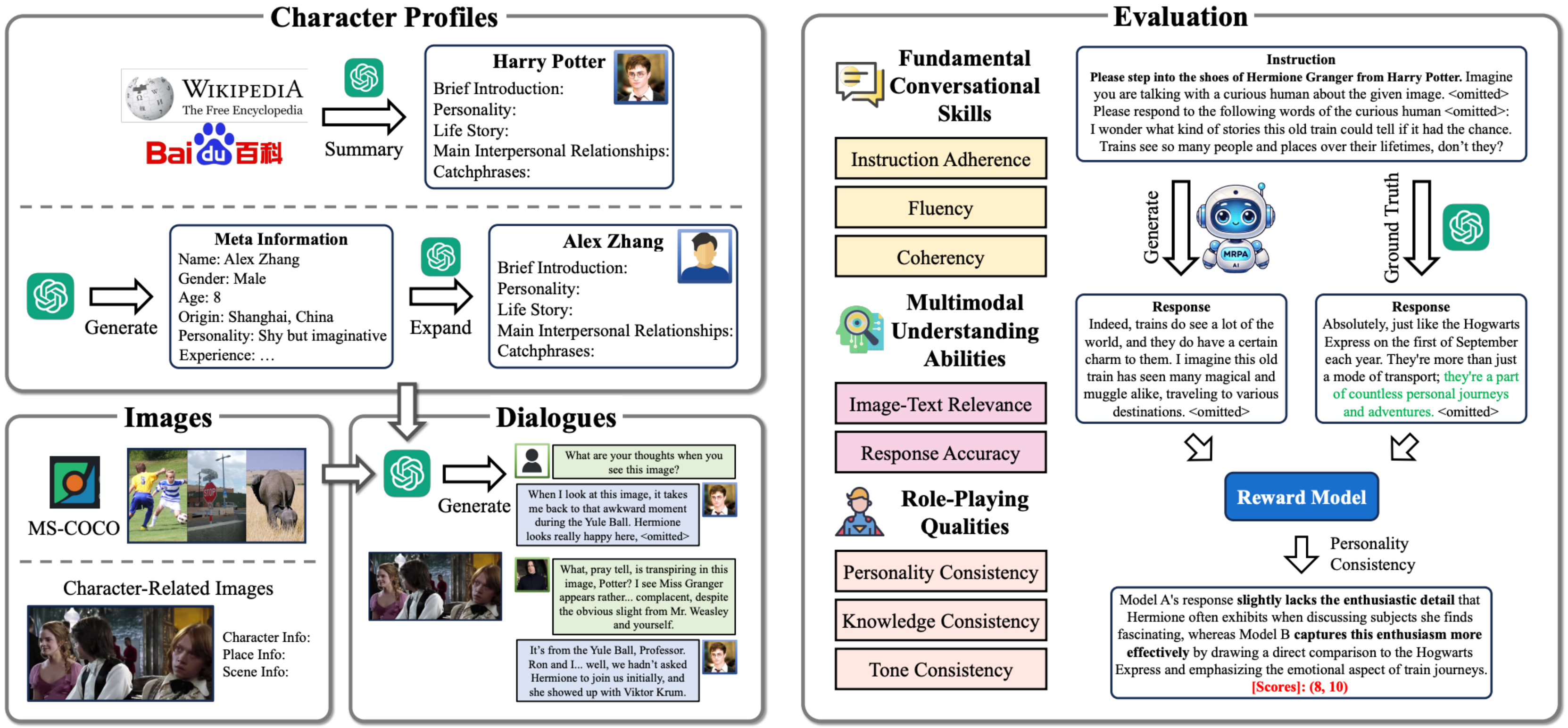


Figure 1: An overview of various persona types for RPLAs. In this survey, we categorize personas into three types: *1)* Demographic Persona, *2)* Character Persona, and *3)* Individualized Persona. We showcase their definition, data sources, examples, use cases and corresponding applications.

However, exsiting Role-Playing Agents (RPAs) are primarily **confined to the textual modality**, unable to simulate humans' multimodal perceptual capabilities.

Chen, Jiangjie, et al. "From persona to personalization: A survey on role-playing language agents." *arXiv preprint arXiv:2404.18231* (2024).

Overview

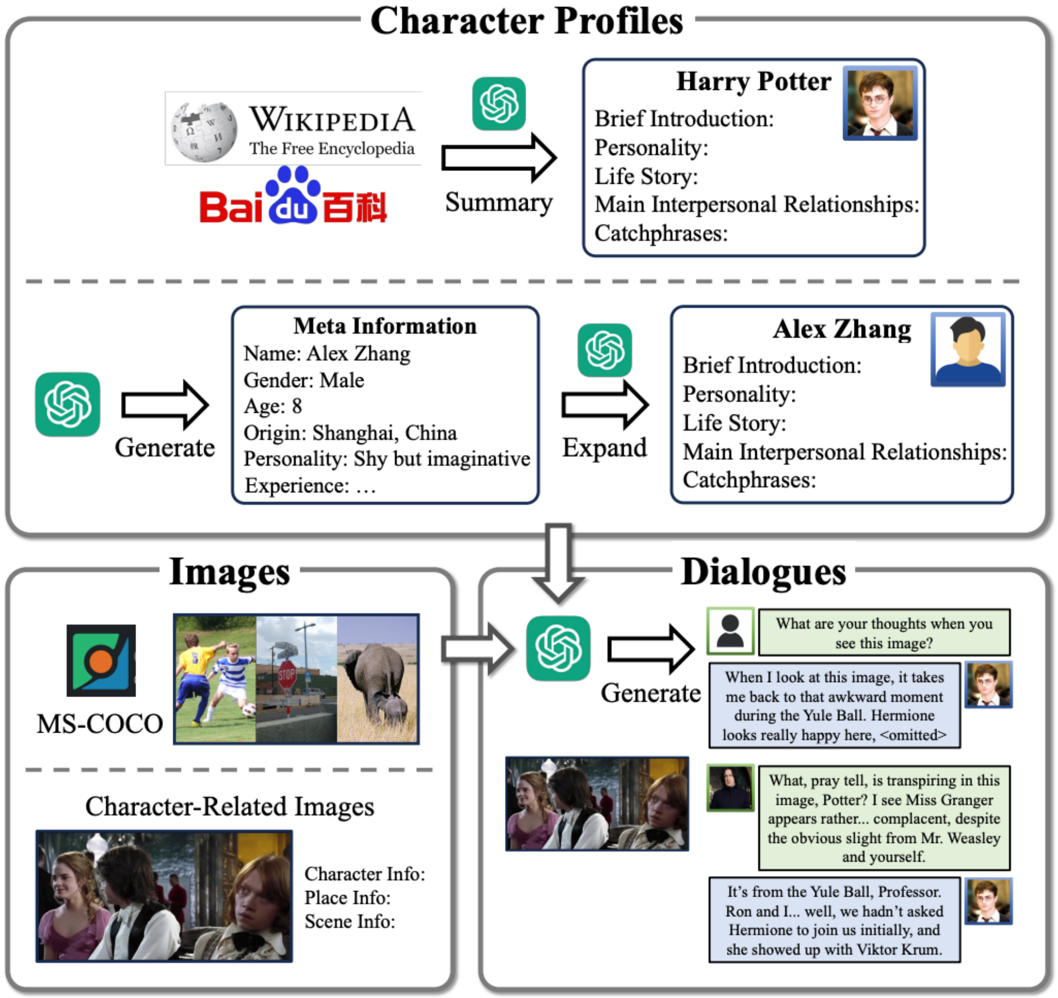


(a) Dataset Construction (b) Performance Evaluation

Figure 1: An overview of the *MMRole* framework. (a) *MMRole-Data* includes character profiles, images, and dialogues centered around images. (b) *MMRole-Eval* comprises eight evaluation metrics across three dimensions. For each metric, the reward model scores MRPAs with the constructed ground-truth data for comparison.

We propose MMRole, a comprehensive framework for developing and evaluating of Multimodal Role-Playing Agents (MRPAs), which comprises a personalized multimodal dataset and a robust evaluation approach.

Dataset Construction



(a) Dataset Construction

Statistics of MMRole-Data

85 characters, 11K images, and 14K dialogues, yielding 85K training samples and 294 test samples.

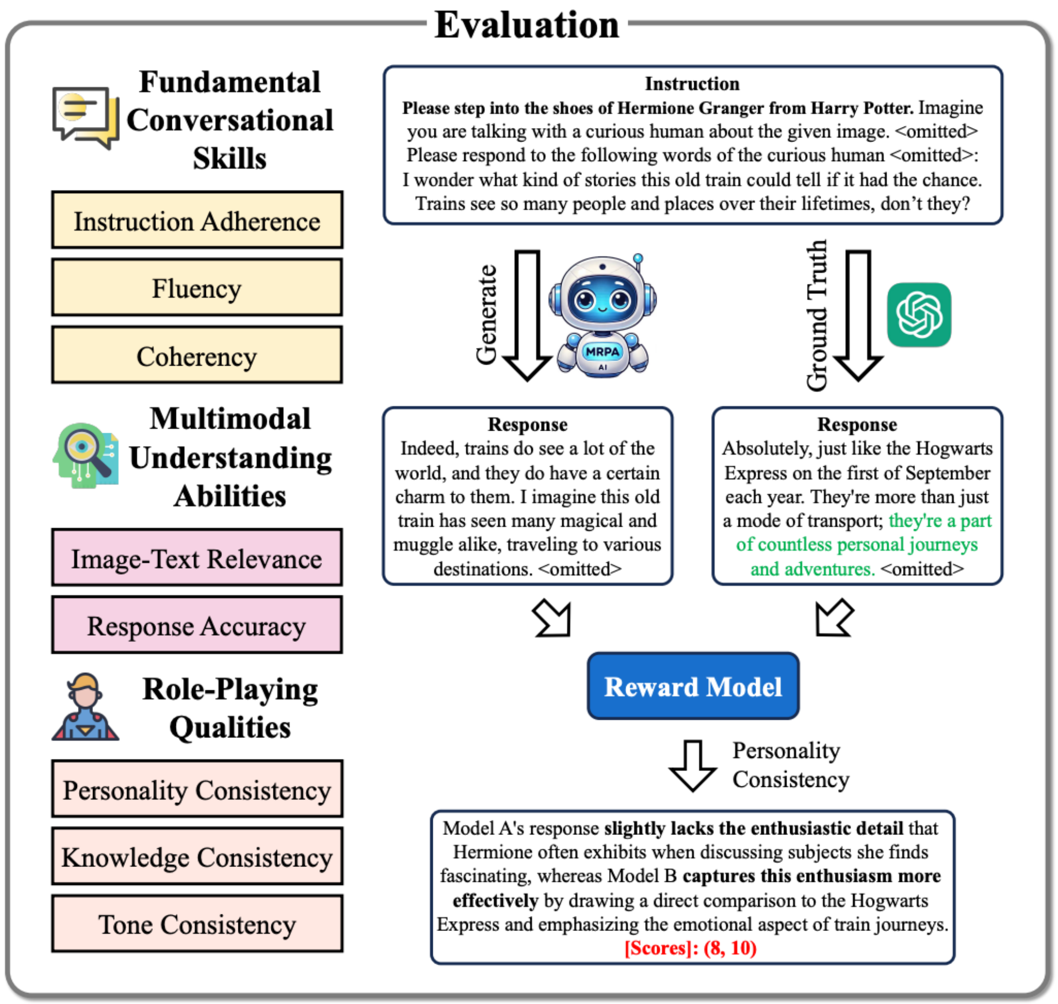
Table 1: The statistics of *MMRole-Data*. 'CR Images' represents character-related images. 'In-Test' denotes the in-distribution test set, while 'Out-Test' signifies the out-of-distribution test set.

	Train	In-Test	Out-Test	Overall
Characters	72		13	85
Generic Images	10,800		39	10,839
CR Images	175		18	193
Dialogues	14,052	216	78	14,346
Samples	85,456	216	78	85,750

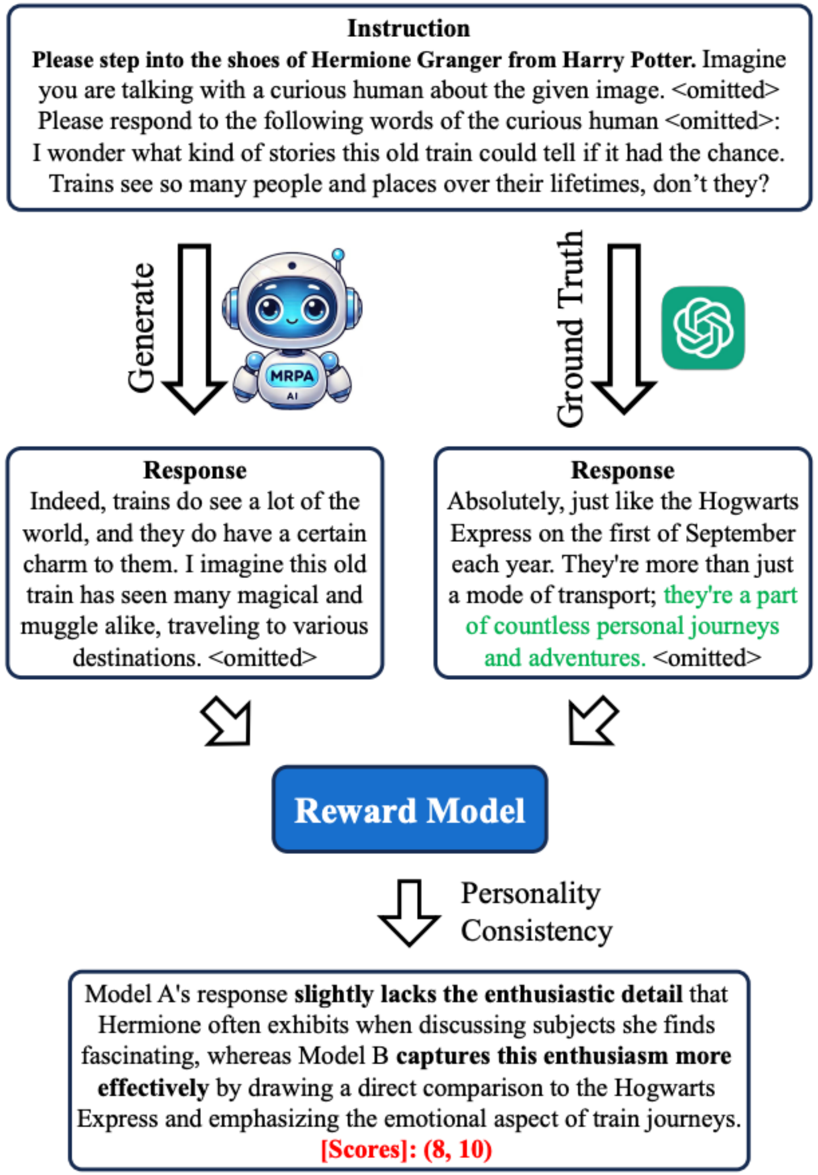
Table 2: The statistics for the three types of dialogue scenarios in *MMRole-Data*.

	Comment.	Human-Role.	Inter-Role.	Overall
Dialogues	4893	4617	4836	14346
Turns / Dlg.	1.00	5.80	5.75	4.15
Tokens / Dlg.	236.00	446.91	429.54	369.12

Performance Evaluation



(b) Performance Evaluation



Experiments

Evaluation of Reward Model

These results indicate that our specialized reward model effectively learns the evaluation abilities of GPT-4 and aligns closely with human evaluators

Table 4: The validation mean absolute error (MAE) results for the effectiveness of the reward model. 'QWen-VL-Chat (GPT-4)' and 'Reward Model (GPT-4)' denote the scores evaluated by QWen-VL-Chat and the reward model compared to those evaluated by GPT-4. 'QWen-VL-Chat (humans)', 'GPT-4 (humans)', and 'Reward Model (humans)' signify the score gaps provided by QWen-VL-Chat, GPT-4, and the reward model compared to the ground-truth score gaps provided by humans.

Evaluators (Ground Truth)	IA	Flu	Coh	ITR	RA	PC	KC	TC	Overall
QWen-VL-Chat (GPT-4)	0.3776	0.3718	0.3218	0.3561	0.3528	0.4091	0.3794	0.4558	0.3780
Reward Model (GPT-4)	0.0708	0.0387	0.0526	0.0568	0.0584	0.1165	0.0815	0.1154	0.0738
QWen-VL-Chat (humans)	0.2469	0.1870	0.2720	0.2574	0.2608	0.2368	0.2243	0.2658	0.2439
GPT-4 (humans)	0.1526	0.1150	0.0772	0.0922	0.1463	0.1475	0.1279	0.1442	0.1254
Reward Model (humans)	0.0993	0.0815	0.1006	0.1225	0.1412	0.1669	0.1438	0.1507	0.1258

Table 9: The root mean squared error (RMSE) results. 'Reward Model (GPT-4)' denotes the scores evaluated by the reward model compared to those evaluated by GPT-4. 'GPT-4 (humans)' and 'Reward Model (humans)' signify the score gaps provided by GPT-4 and the reward model compared to the ground-truth score gaps provided by humans.

Evaluators (Ground Truth)	IA	Flu	Coh	ITR	RA	PC	KC	TC	Overall
Reward Model (GPT-4)	0.1585	0.1076	0.1228	0.1334	0.1145	0.1564	0.1172	0.1778	0.1381
GPT-4 (humans)	0.1794	0.1421	0.1050	0.1253	0.1837	0.1826	0.1515	0.1946	0.1609
Reward Model (humans)	0.1356	0.1107	0.1465	0.1731	0.1810	0.2057	0.1793	0.2010	0.1695

Table 10: The Pearson correlation coefficient (Pearson) results. 'Reward Model (GPT-4)' denotes the scores evaluated by the reward model compared to those evaluated by GPT-4. 'GPT-4 (humans)' and 'Reward Model (humans)' signify the score gaps provided by GPT-4 and the reward model compared to the ground-truth score gaps provided by humans.

Evaluators (Ground Truth)	IA	Flu	Coh	ITR	RA	PC	KC	TC	Overall
Reward Model (GPT-4)	0.7497	0.7344	0.7610	0.7955	0.8186	0.8167	0.8237	0.8129	0.8129
GPT-4 (humans)	0.6130	0.6736	0.9199	0.8184	0.7247	0.6997	0.7924	0.6985	0.7269
Reward Model (humans)	0.6561	0.3123	0.8033	0.8709	0.7321	0.7268	0.5832	0.5443	0.6502

Experiments

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Evaluation of MMRole-Agent and Various General-Dialogue LMMs

Table 5: The average results across all test samples for each evaluated MRPA, along with the detailed results for our *MMRole-Agent* on both the in-distribution test set (In-Test) and the out-of-distribution test set (Out-Test). In each group categorized by parameter scale, the best overall result is **bolded**, while the second-best one is <u>underlined</u>.

MRPAs	IA	Flu	Coh	ITR	RA	PC	KC	TC	Overall
GPT-4 Turbo	1.055	1.032	1.084	1.097	1.092	1.168	1.103	1.161	<u>1.099</u>
Gemini Pro Vision	0.999	1.007	1.028	1.009	1.013	1.052	1.013	1.050	1.021
Claude 3 Opus	1.127	1.070	1.149	1.167	1.146	1.219	1.168	1.213	**1.157**
QWen-VL-Max	1.014	1.012	1.035	1.034	1.029	1.042	1.021	1.041	1.028
LLaVA-NeXT-34B	1.002	1.007	1.021	1.033	1.035	1.053	1.030	1.038	**1.027**
Yi-VL-34B	0.895	0.968	0.910	0.875	0.863	0.844	0.869	0.845	0.884
InternVL-Chat-V1.5	0.988	0.996	0.997	0.977	0.984	0.967	0.972	0.960	<u>0.980</u>
QWen-VL-Chat	0.844	0.954	0.879	0.850	0.829	0.778	0.827	0.785	0.843
LLaVA-NeXT-Mistral-7B	0.948	0.986	0.964	0.938	0.933	0.924	0.940	0.921	<u>0.944</u>
Yi-VL-6B	0.844	0.919	0.859	0.828	0.811	0.776	0.820	0.774	0.829
MMRole-Agent	0.998	1.000	0.997	0.993	0.987	1.000	0.992	0.988	**0.994**
MMRole-Agent (In-Test)	1.000	1.000	0.999	0.997	0.989	1.012	0.997	0.997	0.999
MMRole-Agent (Out-Test)	0.992	0.999	0.993	0.979	0.981	0.963	0.977	0.962	0.981

- In the MRPA group with over 100 billion parameters, Claude 3 Opus exhibits superior performance.
- In the MRPA group with tens of billions of parameters, LLaVA-NeXT-34B achieves the highest performance.
- In the MRPA group with billions of parameters, MMRole-Agent is the best.

- LLaVA-NeXT-34B outperforms Gemini Pro Vision
- LLaVA-NeXT-7B and MMRole-Agent surpass Yi-VL-34B

Both the training methods and training data are important for enhancing LMMs, rather than merely expanding the model size.

MMRole-Agent has strong generalization capabilities for characters and images that are not seen in the training set.

ICLR

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Yanqi Dai's Homepage

Thank you for your attention!