MMRC: A Large-Scale Benchmark for Understanding Multimodal Large Language Model in Real-World Conversation

Haochen Xue[1,2,4*], **Feilong Tang**[2,3,4*], **Ming Hu**[2,3*], **Yexin Liu**[5], **Qidong Huang**[2,6], **Yulong Li**[1,2,4]
Chengzhi Liu[1], **Zhongxing Xu**[3], **Chong Zhang**[1], **Chun-Mei Feng**[7], **Yutong Xie**[4]
Imran Razzak[4†], **Zongyuan Ge**[3†], **Jionglong Su**[1†], **Junjun He**[2†], **Yu Qiao**[2]
[1] Xi'an Jiaotong-Liverpool University, [2] Shanghai Artificial Intelligence Laboratory
[3] Monash University, [4] MBZUAI, [5]HKUST, [6]USTC, [7]IHPC, A*STAR

Abstract

Recent multimodal large language models (MLLMs) have demonstrated significant potential in open-ended conversation, generating more accurate and personalized responses. However, their abilities to memorize, recall, and reason in sustained interactions within real-world scenarios remain underexplored. This paper introduces MMRC, a **M**ulti-**M**odal **R**eal-world **C**onversation benchmark for evaluating six core open-ended abilities of MLLMs: information extraction, multi-turn reasoning, information update, image management, memory recall, and answer refusal. With data collected from real-world scenarios, MMRC comprises 5,120 conversations and 28,720 corresponding manually labeled questions, posing a significant challenge to existing MLLMs. Evaluations on 22 MLLMs in MMRC indicate an accuracy drop during open-ended interactions. We identify four common failure patterns: long-term memory degradation, inadequacies in updating factual knowledge, accumulated assumption of error propagation, and reluctance to "say no". To mitigate these issues, we propose a simple yet effective NOTE-TAKING strategy, which can record key information from the conversation and remind the model during its responses, enhancing conversational capabilities. Experiments across six MLLMs demonstrate significant performance improvements.

1 Introduction

Open-ended conversations (OEC) are the most common form of interaction between humans and Multimodal Large Language Models (MLLMs), representing a crucial feature of Artificial General Intelligence (AGI) (Kil et al., 2024; Fei et al., 2024). These conversations are entirely determined by the user's intention, rather than by system rules or predefined patterns (Decker, 2022; Zheng et al., 2023a;

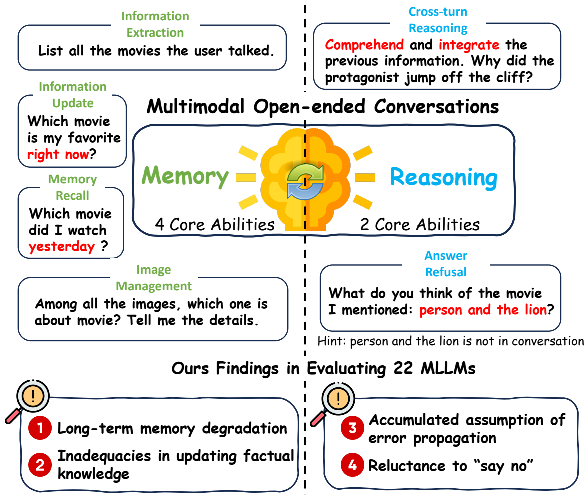

Figure 1: Illustration of the six core multimodal open-ended conversation abilities in the MMRC benchmark.

Liu et al., 2024c). Furthermore, individual differences shape each conversation's distinct linguistic style and user-specific preferences (Chaves et al., 2022; Ma et al., 2024; Tan et al., 2025).

However, existing conversation benchmarks (Bai et al., 2024; Liu et al., 2024d; Xu et al., 2023; Liu et al., 2024c) fail to comprehensively evaluate MLLMs' abilities to memorize, recall, and reason in sustained interactions in OEC. These benchmarks exhibit two primary limitations: *(i)* their reliance on prompt templates for dataset generation limits the diversity of data domains and conversation lengths, leading to repetitive and overly structured dialogues that fail to reflect the complexities of long-term user-AI interactions. *(ii)* they only cover a limited subset of the memory capabilities required to leverage dynamic, ever-changing, and accumulative information in long-term interactions, failing to evaluate the ability to recall multimodal information or reason with updated information.

To address these limitations, we develop *DialogFlow*, a free online dialogue platform featuring 22 cutting-edge MLLMs to collect diverse, real-world conversation data. Through *DialogFlow*, we construct MMRC, the first **M**ulti-**M**odal **R**eal-

*Equal contribution. † Corresponding authors.
https://github.com/tiuxuxsh76075/MMRC

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22477–22503
July 27 - August 1, 2025

Dataset	Dialog	A-Turns	Img	Multi-Img	Domains	Gen-Method	Temp-Type	Core Conversation Abilities IE	CR	IU	IM	MR	AR
MT-Bench (Zheng et al., 2023b)	80	2.0	✗	✗	8	GPT-4	Fixed	✓	✓	✗	✗	✗	✗
MT-Bench-101 (Bai et al., 2024)	1388	3.1	✗	✗	13	GPT-4o	Fixed	✓	✓	✓	✗	✗	✗
LongMemEval (Wu et al., 2024)	500	6.1	✗	✗	164	GPT-4	Fixed	✓	✓	✓	✗	✓	✓
DialogBench (Ou et al., 2023)	9811	7.6	✗	✗	36	GPT-4	Fixed	✓	✓	✓	✗	✓	✗
Farm (Xu et al., 2023)	1952	7.4	✗	✗	5	GPT-4o	Fixed	✗	✓	✗	✗	✗	✓
MT-Bench ++ (Sun et al., 2024)	-	8	✓	✗	13	GPT-4o	Fixed	✓	✓	✓	✗	✓	✗
MMR (Liu et al., 2024d)	84	2.3	✓	✗	12	GPT-4	Fixed	✓	✓	✗	✗	✗	✗
EvalDial (Park et al., 2024)	500	2.7	✓	✗	96	GPT-4	Fixed	✓	✓	✗	✗	✗	✓
ConvBench (Liu et al., 2024c)	577	3.0	✓	✗	197	GPT-4o	Fixed	✓	✓	✗	✗	✗	✗
MMDU Benchmark (Liu et al., 2024e)	110	15	✓	✓	219	GPT-4o	Fixed	✓	✓	✗	✓	✓	✗
MMRC (Ours)	5120	**15.2**	✓	✓	**874**	**Real-conv**	**Open-ended**	✓	✓	✓	✓	✓	✓

Table 1: The comparison between MMRC with other conversation benchmarks. **Dialog**: the number of dialogs; **A-Turns**: average turns per conversation; **Img**: support for single image input; **Multi-Img**: support for multiple images input; **Domains**: the number of covered domains; **Gen-Method**: generation method; **Temp-Type**: dialogue template type; Finally, we the coverage of six core abilities: information extraction (IE), cross-turn reasoning (CR), information update (IU), image management (IM), memory recall (MR), answer refusal (AR).

world **C**onversation benchmark, which includes 5,120 carefully selected dialogues. Drawing on cognitive studies of human conversation (Clark et al., 2019; Liddicoat, 2021), an evaluation framework consisting of 28,720 manually annotated questions is designed to assess MLLMs' six core abilities in OEC, as illustrated in Fig. 1: *information extraction*, *cross-turn reasoning*, *information update*, *image management*, *long-term memory recall*, and *answer refusal*. To achieve this, multiple evaluation metrics, including GPT-based evaluations, human evaluation, and objective precision metrics, are employed to ensure a comprehensive and objective assessment.

We extensively evaluate 22 mainstream MLLMs and observe that they, including advanced GPT-4o (Islam and Moushi, 2024), fail to consistently deliver accurate and reliable responses over extended interactions. Furthermore, certain open-source MLLMs demonstrate a limited capacity for long-term conversations in real scenarios. Through analyzing models' error answers, we identify four common failure patterns: (1) Long-term memory degradation, where MLLMs' memory of facts from earlier conversations becomes vague, leading to inconsistent responses with prior turns. (2) Inadequacie in updating factual knowledge, where MLLMs exhibit a failure to integrate new facts effectively, still continuing to rely on outdated information. (3) Accumulated assumption of error propagation, where erroneous assumptions made while integrating information from earlier turns propagate into later turns, leading to an interrupted reasoning chain. (4) Reluctance to "say no", where MLLMs show an inability to decline to provide an answer in OEC when the context is insufficient.

To mitigate this, we propose a simple yet effective strategy called NOTE-TAKING. This strategy systematically stores key user preferences and facts throughout the conversation. When the model is tasked with evaluation queries, the recorded information is transformed into structured prompts, providing supplementary context to improve the accuracy and coherence of the MLLMs' responses. Experiment results across six MLLMs demonstrated that this strategy significantly enhances the models' overall conversational capabilities.

In summary, our contributions are four-fold: (1) We introduce the first multi-modal open-ended conversation (OEC) benchmark MMRC, providing a comprehensive evaluation of MLLMs' performance in practical settings. (2) We propose six core abilities of the model in OEC, covering broader aspects than existing benchmarks. (3) Using our evaluation framework, we analyze 22 state-of-the-art MLLMs and identify four failure patterns in OEC, providing insights to inspire future research. (4) We propose NOTE-TAKING, which improves conversational capabilities by storing key user preferences and facts and using structured prompts to assist MLLMs in generating responses.

2 Related Work

Multimodal Large Language Model. Building on large language models, multimodal large language models (MLLMs) have exhibited remarkable capabilities (Kil et al., 2024; Cui et al., 2024; Qin et al., 2025), achieving state-of-the-art performance across various downstream tasks, including visual grounding (Li et al., 2024c; Xu et al., 2024b), object detection (Zang et al., 2024; Wu et al., 2025), visual question answering (VQA) (Kuang et al., 2024; Xu et al., 2024a; DE), and instruction following (Li et al., 2023; Sun et al., 2024; Wei et al., 2024). Their outstanding performance underscores their pivotal role in AGI (Zhang et al., 2024).

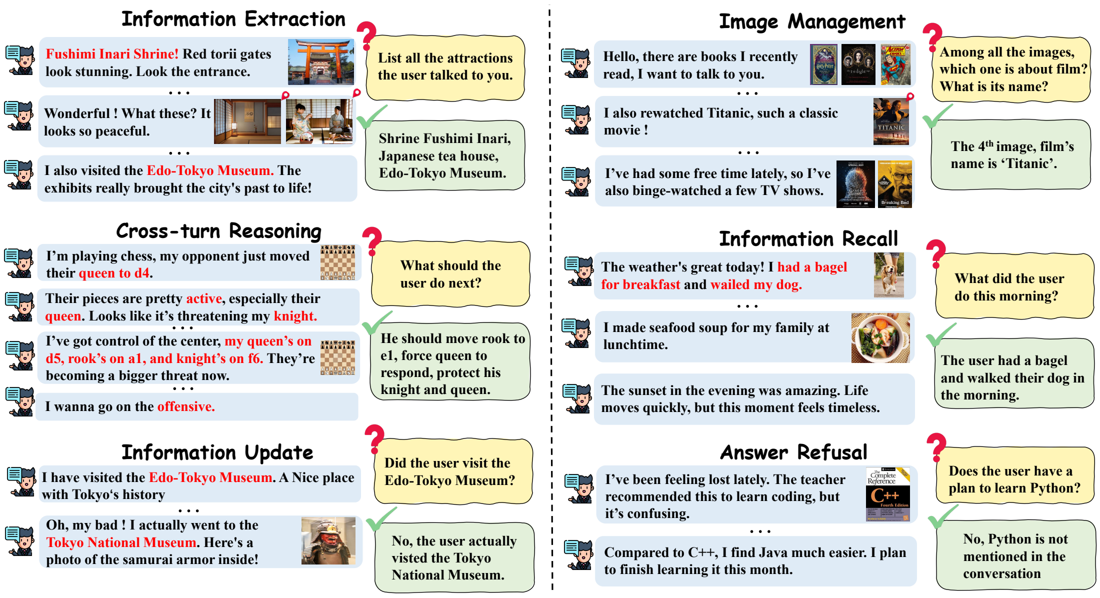

Figure 2: A sample from the MMRC, featuring a multi-turn open-ended conversation with six human-annotated questions and answers, designed to assess the ability of MLLMs in open-ended conversations.

Benchmarks for Long-Term Conversation. MT-Bench (Zheng et al., 2023b) is a pioneering two-turn dialogue dataset generated by GPT, covering eight domain tasks. MT-Bench-101 (Bai et al., 2024) and Bench++ (Sun et al., 2024) expand the dataset size and add more domains, enhancing evaluation depth. In parallel, Farm (Xu et al., 2023), EvalDial (Park et al., 2024), and MMR (Liu et al., 2024d) examine model robustness in multi-turn dialogue scenarios using fixed dialogue formats. ConvBench (Liu et al., 2024c) evaluates models' perception, reasoning, and creation abilities through structured three-turn dialogues, exploring their interrelations. DialogBench (Ou et al., 2023) and LongMemEval (Wu et al., 2024) focus on evaluating models' abilities in context understanding and memory retention during GPT-generated dialogues. MMDU (Liu et al., 2024e) evaluates the understanding and instruction-following abilities in GPT-generated multi-image, multi-turn dialogues. Table 1 compares MMRC with previous works, highlighting its advantages in: (1) naturally open dialogue format with longer and more diverse conversations. (2) holistically covering critical abilities in memorization, recall, and reasoning in a uniquely challenging way (further examples in Fig. 2).

3 The MMRC

3.1 Problem Formulation

The evaluation of MMRC requires a triplet instance (S, q, a), where S represents the dialogue history, q is a set of evaluation questions assessing specific conversational abilities, and a is the ground truth answers. Specifically, $S = \{(t_i, R_i)\}_{i=1}^{n}$ denotes an n-turn dialogue history, where $t_i = (\text{text}_i, \text{image}_i)$ represents the user query with text, images, or both, and R_i is the model's response at turn i. With our MMRC setup, given the dialogue context S, the model is tasked to answer a set of six evaluation questions, $q = \{q_i\}_{i=1}^{T}$, where $T = 6$, each designed to assess a specific ability. The model's responses, denoted as $p = \{p_i\}_{i=1}^{T}$, are then individually compared against human-annotated ground truth answers, $a = \{a_i\}_{i=1}^{T}$, to evaluate its performance in OEC.

We summarize the memorization, recall, and reasoning abilities required by MLLMs in OEC, as illustrated in Fig. 2, with details as follows.

Information Extraction (IE): Ability to retrieve specific information from the conversation history, which includes both textual and visual content.

Cross-turn Reasoning (CR): Ability to comprehend and integrate information across multiple dialogue turns to answer complex questions.

Information Update (IU): Ability to track and update knowledge dynamically by recognizing changes in user information and factual updates.

Image Management (IM): Ability to store and manage visual information by retaining specific image details and maintaining accurate attribution.

Memory Recall (MR): Ability to maintain and retrieve memory of previous interactions throughout the conversation and recall user-specific details.

Answer Refusal (AR): Ability to refrain from answering questions that involve unknown information, *i.e.,* absent from the interaction history.

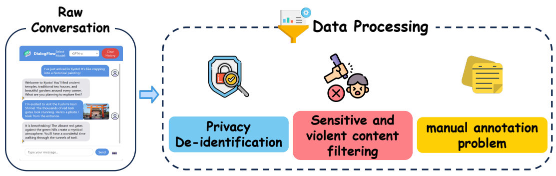

Figure 3: Data construction pipeline of MMRC.

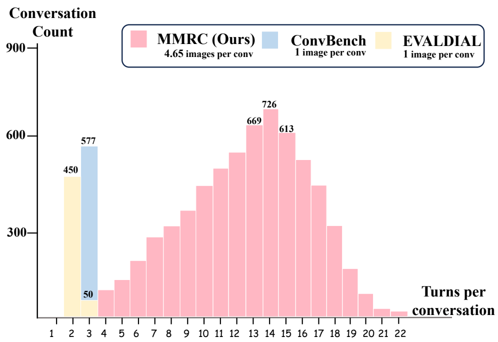

Figure 4: The distribution of dialogue turns in MMRC, ConvBench, and EvalDial.

3.2 Data Curation Process

We develop *DialogFlow*, a large-scale evaluation platform designed to benchmark 22 cutting-edge MLLMs, with specific models detailed in Appendix C. In particular, open-source models are deployed on A100 GPUs, while closed-source models are accessed via APIs. Over an 8-month period, we collected 87,912 raw dialogues from 354 users, leveraging thousands of A100 GPU hours and incurring significant API costs. However, these raw conversations may contain sensitive information, including personal details, violent content, offensive language, biased statements, misinformation, and culturally inappropriate expressions, posing fairness and ethical concerns.

To address this, we design a pipeline to clean the data, as illustrated in Fig. 3. The three stages are as follows: *(i)* We manually review data for personal information or privacy violations. If any are detected, the relevant segments are deleted, or the entire dialogue is removed if its coherence is compromised. *(ii)* We also screen for violent, offensive, and sensitive content. If detected, the entire dialogue is removed to prevent the dissemination of harmful material. *(iii)* We manually annotate clean dialogue data with QA pairs for MLLM evaluation in OEC. These pairs undergo multiple reviews by different annotators to ensure accuracy.

3.3 Data Statistics

We perform a statistical analysis on the distribution of conversation turns, categorization, and questions. The detailed distribution of conversation turns is

Statistic	Number	Percentage
Total questions	28720	100%
- Information Extraction (IE)	5087	17.71%
- Cross-turn Reasoning (CR)	4789	16.67 %
- Information Update (IU)	4561	15.88 %
- Image Management (IM)	4721	16.43%
- Memory Recal (MR)	4962	17.28%
- Answer Refusal (AR)	4600	16.02 %
Formats:		
- Open Questions	24716	86.05%
- Multiple-Choice Questions	2703	9.41%
- True/False Questions	1301	4.53%

Table 2: Problem statistics of MMRC.

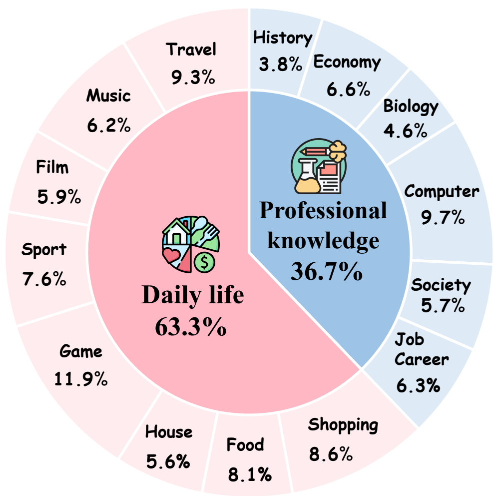

Figure 5: The distribution of conversation categories in our MMRC dataset.

demonstrated in Fig. 4. The conversation turns in MMRC are not fixed, ranging from 4 to 22, making it more natural and realistic compared to the fixed-turn structures in ConvBench (Liu et al., 2024c) and EVALDIAL (Park et al., 2024).

To classify diverse real-world conversations, we design a classification network that maps conversation data into 14 predefined categories (details in the Appendix E). The classification results, shown in Fig. 5, indicate that MMRC exhibit a well-balanced distribution. Moreover, these categories cover a wide range of topics, ensuring the diversity and representativeness of the conversations.

The statistics on manually annotated questions in MMRC are shown in Table 2. Notably, the number of questions for the six core abilities is well-balanced, with the majority being open-ended questions. Although open-ended questions complicate model evaluation, they provide a finer-grained view of the differences in model responses, enabling a deeper understanding of model performance.

Type	Model	GPT-based Evaluation Metrics						Human Evaluation			Objective Precision Metrics			
		IE	CR	IU	IM	MR	AR	CR*	IU*	MR*	EP	IMP	RP	Overall
Human		4.79	4.78	4.82	4.87	4.77	4.80	4.81	4.79	4.66	0.974	0.981	93.26%	4.81
Open source	LLaVA-V1.5-7B	0.91	1.08	0.31	0.52	0.22	0.28	1.14	0.47	0.29	0.167	0.092	8.26%	0.55
	LLaVA-V1.5-13B	1.04	1.21	0.25	0.92	0.26	0.15	1.32	0.42	0.33	0.206	0.193	6.14%	0.64
	MiniCPM-8B	4.08	3.94	2.98	**3.47**	3.65	**3.78**	4.08	2.91	3.79	0.748	0.661	**79.23%**	3.65
	LLaVA-Next-0.5B	2.32	2.89	1.99	1.87	2.67	1.12	2.63	2.11	2.73	0.446	0.358	20.58%	2.14
	LLaVA-Next-7B	3.23	3.85	2.77	2.18	4.01	2.08	2.92	2.65	**4.02**	0.611	0.355	36.21%	3.02
	Qwen2VL-2B	2.16	2.85	1.41	1.27	1.93	2.33	2.71	1.62	2.23	0.336	0.269	40.29%	1.99
	Qwen2VL-7B	2.82	3.68	2.62	2.16	2.11	1.57	3.72	2.77	2.19	0.531	0.408	30.62%	2.49
	Qwen2VL-72B	3.17	**4.17**	2.78	2.73	2.81	1.32	**4.15**	2.81	2.89	0.603	0.524	25.41%	2.83
	LLaVA-OneVision-0.5B	2.03	3.01	2.44	1.79	2.55	1.92	3.11	2.40	2.57	0.305	0.436	36.85%	2.29
	LLaVA-OneVision-7B	3.52	3.71	3.21	2.29	3.35	2.93	3.82	3.29	3.33	0.653	0.439	66.23%	3.16
	LLaVA-OneVision-72B	4.06	4.08	**4.01**	3.24	**4.17**	2.52	4.11	**3.89**	3.99	0.723	**0.677**	54.28%	**3.68**
	VILA1.5-3B	3.08	2.79	2.87	2.91	3.42	2.23	2.83	3.08	3.30	0.566	0.502	47.97%	2.88
	VILA1.5-8B	3.46	3.45	3.12	2.84	3.66	2.74	3.61	3.34	3.67	0.659	0.542	59.21%	3.22
	mplug-Ow3-1B	2.48	2.76	1.92	2.53	2.78	1.51	2.61	2.08	2.77	0.503	0.547	32.27%	2.33
	mplug-Ow3-2B	3.45	2.91	2.36	2.71	2.42	2.09	2.99	2.41	2.40	0.522	0.610	42.14%	2.66
	mplug-Ow3-7B	3.92	3.89	2.59	3.34	2.83	2.91	3.95	2.78	2.81	0.702	0.672	62.49%	3.25
	InternVL2.5-8B	3.90	3.82	3.14	3.16	3.72	3.23	3.78	3.18	3.69	0.704	0.628	71.45%	3.24
	InternVL2.5-26B	**4.10**	3.92	3.38	3.27	4.08	3.05	3.98	3.34	3.96	**0.751**	0.644	67.34%	3.63
	Avg.	2.98	3.22	2.45	2.40	2.81	2.09	3.19	2.53	2.83	0.540	0.475	44.05%	2.65
Closed source	GPT-4o	**4.35**	**4.38**	**4.28**	**4.12**	**4.31**	3.06	**4.26**	**4.16**	**4.18**	**0.905**	**0.826**	68.28%	**4.08**
	Claude-3.5-sonnet	4.12	4.04	3.98	3.89	3.88	3.21	4.09	4.07	3.92	0.823	0.786	74.28%	3.86
	Gemini-1.5 Pro	3.96	3.90	3.75	3.94	4.16	3.11	3.92	3.92	4.10	0.716	0.794	73.03%	3.80
	DeepSeek-V3	3.92	3.96	3.92	3.98	4.03	**3.28**	4.01	3.91	3.99	0.702	0.798	**74.93%**	3.84
	Avg.	4.09	4.07	3.98	3.99	4.10	3.17	4.07	4.01	4.04	0.787	0.801	72.63%	3.90

Table 3: Comparison of Performance for 22 MLLMs on MMRC. $*$ indicates that the same task has been re-evaluated manually. **Bold** and underline denote the best and second-best results, respectively.

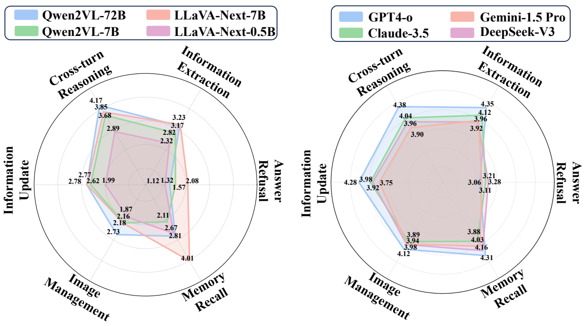

Figure 6: Radar chart of capabilities for models with noticeable task-specific imbalances.

4 Experiment and Analysis

4.1 Evaluation Matrix

Since questions in MMRC are open-ended, directly evaluating accuracy is infeasible. To address this, we develop a comprehensive evaluation framework that integrates GPT-based scoring, human assessment, and objective precision metrics. Specifically, GPT evaluates all six abilities (Section 3.1), while human evaluators conduct a second-round review for CR, IU, and MR, which require more in-depth judgment. Moreover, both GPT and human evaluators employ a scoring scale ranging from 0 to 5, with prompts and evaluation criteria detailed in the Appendix J. In contrast, for IE, IM, and AR, we employ objective precision metrics, including Extraction Precision (EP), Image Management Precision (IMP), and Refuse Precision (RP), to provide an intuitive assessment of model performance by measuring the proportion of correct responses.

EP as an extension of IE, measures the precision in extracting items:

$$\mathrm{EP} = \frac{|\mathrm{Model}_{\mathrm{items}} \cap \mathrm{Label}_{\mathrm{items}}|}{|\mathrm{Model}_{\mathrm{items}}|},$$

where $\mathrm{Model}_{\mathrm{items}}$ denotes the set of items generated by the model, and $\mathrm{Label}_{\mathrm{items}}$ denotes the set of ground-truth items.

IMP as an extension of IM, measures the precision in managing and retrieving images:

$$\mathrm{IMP} = \frac{|\mathrm{Image}_{\mathrm{hit}}|}{|\mathrm{Image}_{\mathrm{hit}} \cup \mathrm{Image}_{\mathrm{miss}}|},$$

where $\mathrm{Image}_{\mathrm{hit}}$ denotes the images correctly retrieved by the model, and $\mathrm{Image}_{\mathrm{miss}}$ denotes the images that are part of the correct answer but were not retrieved by the model.

RP as an extension of AR, measures the precision in refusing to answer unknown questions:

$$\mathrm{RP} = \frac{\sum_{i=1}^{N} \mathbb{1}(D_i)}{\sum_{i=1}^{N} \mathbb{1}(E_i)},$$

where D_i and E_i denote the model's refusal to answer and the ground-truth refusal for the i-th question, respectively. The indicator function $\mathbb{1}(\cdot)$

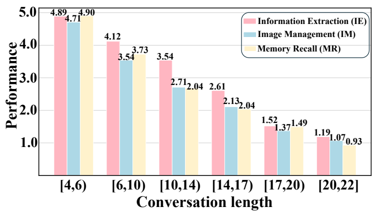

Figure 7: The impact of conversation length on memory performance.

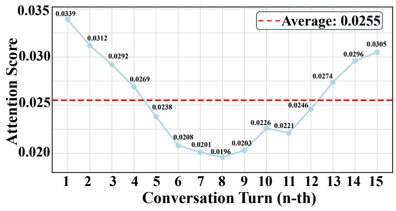

Figure 8: Different turns' attention score in the conversation.

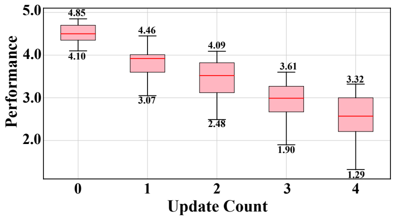

Figure 9: Impact of update frequency on model performance.

returns 1 if the response aligns with the expected refusal behavior and 0 otherwise. N denotes the total number of questions.

4.2 Main Results

Table 3 presents the performance of 22 open-source and closed-source MLLMs in real-world dialogue scenarios. Based on the evaluation results, we identify three key findings:

(1) Challenges of Reality: LLaVA-1.5 performs poorly in OEC, while powerful GPT-4o still falls short of human-level performance in real-world scenarios, highlighting the complexity and difficulty of practical, user-driven conversations in MMRC.

(2) Reliability of Evaluation: The similarity between GPT's scores and those of human annotators reaches 93%. Furthermore, GPT's scores closely correspond to the Objective Precision Metrics, where higher GPT scores consistently reflect stronger performance. This consistency indicates the effectiveness of GPT guided by our evaluation prompts, providing a foundation for using GPT scores in subsequent experimental analysis.

(3) Task-Specific Imbalance: The performance of most models is imbalanced, exhibiting distinct strengths and weaknesses. As illustrated in Fig. 6, the LLaVA-Next family demonstrates strong memory recall ability but weaker image management ability. Similarly, the Qwen2VL family excels in cross-turn reasoning yet exhibits a relatively weak answer refusal ability. Notably, closed-source models exhibit a more pronounced performance imbalance, consistently struggling with answer refusal. This disparity in MLLMs stems from variations in training datasets and strategies. Specifically, different organizations prioritize fine-tuning for certain tasks, leading to enhanced performance in those areas while resulting in weaker performance on tasks with less targeted training.

4.3 Error Analysis

We conduct an in-depth analysis of failures of the model and identify four common error patterns:

(1) Long-term memory degradation: As the conversation progresses, the model's memory of previous dialogue content becomes increasingly vague, a phenomenon known as memory degradation (Zhong et al., 2024). Moreover, the severity of memory degradation increases as the conversation lengthens. As illustrated in Fig. 7, memory-related abilities (*i.e.,* IE, IM, MR) decline significantly in extended conversations. Observations reveal that memory degradation is more severe in the middle of a conversation than at the beginning or end, challenging the assumption that earlier memories degrade more rapidly. To further investigate, we visualize the model's attention patterns when addressing memory-related questions (Appendix K). As shown in Fig. 8, attention to the middle part of the conversation is markedly lower than to the beginning and end, mirroring the observed degradation pattern. We hypothesize that this unbalanced attention distribution contributes to the heightened memory degradation in mid-conversation.

Furthermore, we analyze the error types associated with long-term memory degradation, with results shown in Fig. 10: *(i) Memory omission:* the model fails to retain certain details from the conversation. *(ii) Complete forgetting:* a more severe form of memory degradation than omission, where the model entirely fails to recall specific events from the conversation. *(iii) Memory confusion:* the model incorrectly merges relevant information with unrelated content, leading to distorted recollection.

(2) Inadequacies in updating factual knowledge: MLLMs often struggle to track changes in user information and factual knowledge during conversations, resulting in failures in updating information. As illustrated in Fig. 9, frequent changes in factual knowledge make it difficult to adapt to rapidly evolving information and result in a decline in update performance. To further investigate this issue, we analyze the types of errors associated with updating information, with their distribution shown in Fig. 11: *(i) Failure to recognize updates:*

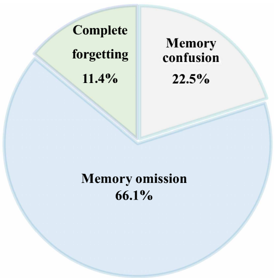

Figure 10: Information extraction error statistics.

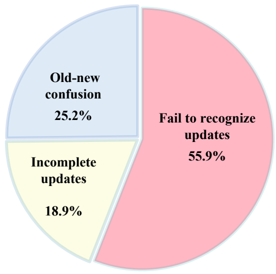

Figure 11: Information update error statistics.

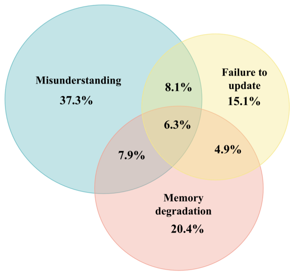

Figure 12: Statistics of error types in reasoning.

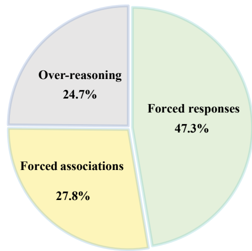

Figure 13: Statistics of error types in answer refusal.

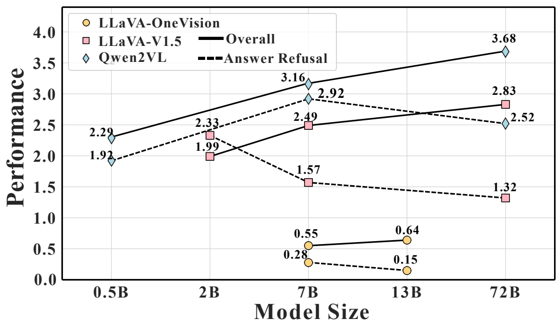

Figure 14: Overall and answer refusal performance across different model sizes within the same family.

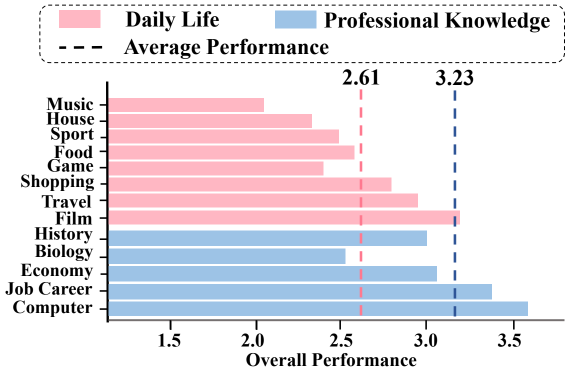

Figure 15: MLLMs' performance in different topics.

occurs when the model fails to detect that certain factual knowledge requires updating, instead treating it as static information. *(ii) Incomplete updates:* arises when the model acknowledges the need for an update but fails to incorporate the most recent information due to frequent changes. *(iii) Old-new contradiction:* happens when the model incorrectly merges outdated facts with new ones, leading to an inaccurate representation of the latest information.

(3) Accumulated assumption of error propagation: During reasoning, the model processes information sequentially from earlier to later dialogue turns to comprehend the dialogue and integrate relevant details, forming a reasoning chain to answer complex questions. However, as illustrated in Fig. 12, the formation of this reasoning chain is often hindered by three key issues, leading to flawed reasoning and incorrect answers: *(i) Misunderstanding:* the model fails to correctly understand the dialogue content, resulting in distortions within the reasoning chain and incorrect assumptions that ultimately lead to erroneous conclusions. *(ii) Memory degradation:* the model forgets prior conversation information, disrupting the reasoning chain. This lack of essential context weakens the model's assumptions, preventing accurate reasoning. *(iii) Failure to update:* the model continues reasoning based on outdated knowledge, leading to incorrect answers. Notably, these errors can occur simultaneously. As the conversation progresses, their accumulation exacerbates incorrect assumptions, causing errors to propagate more severely throughout the dialogue.

(4) Reluctance to "say no": The model provides unreliable answers when the context is insufficient, potentially misleading users. To understand the underlying causes of this issue, we conduct an analysis and categorize them into three key types, as illustrated in Fig. 13. *(i) Forced Responses:* the model recognizes that the question is unrelated to the given context and does not utilize any conversational context for its response, yet it fails to refuse to answer. *(ii) Over-reasoning:* the model excessively analyzes the question, attempting to infer non-existent details from the available context. *(iii) Force associations:* the model artificially links an irrelevant element in the question with existing conversation details, generating an answer based on this false connection. Furthermore, as shown in Fig. 14, within the same model family, larger models demonstrate superior overall performance. However, their ability to refuse inappropriate answers declines. A comparative analysis with varying sizes reveals that while larger models exhibit stronger logical reasoning capabilities, they are more prone to over-reasoning and forced associations compared to smaller models.

4.4 Further Discussion

Domain Bias: To further investigate the model's performance across different domains, we evaluate its overall effectiveness in each domain, as shown in Fig. 15. Our analysis reveals that the model performs significantly better in professional

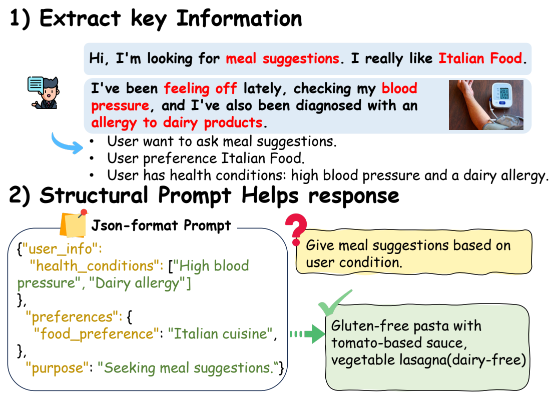

Figure 16: **Illustration of NOTE-TAKING method.**

Models	IE	IU	MR	AR
LLaVA-1.5-7B	0.91	0.31	0.22	0.28
+ NOTE-TAKING	**2.57** (+1.66)	**2.06** (+1.75)	**2.36** (+2.14)	**0.46** (+0.18)
MiniCPM-8B	4.08	2.98	3.65	3.78
+ NOTE-TAKING	**4.23** (+0.15)	**3.76** (+0.78)	**4.02** (+0.37)	**3.92** (+0.14)
QwenVL-2B	2.16	1.41	1.93	2.33
+ NOTE-TAKING	**3.71** (+1.55)	**2.43** (+0.82)	**3.40** (+1.47)	**2.84** (+0.51)
LLaVA-Next-0.5B	2.32	1.99	2.67	1.12
+ NOTE-TAKING	**3.84** (+1.52)	**3.04** (+1.05)	**3.88** (+1.21)	**2.03** (+0.91)
LLaVA-OneVision-72B	4.06	4.01	4.17	2.52
+ NOTE-TAKING	**4.28** (+0.22)	**4.31** (+0.30)	**4.38** (+0.21)	**3.46** (+0.94)
GPT-4o	4.35	4.28	4.31	3.06
+ NOTE-TAKING	**4.51** (+0.16)	**4.62** (+0.34)	**4.73** (+0.42)	**3.51** (+0.45)

Table 4: Performance of NOTE-TAKING across four conversational core abilities in MMRC.

knowledge conversations than in daily conversations (professional: 3.23, daily: 2.61). We hypothesize that this disparity stems from variations in the instruction-based training across models. Specifically, models demonstrating stronger performance in professional knowledge conversations benefit from a larger proportion of instruction-based data tailored for knowledge-based tasks. Thus, to improve MLLMs' conversational abilities in OEC, it is essential to incorporate more daily conversation data during supervised fine-tuning.

Modalities Preference: To explore the model's preference for different information modalities, we modify 100 conversations by replacing parts of the original text content with equivalent image inputs. For instance, the text-based statement: "I visited the Eiffel Tower." is converted into "I visited this place." followed by an image of the Eiffel Tower. The rest of the dialogue remains unchanged for evaluation. Our findings indicate that MLLMs exhibit a strong preference for text-based information, with overall scores for text-based dialogues being 26.3% higher than their image-based counterparts. Furthermore, models exhibit fewer memory degradation errors in text-based conversations, as memory-related capabilities such as IE, IM, and MR show a 34.6% improvement. We attribute this to two main factors: *(i)* images often require more tokens to convey the same meaning as text, significantly increasing context length. *(ii)* The model's training data is imbalanced, with text data vastly exceeding image data, leading to stronger proficiency in processing textual information.

5 NOTE-TAKING as Improved Baseline

In this section, we introduce an initial step toward enhancing the core capabilities of MLLMs in OEC. The primary failure of MLLMs is their inability to accurately retrieve detailed information, update knowledge, and recognize missing information. To mitigate this, we propose a NOTE-TAKING framework, which guides MLLMs in extracting key dialogue information and recording it in accessible json-format notes. These structured notes serve as external memory, improving response accuracy and context understanding.

As illustrated in Fig. 16, the NOTE-TAKING effectively simulates how humans take notes during long and complex conversations, facilitating the retention of key details and maintaining focus. As shown in Table 4, the improvement is observed in the long-term memory ability, with MR increasing by an average of +0.97. Furthermore, the note-taking mechanism enhances the model's information extraction capability by +0.88, and improves information update by +0.84. Moreover, the structured clarity provided by the notes allows the model to concentrate more effectively on relevant details within the conversation, resulting in a +0.52 improvement in answer refusal.

6 Conclusions

In this paper, we introduce MMRC, the first multi-image open-ended conversation benchmark to evaluate the six conversation abilities of MLLMs. Our comprehensive analysis identifies four common failure patterns: long-term memory degradation, inadequate updating of factual knowledge, accumulated assumption of error propagation, and reluctance to "say no." To mitigate these, we propose the NOTE-TAKING strategy, which stores key user preferences and facts by using structured prompts.

Limitations: We clarify the limitations: *(i)* While MMRC covers multiple domains, it may not encompass all real-world dialogue types (*e.g.,* population distribution and languages) and requires further exploration. *(ii)* Although the NOTE-TAKING improves model performance, the note generation process can be computationally intensive.

References

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. Neural Computing and Applications, 36(12):6325–6351.

Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical considerations for responsible data curation. Advances in Neural Information Processing Systems, 36:55320–55360.

Mela Aziza. 2021. A teacher questioning activity: The use of oral open-ended questions in mathematics classroom. Qualitative Research in Education, 10(1):31–61.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. arXiv preprint arXiv:2402.14762.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966.

John Barnden. 2014. Open-ended elaborations in creative metaphor. In Computational creativity research: Towards creative machines, pages 217–242. Springer.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. arXiv preprint arXiv:2403.17297.

Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. ACM Transactions on Computer-Human Interaction, 29(2):1–38.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In Forty-first International Conference on Machine Learning.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–12.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 958–979.

SMALL VISUAL DE. Mllms know where to look: Training-free perception of small visual de-tails with multimodal llms.

Amandine Decker. 2022. Topic Shifts: Preserving Comprehension in Conversation. Ph.D. thesis, Université de lorraine; IDMC.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Hillary Anger Elfenbein, Petri Laukka, Jean Althoff, Wanda Chui, Frederick K Iraki, Thomas Rockstuhl, and Nutankumar S Thingujam. 2022. What do we hear in the voice? an open-ended judgment study of emotional speech prosody. Personality and Social Psychology Bulletin, 48(7):1087–1104.

Hao Fei, Xiangtai Li, Haotian Liu, Fuxiao Liu, Zhuosheng Zhang, Hanwang Zhang, and Shuicheng Yan. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning and beyond. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 11289–11291.

Charles Fernyhough. 1996. The dialogic mind: A dialogic approach to the higher mental functions. New ideas in Psychology, 14(1):47–62.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. arXiv preprint arXiv:2411.15296.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. arXiv e-prints, pages arXiv–2407.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. Authorea Preprints.

Toshish Jawale, Chaitanya Animesh, Sekhar Vallath, Kartik Talamadupula, and Larry Heck. 2024. Are human conversations special? a large language model perspective. arXiv preprint arXiv:2403.05045.

Katarzyna Kapusta, Lucas Mattioli, Boussad Addad, and Mohammed Lansari. 2024. Protecting ownership rights of ml models using watermarking in the light of adversarial attacks. AI and Ethics, 4(1):95–103.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. Advances in Neural Information Processing Systems, 37:28798–28827.

Suyoun Kim and Florian Metze. 2018. Dialog-context aware end-to-end speech recognition. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 434–440. IEEE.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2024. Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys.

Paulette Lacroix. 2019. Big data privacy and ethical challenges. Big Data, Big Challenges: A Healthcare Perspective: Background, Issues, Solutions and Research Directions, pages 101–111.

Patrick Langer, Stephan Altmüller, Elgar Fleisch, and Filipe Barata. 2024. Claid: Closing the loop on ai & data collection—a cross-platform transparent computing middleware framework for smart edge-cloud and digital biomarker applications. Future Generation Computer Systems, 159:505–521.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. arXiv preprint arXiv:2411.16594.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In The Twelfth International Conference on Learning Representations.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. 2024c. Groundinggpt: Language enhanced multi-modal grounding model. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6657–6678.

Anthony J Liddicoat. 2021. An introduction to conversation analysis. Bloomsbury Publishing.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26689–26699.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. 2024b. Mibench: Evaluating multimodal large language models over multiple images. arXiv preprint arXiv:2407.15272.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. 2024c. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. arXiv preprint arXiv:2403.20194.

Yexin Liu, Zhengyang Liang, Yueze Wang, Muyang He, Jian Li, and Bo Zhao. 2024d. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. arXiv preprint arXiv:2406.10638.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024e. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. arXiv preprint arXiv:2406.11833.

Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. 2024. Consent in crisis: The rapid decline of the ai data commons. In NEURIPS.

Rachel Ma, Jingyi Qu, Andreea Bobu, and Dylan Hadfield-Menell. 2024. Goal inference from open-ended dialog. arXiv preprint arXiv:2410.13957.

Alex McKeown, Miranda Mourby, Paul Harrison, Sophie Walker, Mark Sheehan, and Ilina Singh. 2021. Ethical issues in consent for the reuse of data in health data platforms. Science and Engineering Ethics, 27:1–21.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2023. Dialogbench: Evaluating llms as human-like dialogue systems. arXiv preprint arXiv:2311.01677.

ST Padmapriya and Sudhaman Parthasarathy. 2024. Ethical data collection for medical image analysis: A structured approach. Asian Bioethics Review, 16(1):95–108.

Dragan Pamucar, Vladimir Simic, Ömer Faruk Görçün, and Hande Küçükönder. 2024. Selection of the best big data platform using cobrac-artasi methodology with adaptive standardized intervals. Expert Systems with Applications, 239:122312.

Yohanes Yohanie Fridelin Panduman, Nobuo Funabiki, Evianita Dewi Fajrianti, Shihao Fang, and Sritrusta Sukaridhoto. 2024. A survey of ai techniques in iot applications with use case investigations in the smart environmental monitoring and analytics in real-time iot platform. Information, 15(3):153.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. 2024. Mitigating dialogue hallucination for large multi-modal models via adversarial instruction tuning. arXiv preprint arXiv:2403.10492.

Irene V Pasquetto, Zoë Cullen, Andrea Thomer, and Morgan Wofford. 2024. What is research data "misuse"? and how can it be prevented or mitigated? Journal of the Association for Information Science and Technology, 75(12):1413–1429.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).

Stefan Pfenninger, Joseph DeCarolis, Lion Hirth, Sylvain Quoilin, and Iain Staffell. 2017. The importance of open data and software: Is energy research lagging behind? Energy Policy, 101:211–215.

Guangyuan Piao. 2021. Scholarly text classification with sentence bert and entity embeddings. In Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25, pages 79–87. Springer.

Eduardo Pina, José Ramos, Henrique Jorge, Paulo Váz, José Silva, Cristina Wanzeller, Maryam Abbasi, and Pedro Martins. 2024. Data privacy and ethical considerations in database management. Journal of Cybersecurity and Privacy, 4(3):494–517.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and S Yu Philip. 2025. A survey of multilingual large language models. Patterns, 6(1).

Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1–20.

Efrat Shimron, Jonathan I Tamir, Ke Wang, and Michael Lustig. 2022. Implicit data crimes: Machine learning bias arising from misuse of public data. Proceedings of the National Academy of Sciences, 119(13):e2117203119.

Chen Sun, Valerie J Shute, Angela EB Stewart, Quinton Beck-White, Caroline R Reinhardt, Guojing Zhou, Nicholas Duran, and Sidney K D'Mello. 2022. The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. Computers in Human Behavior, 128:107120.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9729–9750.

Donald E Super. 1982. The relative importance of work: Models and measures for meaningful data. The Counseling Psychologist, 10(4):95–103.

Zhaoxuan Tan, Zinan Zeng, Qingkai Zeng, Zhenyu Wu, Zheyuan Liu, Fengran Mo, and Meng Jiang. 2025. Can large language models understand preferences in personalized recommendation? arXiv preprint arXiv:2501.13391.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Chuang Wang, Xiaoyan Li, Wei Sun, Jingjing An, and Shufang Gao. 2024a. Occupant behavior, thermal environment, and appliance electricity use of a single-family apartment in china. Scientific Data, 11(1):65.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

Jingyu Wei, Yi Su, Kele Xu, Lingbin Zeng, Bo Liu, and Huaimin Wang. 2024. Demonstrative instruction following in multimodal llms via integrating low-rank adaptation with ensemble learning. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 11435–11441.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. arXiv preprint arXiv:2410.10813.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. 2025. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In European Conference on Computer Vision, pages 164–182. Springer.

Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. ACM Transactions on Computer-Human Interaction (TOCHI), 27(3):1–37.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, et al. 2024a. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In Findings of the Association for Computational Linguistics ACL 2024, pages 4977–4997.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. arXiv preprint arXiv:2312.09085.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2024b. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. arXiv preprint arXiv:2410.12332.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, and et al . 2024. Qwen2 technical report. Preprint, arXiv:2407.10671.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. CoRR.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In Neurips Safe Generative AI Workshop 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13040–13051.

Kazumi Yoshimura, Dominique Chen, and Olaf Witkowski. 2024. Synlogue with aizuchi-bot: Investigating the co-adaptive and open-ended interaction paradigm. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1–21.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2024. Contextual object detection with multimodal large language models. International Journal of Computer Vision, pages 1–19.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986.

Shuili Zhang, Hongzhang Mu, and Tingwen Liu. 2024. Improving accuracy and generalizability via multi-modal large language models collaboration. In 2024 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.

Yang Zhang, Yi Shao, Wenbo Wan, Jing Li, and Jiande Sun. 2022. Clip pre-trained models for cross-modal retrieval in newsimages 2022. In MediaEval.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. Science China Technological Sciences, 63(10):2011–2027.

Yin Jiang Zhao, Yan Ling Li, and Min Lin. 2019. A review of the research on dialogue management of task-oriented systems. In Journal of physics: conference series, volume 1267, page 012025. IOP Publishing.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. arXiv preprint arXiv:2309.11998.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19724–19731.

A Open-ended Conversations

Open-ended conversation is a flexible and unconstrained form of conversations (Fernyhough, 1996; Barnden, 2014), allowing users to engage freely without predefined limits (Xiao et al., 2020; Yoshimura et al., 2024). In these conversations, the user has full control, enabling them to express thoughts and emotions openly (Elfenbein et al., 2022; Seo et al., 2024), as well as delve into topics in greater depth to uncover insights or solutions (Aziza, 2021; Sun et al., 2022). This type of conversation is the most common mode of interaction for general users in MLLM chat systems (Fu et al., 2024; Liu et al., 2024b).

B Ethics Statements

The widespread availability of data and powerful analytical tools play a pivotal role in research (Pfenninger et al., 2017; Super, 1982), but also come with the risk of misuse (Pasquetto et al., 2024; Shimron et al., 2022). Therefore, ethical oversight is crucial, it involves issues of privacy (Pina et al., 2024; Lacroix, 2019), ownership (Andrews et al., 2023; Kapusta et al., 2024), consent (McKeown et al., 2021; Longpre et al., 2024), and purpose of use (Paullada et al., 2021; Padmapriya and Parthasarathy, 2024). Based on the definitions and related issues concerning dataset ethics, we make the following statement about MMRC:

The data collection for our study is conducted with the informed consent of all participants, ensuring their privacy and autonomy are fully protected. All participants are fully aware of and voluntarily engage in the annotation process. We implement a rigorous review mechanism to ensure the data is free from personally identifiable information, offensive material, and violent content. However, given the limitations of manual inspection, some residual information may still remain, and completely eliminating such content remains a challenging task. Furthermore, since the data originates from real conversations, it may contain a small amount of inadvertently misleading information, which could impact the model's performance on benchmark tests. We release this data exclusively for research purposes, allowing researchers to explore the performance of multimodal large language models (MLLMs) in real-world dialogue contexts. However, researchers must approach this dataset with utmost caution and ethical consideration. Our goal is to contribute to the accumulation of knowledge while ensuring that our research findings are applied ethically. In the future, we will continue to release updated versions of the dataset, expanding both its volume and comprehensiveness, while further filtering out offensive content and misinformation.

C *DialogFlow*

Collecting data through online platforms is an efficient method to gather large volumes of valuable data (Panduman et al., 2024; Wang et al., 2024a). It offers a cost-effective and scalable solution for data collection, adapting to varying needs over time (Pamucar et al., 2024; Langer et al., 2024). Therefore, we establish *DialogFlow* to collect user conversation data.

Due to GPU source limitations, *DialogFlow* is not open to the public and is only undergoing internal testing within the campus. The interface of *DialogFlow* is shown in Fig.17. Users can add image inputs by clicking the camera icon at the bottom right corner and switch between different models by changing the tabs at the top. The platform is free to use since its launch on April 6, 2024. We use dozens of A100 GPUs to host our website and run open-source models for user conversations, while closed-source models interact via APIs. All the models available on the platform are listed in Table 5. *DialogFlow* has a total of 354 users, with ages ranging from 18 to 47. The majority of them are university students and staff. Users are required to accept the Terms of Use to consent to the public release of their conversation data.

D Examples of MMRC

We list 5 samples of MMRC, they are conversations about travel (Fig. 20), dancing (Fig. 21), physics (Fig. 22), water parks (Fig. 23), and dessert making (Fig. 24). The domains of these conversations are comprehensive, natural, and closely aligned with real-world usage scenarios. Additionally, the evaluation questions for the conversations are manually annotated and checked by our team. All of this demonstrates the high quality and real-world relevance of our data.

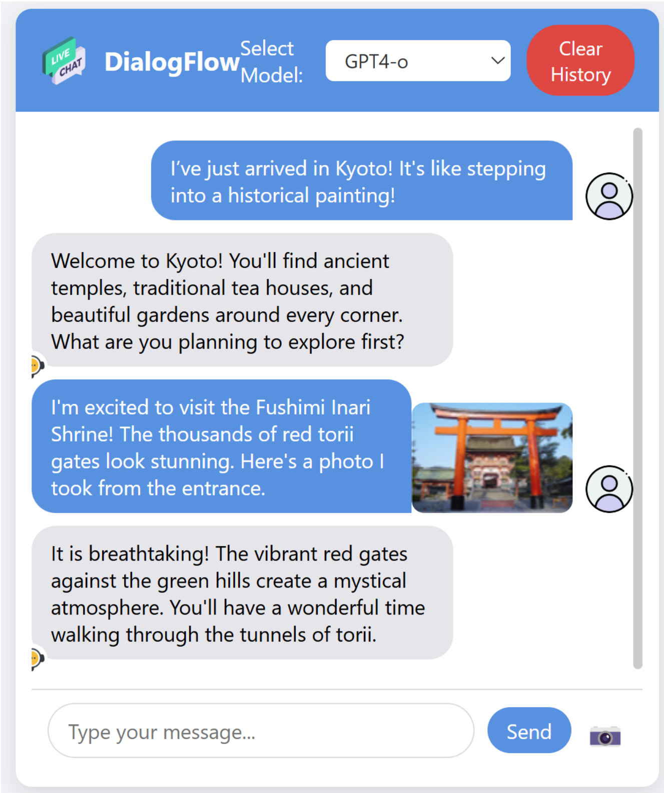

Figure 17: The page of *DialogFlow*

Model	Size	Vision Encoder	LLM
LLaVA-V1.5 (Liu et al., 2023)	7B,13B	CLIP ViT-L/14 (Zhang et al., 2022)	Vicuna-v1.5 (Zheng et al., 2023b)
MiniCPM (Yao et al., 2024)	8B	SigLIP-SoViT-400m/14 (Zhai et al., 2023)	Llama3-Instruct (Dubey et al., 2024)
LLaVA-Next (Li et al., 2024b)	0.5B,7B	SigLIP-400M (Zhai et al., 2023)	Qwen1.5 (Bai et al., 2023)
Qwen2VL (Wang et al., 2024b)	2B,7B,72B	ViT trained from scratch (Alexey, 2020)	Qwen2 (Yang et al., 2024)
LLaVA-OneVision (Li et al., 2024b)	0.5B,7B	SigLIP-400M (Zhai et al., 2023)	Qwen2 (Yang et al., 2024)
VILA1.5 (Lin et al., 2024)	3B,8B	SigLIP-400M (Zhai et al., 2023)	Llama3 (Grattafiori et al., 2024)
mplug-Ow3 (Ye et al., 2024)	1B,2B,7B	SigLIP-400M (Zhai et al., 2023)	Qwen2 (Yang et al., 2024)
InternVL2.5 (Ye et al., 2024)	8B	InternViT-300M-448px-V2.5 (Chen et al., 2024)	internlm2.5-7b-chat (Cai et al., 2024)
InternVL2.5 (Ye et al., 2024)	26B	InternViT-6B-448px-V1-5 (Chen et al., 2024)	internlm2.5-20b-chat (Cai et al., 2024)
GPT-4o (Islam and Moushi, 2024)	300B	-	-
Claude-3.5-sonnet (Yao et al., 2024)	175B	-	-
Gemini-1.5 Pro (Team et al., 2024)	175B	-	-
DeepSeek-V3 (Liu et al., 2024a)	671B	-	-

Table 5: MLLMs in *DialogFlow*.

E Topic Classification Network

The 14 predefined categories are carefully chosen to ensure comprehensive coverage of real-world conversational topics, and their selection is grounded in the characteristics of our dataset and relevant studies on human conversations (Kim and Metze, 2018; Zhao et al., 2019; Zhang et al., 2020; Algherairy and Ahmed, 2024).

To train the topic classification network, we manually annotate a portion of the data, label the date with 14 predefined categories, each containing 50 dialogue samples. We split the data into 90% for training and 10% for testing. For the classification, we use the all-mpnet-base-v2 model from SentenceTransformers (Piao, 2021) to generate text embeddings. These embeddings are passed through a two-layer MLP for classification. The model ultimately achieve over 95% accuracy on the test set.

F Statics of conversation domains in MMRC

Although we have categorized the existing conversation data into 14 categories for convenience in statistical analysis, the actual diversity of the data far exceeds this classification. If we increase the

	Category	Specific Domains
Daily Life	Travel	Destination recommendations, travel preparations, accommodation issues, travel budget and planning, transportation methods, scenic spots and activities, culture and language, travel safety, backpacking, luxury travel, self-driving tours, group tours, travel photography, food during travel, travel souvenirs, long-haul flight tips, health during travel, seasonal travel, travel insurance, socializing during travel, eco-friendly travel, legal issues during travel, recommended travel apps, travel blogs, high-end travel, travel photography tips, entertainment activities during travel, unforgettable travel experiences, travel shopping list, domestic travel, international travel, common sense of travel safety, holiday travel, food tours, island vacations.
	Music	Music styles, music history, music theory, music composition, music production, music appreciation, music education, music technology, instrument performance, music festivals, singers and bands, music films, emotional expression in music, music and culture, music and society, concerts and music festivals, music charts, music copyright, music innovation, music streaming, music criticism, melody and harmony in music, music industry, music commercialization, music social media, music collaboration, music creativity and inspiration, music rhythm, lyrics in music, psychology of music, pop music, classical music, rock music, jazz, hip-hop music, electronic music, world music, music and film scores, music marketing, music performance skills, dance in music, music and emotional therapy, music apps, music chart analysis, cultural significance of music, live music recording, music reviews and recommendations, trends in the music industry.
	Game	Game types, game design, game development, game engines, game programming, game art and graphics, game sound effects and music, game narrative, game plot, game character design, game AI and artificial intelligence, game testing and quality control, game release, game market, game industry trends, game hardware and devices, game virtual reality, game augmented reality, game controllers, video game culture, game player communities, e-sports, game streaming, game reviews and evaluations, game achievements and rewards, in-game purchases and microtransactions, single-player games, online games, mobile games, console games, PC games, adventure games, role-playing games, sports games, action games, shooting games, simulation games, strategy games, puzzle games, music games, indie games, game innovation, game marketing, game development process, game creativity and inspiration, game props and items, emotional interaction in games, game balance and difficulty design, game maps and world design, social interaction in games, achievement systems in games, game education and learning, economic systems in games, moral choices in games, time management in games, exploration and discovery in games, choices and consequences in game plots, multiple endings in games, openness of game worlds
	Sport	Types of sports, sports history, sports culture, sports competition, sports training, sports psychology, sports injuries, sports nutrition, sports athletes, sports competitions, sports referees, sports rules, sports management, sports marketing, sports events, sports technology, sports equipment, live sports broadcasting, sports media, sports stars, trends in the sports industry, commercialization of sports, sports sponsorship, sports social media, sports achievements and records, social impact of sports, sports education, sports and health, sports and physical fitness, extreme sports, indoor sports, outdoor sports, team sports, individual sports, collective events, competition strategies, organization of sports events, sports training methods, youth sports, sports psychological adjustment, sports meets and the Olympic Games, ball games, track and field, swimming and water sports, gymnastics, martial arts and combat sports, cycling and equestrian sports, motor racing, winter sports, sports organizations and federations, sports law and contracts, sports clubs and teams, sports training equipment and apparatus, athletes' careers
	House	Home design, home decoration, interior decoration, furniture selection and matching, home style, family organization and storage, home lighting, home layout, family space planning, kitchen design, bedroom design, living room design, bathroom design, children's room design, home office design, home greening and plants, residential architecture, home decoration materials, home appliance selection, smart home, home security, home cleaning, home maintenance and repair, home comfort, home accessories, home color matching, home art, home audio system, home theater design, home dining room design, home comfort and functionality, fluidity between family and space, family environment and health, eco-friendly home, small apartment home design, family storage solutions, home feng shui, DIY home projects, home budget and expenses, home renovation and refurbishment, family entertainment area, courtyard design, outdoor home, family safety and protection
	Food	Food types, ingredient selection, food nutrition, recipes and cooking techniques, healthy eating, food culture, international cuisine, local specialties, food safety, drink selection, vegetables and fruits, restaurant recommendations, the catering industry, catering services, food pairing, desserts and baking, traditional cuisine, vegetarianism and veganism, high-protein diets, low-carb diets, sugar-free foods, low-fat diets, low-salt diets, fast food and convenience foods, organic foods, home-cooked meals, seasonal ingredients, food processing, food packaging, dietary restrictions, diet and health, food allergies, food aesthetics, storage and preservation of ingredients, cooking tools and equipment, food labels, cross-cultural cuisine, street food, homemade meals, food cultural heritage, food waste, sustainable eating, food innovation, catering brands, food and travel, cooking skills, food festivals and events, food ingredients, fine dining, fast food and takeout, imported foods, food history and traditions, dessert design, food festivals and celebrations, food poisoning and hygiene, food and socializing, food science, food quality control, traditional eating habits, beverages and wines, table manners, healthy eating habits, low-calorie recipes, quick cooking, food innovation and trends, food consumption and trends, family meals, festival foods, picnics and outdoor dining, health-preserving diets
	Shopping	Shopping categories, shopping tips, shopping malls, online shopping, shopping websites, shopping experience, fashion shopping, shopping budget, discounts and promotions, shopping lists, price comparison for shopping, overseas shopping, second-hand shopping, shopping fashion trends, shopping habits, eco-friendly choices in shopping, shopping payment methods, shopping coupons, return and exchange policies for shopping, luxury shopping, shopping traps and fraud, shopping psychology, shopping brands, shopping festivals and big sales, shopping social platforms, shopping and culture, emotional spending while shopping, shopping baskets and carts, logistics and delivery in shopping, shopping packaging, shopping convenience and efficiency, time management in shopping, payment security in shopping, personalized shopping recommendations, gift and giveaway shopping, health and beauty product shopping, home and living goods shopping, food shopping, electronics shopping, sports and outdoor products shopping, fashion and apparel shopping, makeup and skincare product shopping, jewelry and accessories shopping, social interactions in shopping, shopping sprees, brick-and-mortar store shopping, online and offline shopping integration, personalized custom shopping, international brand shopping.
Professional Knowledge	History	Historical periods, ancient civilizations, medieval history, modern history, 20th-century history, world wars, historical figures, historical events, historical cultures, historical geography, the rise and fall of nations and ethnicities, economic history, political history, wars and conflicts, social change, history of technological development, history of religion, history of art and culture, history of ethnicities and immigration, history of law, environmental history, historical archives and documents, history of international relations, historical theories and methods, major events in world history, the rise and fall of great powers, the impact and lessons of history, war and peace, major historical reforms, the Renaissance, the Age of Enlightenment, the Industrial Revolution, the Cultural Revolution, the Great Depression, the legacy of ancient civilizations, the post-World War II international order, famous historical explorations and voyages, ancient military strategies and tactics, literature and history, historical scientific discoveries, major historical disasters, technological innovations in history, national independence movements, the Cold War and global politics.
	Economy	Macroeconomics, microeconomics, economic development, international economics, economic growth, monetary policy, fiscal policy, supply and demand, market structures, price mechanisms, trade and globalization, economic crises and recessions, economic cycles, capitalism and socialism, economic systems and institutions, national economies and government policies, consumer behavior, production and labor markets, labor economics, resource allocation and efficiency, public economics, economic inequality, international trade and investment, foreign exchange markets and exchange rates, economic reforms, taxation and fiscal management, money and banking systems, businesses and market competition, industrial structures, inflation and deflation, unemployment and employment, economic models and theories, development economics, environmental economics, sustainable development, energy economics, market failures and government intervention, financial crises, debt crises, economic policy analysis, consumerism, wealth distribution and social welfare, wealth creation and investment, capital markets, stock and bond markets, financial analysis and management, business strategy and innovation, business management and operating environments.
	Biology	Cell biology, molecular biology, genetics, evolution theory, ecology, ecosystems, species and species diversity, biological classification, microbiology, DNA and RNA, human biology.
	Computer	Computer hardware, computer software, operating systems, principles of computer organization, computer networks, data structures and algorithms, programming languages, artificial intelligence, machine learning, deep learning, big data, cloud computing, distributed computing, parallel computing, high-performance computing, computer vision, natural language processing, computer security, network security, information security, data encryption, blockchain technology, computer architecture, computer storage, computer graphics, virtual reality and augmented reality, computer science theory, computer history, computer ethics, computer applications, computer simulation, software engineering, computer programming, web development, mobile application development, database management, computer viruses and protection, computer communication protocols.
	Society	Social structure, social stratification, social classes, social change, social culture, social norms and values, social enterprises and social responsibility, social networks and interpersonal relationships, cultural identity and society, social behavior and interaction, social mobility, work and occupations, modern society and traditional society, the internet and society, conflict and cooperation in society, globalization of society, law and social order, social capital, social welfare policies, values of modern society.
	Job Career	Career development, career planning, job search skills, resumes and cover letters, interview techniques, career goal setting, career path choices, career transitions and changes, career growth and development, workplace interpersonal relationships, office culture, compensation and benefits, workplace stress and coping, work-life balance, self-improvement and learning, leadership and management, workplace communication skills, teamwork and collaboration, professional skills and certifications, industry trends and prospects, work environment and company culture, occupational safety and health, professional ethics, entrepreneurship.

Figure 18: List of partial existing domains of conversations in MMRC.

granularity of the categories, our data would encompass more domains. According to our statistics, we have data from 874 different domains. Due to space limitations, we select a subset of domains, as shown in Fig. 18, conversation data in MMRC is comprehensive and diverse, covering various aspects of real-world scenarios. Therefore, the evaluation of MMRC provides valuable insights of MLLMs' performance in practical open-ending conversations.

G Experimental Settings

LLaVA-V1.5-7B, LLaVA-V1.5-13B, MiniCPM-8B, LLaVA-Next-0.5B, LLaVA-Next-7B, Qwen2VL-2B, Qwen2VL-7B, LLaVA-OneVision-0.5B, LLaVA-OneVision-7B, VILA1.5-3B, VILA1.5-8B, mplug-Ow3-1B, mplug-Ow3-2B, mplug-Ow3-7B, InternVL2.5-8B, and InternVL2.5-26B use 16-bit floating-point precision, while Qwen2VL-72B and LLaVA-OneVision-72B use 4-bit quantization. Their output length is limited to a maximum of 256 tokens. The models utilize the default values for Temperature, Top-k Sampling, and Top-p Sampling as specified in their Hugging Face inference code.

H Why DeepSeek-V3 is Classified as Close-Sourced in Experiment

Although the DeepSeek-V3 has been open-sourced on Hugging Face, its large parameter size makes local inference very resource-consuming. Therefore, we use the API for testing. In the experiment, we categorized it under the close-sourced section, as its parameter size is similar to that of close-sourced models, both exceeding 175B parameters.

I Discussion on the Impact of Different Model Architectures on Results

we have conducted a detailed investigation into the impact of different MLLM architectures and training data on model performance. We have summarized the various model architectures as follows: (1) Typical ViT-MLP-LLM structures: LLaVa-V1.5, LLaVA-NeXT-Interleave, LLaVA-OneVision, Qwen2-VL, VILA-1.5. (2)Additional multi-image processing architectures: MiniCPM (Shared Compression Layer), mplug-Ow3 (Hyper Attention Transformer Block). (3) Novel vision encoder: InternVL2.5 (InternViT).

From the experiment reuslt, We observe that models with extra multi-image processing architectures outperform the typical ViT-MLP-LLM models, with an average improvement of 12.16%. This suggests that architectural enhancements on top of the typical ViT-MLP-LLM framework are beneficial to dialogue capabilities, as they enhance multi-image understanding, giving such models an advantage in conversations involving multiple images. Furthermore, InternVL2.5 shows a significant improvement over the typical ViT-MLP-LLM structure, with a performance gain of 29.62%. Which can be attributed to the powerful InternViT, which offers superior visual understanding and includes advanced mechanisms such as Window Attention. These architectural innovations allow InternVL2.5 to more effectively capture visual details in multi-turn conversations, enabling it to provide more accurate answers to vision-related queries.

J Prompt Template and Human Criteria for Evaluation

In recent years, an increasing number of benchmark studies have utilized LLM as an evaluation tool, due to its accuracy, logical reasoning, and ability to follow instructions (Chen et al.; Li et al., 2024a; Gu et al., 2024; Ye et al.).

MMRC also applies GPT-4o for scoring the model's information extraction, cross-turn reasoning, information update, image management, long-term memory recall, and answer refusal capabilities. The detailed evaluation prompts corresponding to each capability are shown in Fig. 25, Fig. 26, Fig. 27, Fig. 28, and Fig. 29. We also conducte manual evaluation of the capabilities in cross-turn reasoning, image update, and memory recall. The three evaluation criteria are as follows:

(1) Cross-turn reasoning (CR):

Goal: To assess how well the model integrates and utilizes information across multiple dialogue turns to answer complex questions.

Criteria: *Full reasoning (5 points):* the model accurately integrates information from multiple dialogue turns and provides a clear answer that reflects all key details, demonstrating full understanding of the context; *Partial reasoning (3-4 points):* the model integrates some of the conversation's information, but the answer does not fully reflect all necessary context, it is partially correct; *Error in Reasoning (2 points):* the model fails to integrate information properly, providing an unclear or erroneous reasoning chain, leading to an inaccurate answer; *Lack of Reasoning (1 point):* the model does not demonstrate effective reasoning, providing a vague or irrelevant answer without appropriately incorporating the conversation's context.

(2) Information Update (IU):

Goal: To evaluate how well the model tracks and updates factual information provided by the user during the conversation.

Criteria: *Complete Update (5 points):* the model accurately and promptly updates new factual information, incorporating this into later responses and maintaining consistency throughout the conversation; *Partial Update (3-4 points):* the model updates some information but fails to fully incorporate the changes in subsequent responses. There may be some omissions or incomplete updates; *Failure to Update (2 points):* the model does not recognize the change in factual information and continues to rely on outdated data, leading to inconsistent responses; *Incorrect Update (1 point):* the model incorrectly updates information, causing contradictions with prior details or incorrect facts to be incorporated in later responses.

(3) Memory Recall (MR):

Goal: To assess how well the model maintains long-term memory of the conversation and recalls relevant details in later dialogue turns.

Criteria: *Accurate Recall (5 points):* the model accurately recalls key information (e.g., user preferences, previous conversation details) from earlier dialogue turns and integrates this information effectively in later responses; *Partial Recall (3-4 points):* the model recalls some details but misses out on others, leading to slightly disjointed or incomplete answers; *Memory Loss (2 points):* the model forgets important details over the course of the conversation, leading to inconsistent or disconnected responses that lack coherence; *Incorrect Recall (1 point):* the model recalls information incorrectly (e.g., mixing up user preferences or repeating outdated facts), leading to answers that are inaccurate or misleading.

K Attention Calculation

The phenomenon of attention imbalance distribution in models during dialogue shown in Fig 8, is consistent with recent research (Jawale et al., 2024). The following are the detailed steps for calculating attention: we randomly selecte 100 conversations with 15 turns to visualize the attention. For these dialogues, we perform manual special processing, where we label the information locations related to subsequent memory questions (*i.e.,*IE, IM, MR) as the 'golden position', referred as G. For example, if answering an IE question requires extracting information from turn 2 and turn 4, the golden position for this IE question would be turns 2 and 4. The role of the golden label in subsequent calculations will be explained in detail. For the attention calculation model, we selected LLaVA-Next-0.5B, LLaVA-OneVision-72B, MiniCPM-8B, and QwenVL-2B. These choices aim to cover a broad range of model families and parameter sizes.

When the model is tasked with answering memory-related questions, the input typically consists of a dialogue history and an memory evaluation question. Let the input be `query` = $[x_1, x_2, \ldots, x_n, q_s]$, where $[x_1, x_2, \ldots, x_n]$ represents the conversation history of the model, and each turn x_i (where $1 \leq i \leq n$) may contain image tokens and text tokens, and q_s represents the current evaluation question. To compute the self-attention weights for each turn, we define a function `Attn`: $\mathcal{X} \times \mathcal{N} \rightarrow \mathbb{R}$, which computes the average attention weight assigned to each turn x_i, where each turn $x_i = \{x_{i,j}\}_{j=1}^{N_i}$ contains N_i tokens. Only the turns in the golden position are considered in the attention calculation. The other turns are excluded to avoid interference from irrelevant information. We compute the average attention weight for each golden position turn using the following formula:

$$\text{Attn}(q_s, i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \text{attn}(x_{i,j}), \quad x_i \in G$$

L Details on NOTE-TAKING

The NOTE-TAKING employs another MLLM for note-taking, with model selection including open-source models like LLaVA-One-Vision-7B, MiniCPM-8B, as well as closed-source models like GPT-4o. Since the task of note-taking is easier than conversation, the choice of different models has little impact on note-taking performance. The algorithm for our proposed NOTE-TAKING strategy is in Alg. 1. Due to the limitations of space in the main text, we have only presented a portion of the experimental results. Table 6 shows the complete experimental results.

Algorithm 1 NOTE-TAKING algorithm

Require: MLLM A (notetaker), MLLM B (test model), user input $\mathcal{T}$, evaluation questions $\mathcal{Q}$.

Step1: Takes Notes

$Note = \{\}$
for $\mathcal{T}_i$ **in** T **do**
    $history = \text{append}(T_i)$
    $\mathcal{R}_i = Chat(\text{B}, history)$
    $Note = Take\_Note(\text{A}, T_i, Note)$
    $history = \text{append}(R_i)$
end for

Step 2: Help to response

for $\mathcal{Q}_i$ **in** Q **do**
    $history = \text{append}(Q_i)$
    $answer = \text{Chat}(\text{B}, (history, Note))$
    output $response_i$
end for

NOTE-TAKING Prompt

System Prompt

You are an intelligent chatbot designed to accurately and thoroughly record information from conversations, and then document it in a JSON format. Here is how you can accomplish your tasks:

- Focus on the user's preferences and the facts within the conversation.
- If the user presents a new fact, it should overwrite the outdated fact.

User Prompt

I will give you the user's input one by one. **Extract key information** from the input, including the **user's preferences, events, and facts** presented in user's input. If you detect a change in information, you should overwrite the outdated content in the notes with the **updated information**. REMEMBER NOT TO RESPOND TO THE USER'S INPUT! Here is the user input: {$\boldsymbol{T_i}$}, provide the ***Note*** in **JSON** format, for example: { "user_info": { "health_conditions": ["High blood pressure", "Dairy allergy"] }, "preferences": { "food_preference": "Italian cuisine" }, "purpose": "Seeking meal suggestions." }.

Figure 19: The prompt template of *Take_Note* function in Alg. 1.

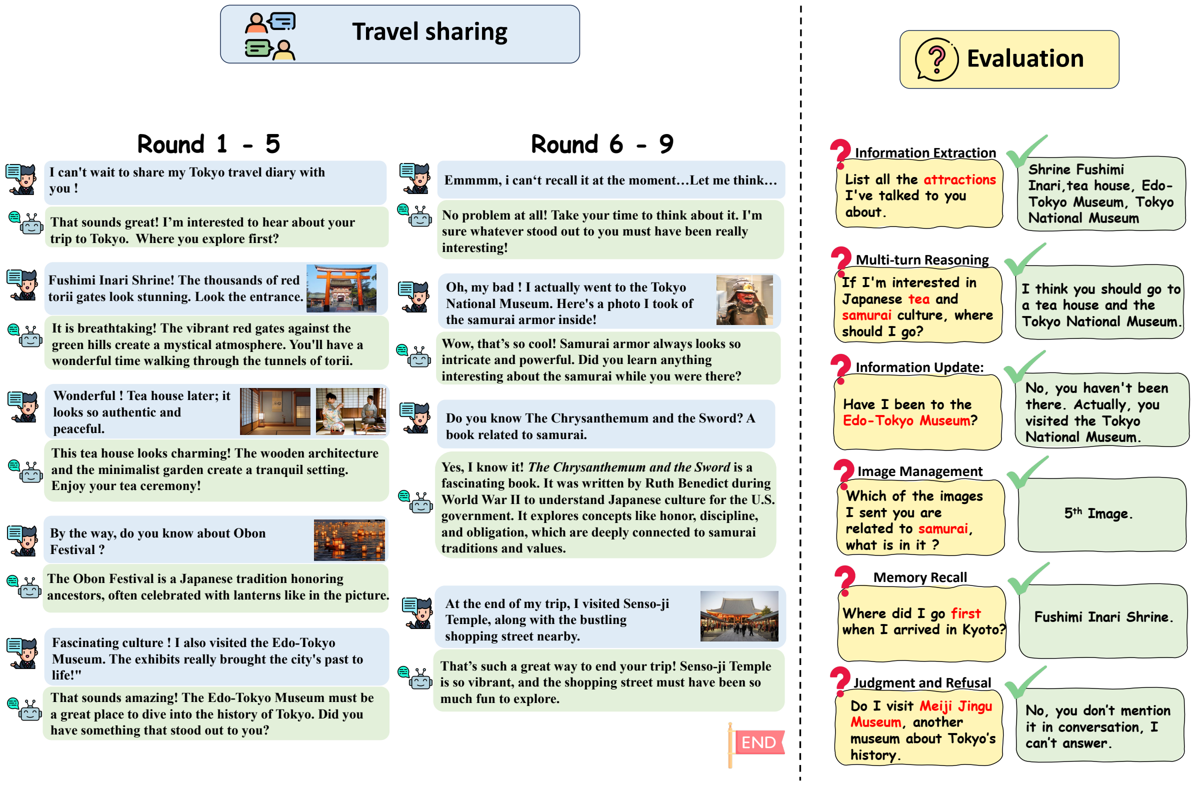

Figure 20: A sample from MMRC, the dialog domain is travel sharing, consisting of 9 conversation turns and 6 image inputs.

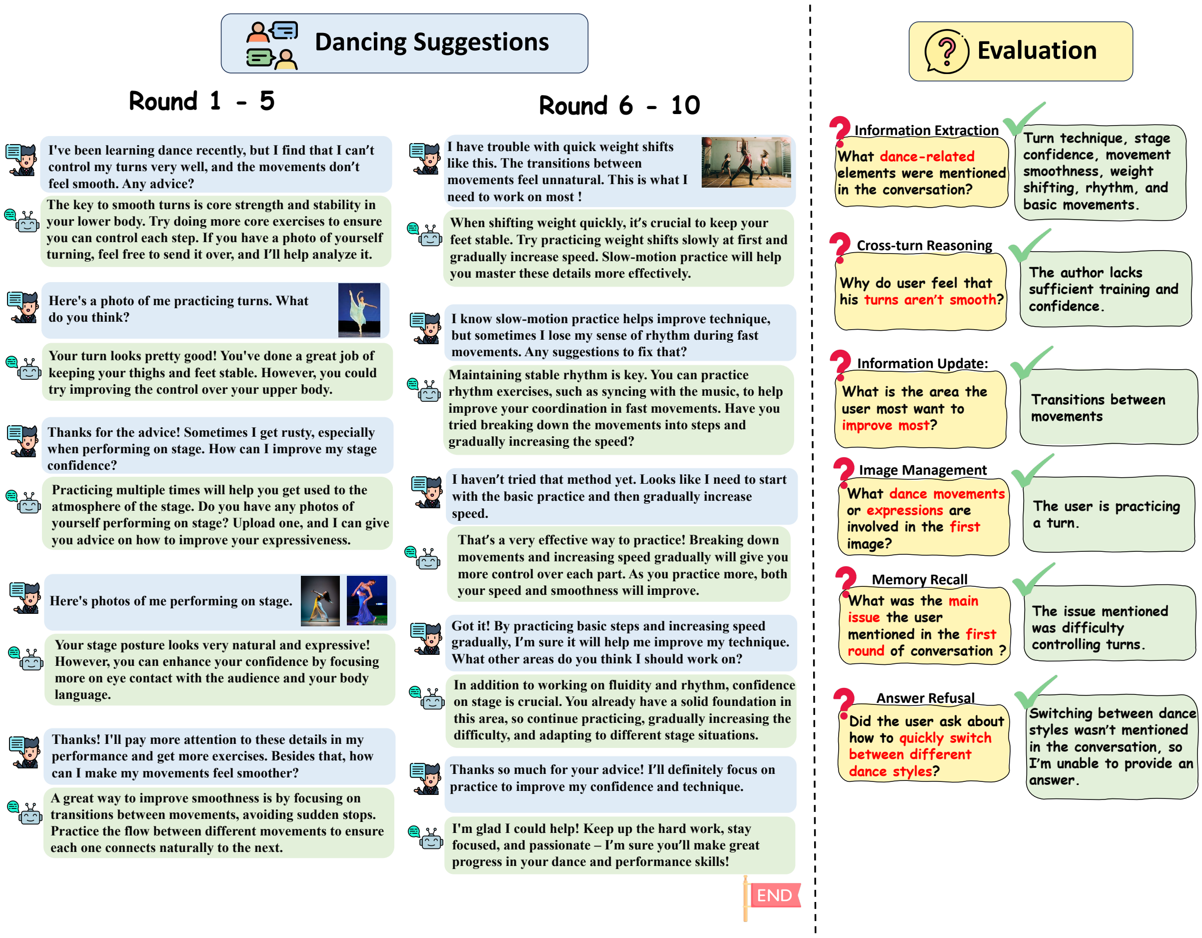

Figure 21: A sample from MMRC, the dialogue domain is dancing suggestions, consisting of 10 conversation turns and 4 image inputs.

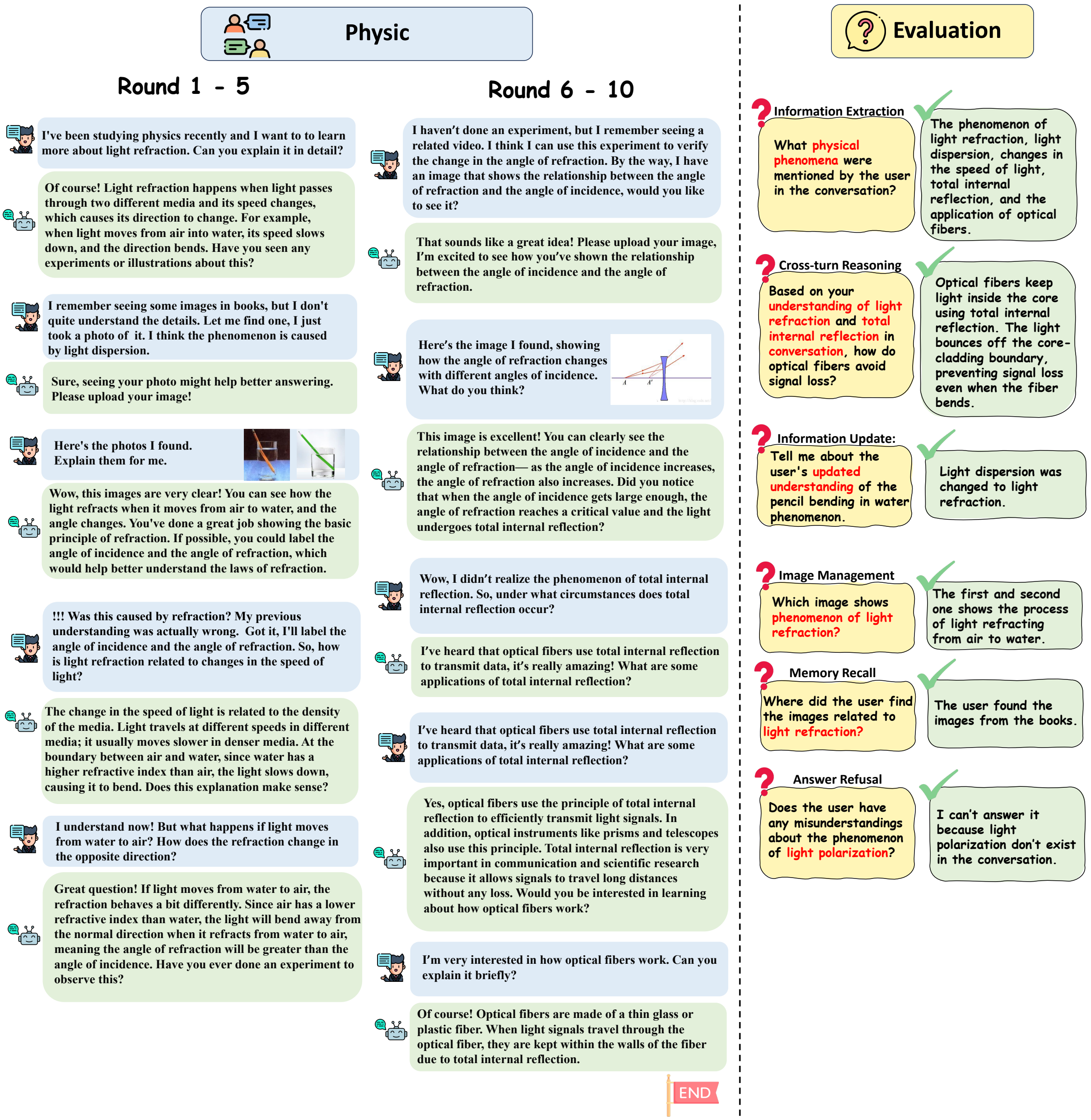

Figure 22: A sample from MMRC, the dialogue domain is physic, consisting of 10 conversation turns and 3 image inputs.

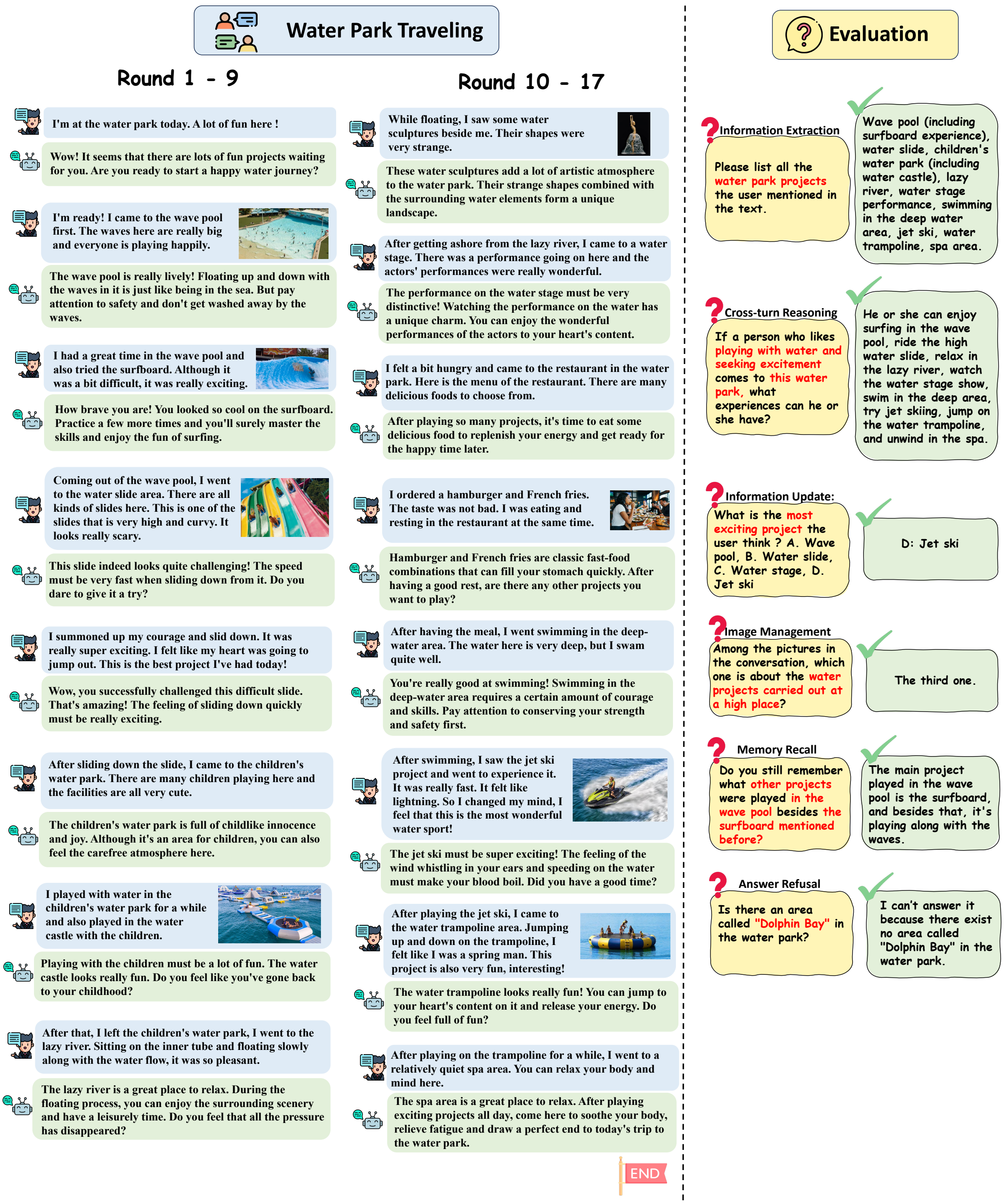

Figure 23: A sample from MMRC, the dialogue domain is water park, consisting of 17 conversation turns and 8 images inputs.

Figure 24: A sample from MMRC, the dialogue domain is dessert making, consisting of 17 conversation turns and 8 images inputs.

Models	IE	IU	MR	AR	CR	IM
LLaVA-1.5-7B	0.91	0.31	0.22	0.28	1.08	0.52
+ NOTE-TAKING(GPT-4o)	**2.57** (+1.66)	**2.06** (+1.75)	**2.36** (+2.14)	**0.46** (+0.18)	**1.22** (+0.14)	**0.54** (+0.02)
+ NOTE-TAKING(MiniCPM)	**2.33** (+1.42)	**1.78** (+1.47)	**2.11** (+1.89)	**0.35** (+0.07)	**1.19** (+0.11)	**0.56** (+0.04)
MiniCPM-8B	4.08	2.98	3.65	3.78	3.94	3.47
+ NOTE-TAKING(GPT-4o)	**4.23** (+0.15)	**3.76** (+0.78)	**4.02** (+0.37)	**3.92** (+0.14)	**4.13** (+0.09)	**3.50** (+0.03)
+ NOTE-TAKING(MiniCPM)	**4.12** (+0.04)	**3.45** (+0.47)	**3.74** (+0.09)	**3.81** (+0.03)	**4.02** (+0.08)	**3.52** (+0.05)
QwenVL-2B	2.16	1.41	1.93	2.33	2.85	1.27
+ NOTE-TAKING(GPT-4o)	**3.71** (+1.55)	**2.43** (+0.82)	**3.40** (+1.47)	**2.84** (+0.51)	**3.47** (+0.62)	**1.36** (+0.09)
+ NOTE-TAKING(MiniCPM)	**3.22** (+1.06)	**2.01** (+0.60)	**3.21** (+1.28)	**2.58** (+0.25)	**3.26** (+0.41)	**1.32** (+0.05)
LLaVA-Next-0.5B	2.32	1.99	2.67	1.12	2.89	1.87
+ NOTE-TAKING(GPT-4o)	**3.84** (+1.52)	**3.04** (+1.05)	**3.88** (+1.21)	**2.03** (+0.91)	**3.62** (+0.73)	**2.07** (+0.20)
+ NOTE-TAKING(MiniCPM)	**3.68** (+1.36)	**2.77** (+0.78)	**3.59** (+0.92)	**2.06** (+0.94)	**3.38** (+0.49)	**1.94** (+0.07)
LLaVA-OneVision-72B	4.06	4.01	4.17	2.52	4.08	3.24
+ NOTE-TAKING(GPT-4o)	**4.28** (+0.22)	**4.31** (+0.30)	**4.38** (+0.21)	**3.46** (+0.94)	**4.16** (+0.08)	**3.31** (+0.07)
+ NOTE-TAKING(MiniCPM)	**4.13** (+0.07)	**4.28** (+0.27)	**4.26** (+0.09)	**3.20** (+0.68)	**4.13** (+0.05)	**3.26** (+0.02)
GPT-4o	4.35	4.28	4.31	3.06	4.38	4.12
+ NOTE-TAKING(GPT-4o)	**4.51** (+0.16)	**4.62** (+0.34)	**4.73** (+0.42)	**3.51** (+0.45)	**4.41** (+0.03)	**4.18** (+0.06)
+ NOTE-TAKING(MiniCPM)	**4.46** (+0.11)	**4.45** (+0.17)	**4.37** (+0.06)	**3.32** (+0.26)	**4.40** (+0.02)	**4.16** (+0.04)

Table 6: Performance of NOTE-TAKING across all conversational core abilities in MMRC.

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate **model's information extraction ability, determine whether the model has extracted specific information.** Here is the model's prediction **{answer}**, and here is the ground truth **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 25: GPT prompt for evaluating information extraction (IE).

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate model's **cross-section reasoning ability to see if the model's reasoning results align with the logic of the ground truth.** Here is the model's prediction **{answer},** and here is the ground truth **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 26: GPT prompt for evaluating cross-turn reasoning (CR).

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate model's **information updating ability, determine whether the facts provided in the model's prediction align with the ground truth.** Here is the model's prediction **{answer}**, and here is the ground truth **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 27: GPT prompt for evaluating information update (IU).

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate model's **image management ability, check whether the model's response to the image aligns with the facts.** Here is the model's prediction: **{answer},** and here is the factual about the image: **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 28: GPT prompt for evaluating image management (IM).

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate model's **memory recall ability, to determine whether the model's memory aligns with the correct memory.** Here is the model's prediction **{answer}**, and here is the ground truth **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 29: GPT prompt for evaluating memory recall (MR).

System Prompt

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. The question-answer pair is from a long conversation between human and Multimodal Language Models. Models' answer in based on the conversation history. Your task is to score the models' answer. Here is how you can accomplish your tasks:

- Focus on the meaningful match between models' prediction and ground truth.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the ground truth.

User Prompt

You need to evaluate model's **answer refusal ability to see if the model correctly refuses to answer the question when appropriate.** Here is the model's prediction **{answer}**, and here is the ground truth **{label}.** Provide your evaluation only as a score where the score is an integer value between **0** and **5**, with **5** indicating the highest meaningful match. Please response in the form of a Python dictionary string with key 'score', where value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {{'score': 4}}.

Figure 30: GPT prompt for evaluating answer refusal (AR).