InterFeedback: Unveiling Interactive Intelligence of Large Multimodal Models via Human Feedback

Henry Hengyuan Zhao *, **Wenqi Pei***, **Yifei Tao***, **Haiyang Mei**, **Mike Zheng Shou**†

Show Lab, National University of Singapore

*Equal contribution †Corresponding author

Abstract

Existing benchmarks do not test Large Multimodal Models (LMMs) on their interactive intelligence with human users which is vital for developing general-purpose AI assistants. We design InterFeedback, an interactive framework, which can be applied to any LMM and dataset to assess this ability autonomously. On top of this, we introduce InterFeedback-Bench that evaluates interactive intelligence using two representative datasets, MMMU-Pro and MathVerse, to test 10 different open-source LMMs. Additionally, we present InterFeedback-Human, a newly collected dataset of 120 cases designed for manually testing interactive performance in leading models such as OpenAI-o1 and Claude-3.5-Sonnet. Our evaluation results show that state-of-the-art LMM (e.g., OpenAI-o1) can correct their results through human feedback less than 50%. Our findings point to the need for methods that can enhance LMMs' capabilities to interpret and benefit from feedback.

1 Introduction

In this paper, we are curious about the question ***"How do Large Multimodal Models perform with human feedback?"*** It is central to developing general-purpose AI assistants with Large Multimodal Models (LMMs). While these models are increasingly used to tackle multimodal tasks, their ability to interact with humans remains largely unknown. We argue that an LMM functioning as the general assistant should possess two capabilities: 1) exceptional problem-solving skills and 2) the ability to improve itself through feedback (e.g., human feedback, execution results). In this work, we focus on the latter capability, which has been rarely examined in existing benchmarks.

Humans are remarkably adaptive, continuously refining their skills by learning from feedback—a process fundamental to acquiring knowledge and solving problems. Similarly, advanced LMM models should also be capable of learning from feedback, thereby enhancing their problem-solving abilities as illustrated in Figure 1.

A surge of large multimodal models (LMMs) (OpenAI, 2023; Wang et al., 2024; Deitke et al., 2024; Wang et al., 2023; Zhao et al., 2024; Chen et al., 2024b) has emerged, designed to handle various tasks, including general vision-language understanding (Liu et al., 2023b; Li et al., 2023), expert-level multimodal understanding (Yue et al., 2024a;b), and scientific reasoning (Lu et al., 2022; 2024; Zhang et al., 2024). However, these LMMs are tested in a static way (Zhang et al., 2024; Yue et al., 2024a), overlooking their great potential in human-AI interaction (HAI). Consequently, a standard benchmark to test these LMMs for HAI problem-solving remains underexplored.

The key challenge in evaluating this interactive intelligence of LMMs is the automatic model tests. In practice, for the same query, different LMMs often produce varied responses, necessitating that humans offer tailored feedback for each conversation round. To address this issue, we propose **InterFeedback** a straightforward problem-solving framework that enables any LMM to tackle multimodal tasks interactively by leveraging leading models such as GPT-4o (OpenAI, 2023) to simulate humans, inspired in previous studies (Yao et al., 2025; Chen et al., 2024a; Yoon et al., 2024).

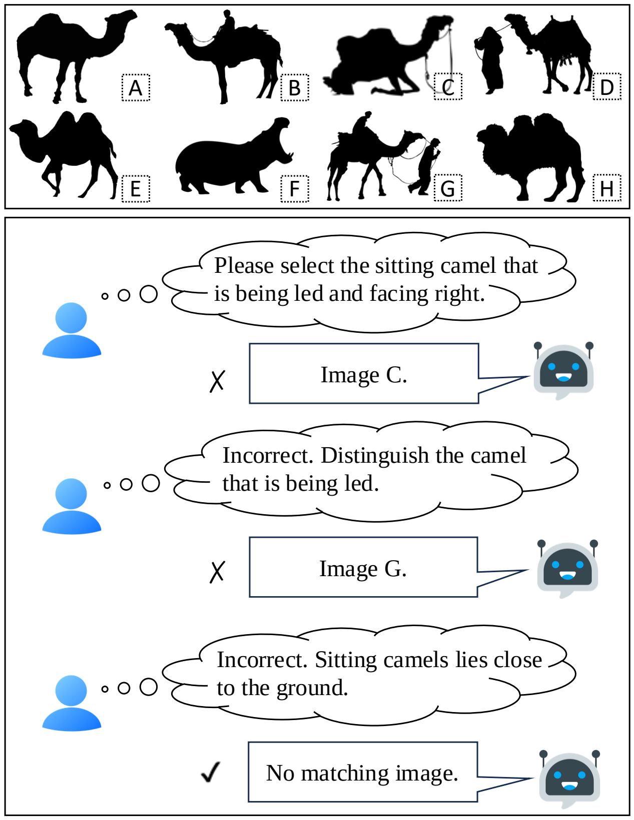

Figure 1: Illustration of an interactive feedback scenario. When models generate incorrect responses, human users provide pertinent feedback to iteratively refine the answers.

On top of this framework, we present **InterFeedback-Bench**, a benchmark designed to comprehensively evaluate LMMs for two purposes: 1) **the ability to interactively solve problems** and 2) **the capability of interpreting the feedback to improve themselves**. We demonstrate with two challenging pre-existing datasets: MMMU-Pro (Yue et al., 2024b) and Mathverse (Zhang et al., 2024). Additionally, for a more in-depth investigation, we conduct human evaluation on four closed-source leading models: GPT-4o (OpenAI, 2023), OpenAI-o1 (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024), and Gemini-2.0 (Gemini, 2025) with a trained user acting as the feedback provider. Finally, we manually collected a dataset **InterFeedback-Human** containing 120 samples for this assessment.

Our experimental results reveal several compelling insights: 1) Interactive process could improve the performance of most LMMs in solving challenging problems; 2) Existing LMMs exhibit suboptimal performance in interpreting and incorporating feedback; 3) Engaging in additional iterations does not necessarily guarantee the derivation of correct solutions; 4) High-quality feedback is essential, as subpar feedback can degrade performance even more than a simple binary (0/1) correctness signal; 5) LMM may not truly reasoning, we find out that LMMs resort to guessing answer even on a simple question according to human. These findings point to the need for methods that can enhance the LMM's capability to interpret and benefit from feedback. In summary, our contributions are:

- We take the first step toward exploring the interactive intelligence of LMMs in improving themselves through human feedback.
- We propose a straightforward and extensible framework InterFeedback which allows any LMM to interactively solve problems.
- We construct InterFeedback-Bench, a novel and universal benchmark for assessing the ability of interactive problem-solving of LMMs.
- We conduct comprehensive evaluations and in-depth analysis, providing several key insights for future development.

2 Related Work

2.1 Large Multimodal Models

The LLaVA-series works (Liu et al., 2023a; 2024a;b; Li et al., 2024a) demonstrate that training with supervised fine-tuning (SFT) multimodal data and expand the vision lens would produce compatible multimodal reasoning ability. By adopting a large-scale image-text corpus for instruction tuning, Qwen2-VL (Wang et al., 2024), CogVLM (Wang et al., 2023), InternVL2 (OpenGVLab, 2024) have achieved exceptional performance on various multimodal abilities. Moreover, Molmo (Deitke et al., 2024) proposes to train an LMM from scratch with only the human-annotated data. Unlike these large models, MiniCPM-V (Yao et al., 2024) and Phi-3.5-Vision (Abdin et al., 2024) propose to train lightweight yet SOTA LMMs. Despite these LMMs have demonstrated their understanding and reasoning ability on various difficulty-level multimodal benchmarks such as MMMU-Pro (Yue et al., 2024b) and MathVista (Lu et al., 2024), it is still unknown how well the interactive intelligence in an Human-AI Interaction scenario. In this paper, we conduct the evaluation of these LMMs to explore this basic yet vital capability (i.e., improving themselves from human feedback).

2.2 Multimodal Benchmarks

Traditional vision-language benchmarks focus on visual question answering (Goyal et al., 2017), image captioning (Chen et al., 2015; Plummer et al., 2015; Agrawal et al., 2019), as well as other benchmarks for specialized scenarios such as scene text understanding (Singh et al., 2019; Sidorov et al., 2020), commonsense reasoning (Zellers et al., 2019), outside knowledge (Marino et al., 2019; Schwenk et al., 2022). The recent development of LMM posts a strong need for modernized multimodal benchmarks (Fu et al., 2023; Liu et al., 2023b; Li et al., 2023; Yu et al., 2023; Yue et al., 2024a; Lu et al., 2024; Zhang et al., 2024) such as MMBench (Liu et al., 2023b), MMMU-pro (Yue et al., 2024b), and MathVerse (Zhang et al., 2024) which involve comprehensively evaluating current LMMs on various multimodal abilities. However, these benchmarks primarily focus on static testing processes, overlooking the interactive testing process that is vital in human-AI interaction scenarios.

2.3 Human-AI Interaction

Investigating how humans and AI systems communicate and collaborate is critical for shaping applications such as virtual assistants (Virvou, 2022), personalized recommendations (Dodeja et al., 2024), autonomous vehicles (Zhang et al., 2021), and healthcare diagnostics (McKinney et al., 2020). Recent LLMs-driven techniques such as memory (Park et al., 2023) and iterative (Zhang et al., 2023) mechanisms offer expert-level collaboration. While LMMs excel in multimodal tasks (Deitke et al., 2024; Wang et al., 2024), their potential for HAI problem-solving remains underexplored. By offering a unified framework and meticulously curated data, our InterFeedback-Bench enables evaluation of LMMs on these capabilities and lays a foundation for advancing multimodal HAI problem-solving.

3 InterFeedback-Bench

In this section, we begin by introducing the interactive benchmarking component of our InterFeedback-Bench in Section 3.1. Here, we propose an interactive human-AI framework, InterFeedback, designed as the evaluation tool for assessing LMM performance with feedback. Next, in Section 3.2, we detail the human benchmarking aspect of our benchmark, including the data sources and testing standards.

3.1 Interactive Benchmarking

3.1.1 Formulation

The InterFeedback-Bench formalizes the interactive problem-solving process with feedback in a partially observable Markov decision process (POMDP) $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ with state space $\mathcal{S}$, observation $\mathcal{O}$, action space $\mathcal{A}$, transition function $\mathcal{T}$: $\mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, and reward function $\mathcal{R}$: $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$.

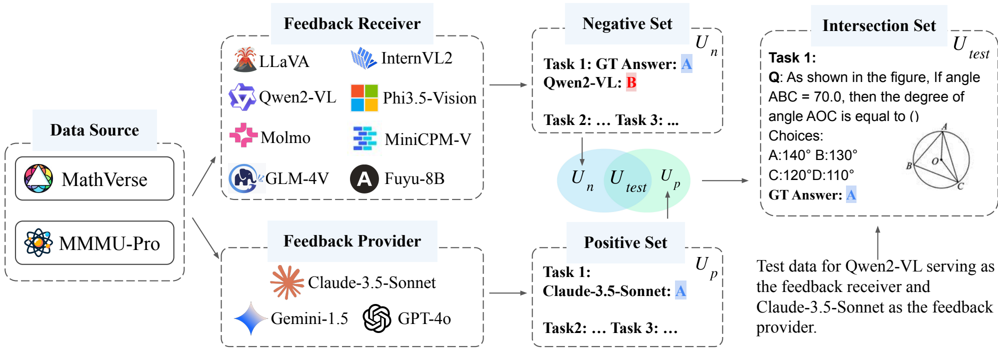

Figure 2: Overview of the test data construction process for InterFeedback-Bench. For each LMM serving as the feedback receiver, we process each instance from a target dataset (e.g., MathVerse) and collect the error cases to form a negative set. The feedback provider then processes the same instances to build a positive set. Finally, we curate test data by selecting the intersection of both sets.

In our setting, given a natural language question q (eg., *Please select the sitting camel that is being led and facing right*) and the input image v, the model first gets the observation $o_t \in \mathcal{O}$ from the state $s_t \in \mathcal{S}$ in the execution environment and then generate the action $a_t \in \mathcal{A}$. The a_t is the response from models in natural language. The reward function $\mathcal{R}$: $\mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ here returns a binary integer indicating the task correctness status. It is implemented by the exact match that compares the ground-truth answer and the predicted answer. The observation o_t includes both the correctness signal from the reward function and the feedback from the humans.

3.1.2 Data Sources

To ensure the quality and difficulty of multimodal tasks, inspired by previous benchmarks demonstrated on pre-existing datasets (Yang et al., 2023; Li et al., 2024c), we choose to test LMMs on two challenging datasets: MathVerse (Zhang et al., 2024) and MMMU-Pro (Yue et al., 2024b). MathVerse is a visual math benchmark that includes various mathematic problems, and 3,940 samples were used in our work. MMMU-Pro is a comprehensive multimodal benchmark with 1,730 expert-level questions. Both datasets are challenging even for the model GPT-4o (i.e., 64.7% accuracy).

3.1.3 Data Construction Process

We choose to use leading LMMs, such as GPT-4o, for stimulating the humans to give feedback mimicking human-AI interactions. The primary challenge, however, is ensuring that the feedback generated by these models is reliable as even the SOTA LMM like GPT-4o and Claude-3.5-Sonnet perform not all correctly on all test samples. Therefore, we construct the test data by selecting the intersection set that feedback provider M_p solves correctly while M_r does not as shown in Figure 2. Specifically, the pipeline includes three parts: 1) feedback receiver LMM locally-running; 2) feedback provider LMM API-calling; and 3) Intersection set selection. Such a data construction process leads to each tested LMM having a different test data set.

Given a test dataset D, we begin by having the feedback receiver model M_r process every instance in D to produce a negative set U_n consisting of tasks it fails to solve correctly. Next, the feedback provider model M_p processes the same dataset to generate a positive set U_p comprising tasks it solves correctly. We then define U_{test} as the intersection of U_n and U_p, i.e.,

$$U_{\text{test}} = U_n \cap U_p,$$

which means that U_{test} contains tasks that M_p solves correctly but M_r does not. This approach ensures that the feedback generated by M_p is both relevant and reliable.

3.1.4 InterFeedback Framework

To make the problem-solving process in an interactive way, we propose a new straightforward framework **InterFeedback**. It includes two roles: feedback receiver M_r and feedback provider M_p, as shown in Figure 3. The feedback receiver is the candidate LMMs (e.g., Qwen2-VL) ready for the benchmark and the feedback provider is the SOTA LMM (e.g., GPT-4o) for providing the pertinent feedback in each time step in place of a human. Consider in time t, the output of M_r is a_t, and the feedback provider M_p has to follow the policy that provides the feedback f_t from the mapping : $F(a_t, s_t) \rightarrow f_t$. The s_t denotes the correctness signal from the verification process via the reward function. We record the model outputs for the final evaluation.

3.2 Human Benchmarking

As use SOTA LMMs play the role of feedback provider, how do these LMMs perform when they are feedback receivers? We begin to assess the SOTA LMMs with a human-in-the-loop process. The feedback provider M_p is a trained user who fully understands all the questions in the newly curated dataset InterFeedback-Human. The feedback receiver M_r is the closed commercial LMM such as OpenAI-o1, GPT-4o, Gemini-2.0, and Claude-3.5-Sonnet. This evaluation aims to assess how effectively these leading models can serve as assistants in a human-AI interaction system.

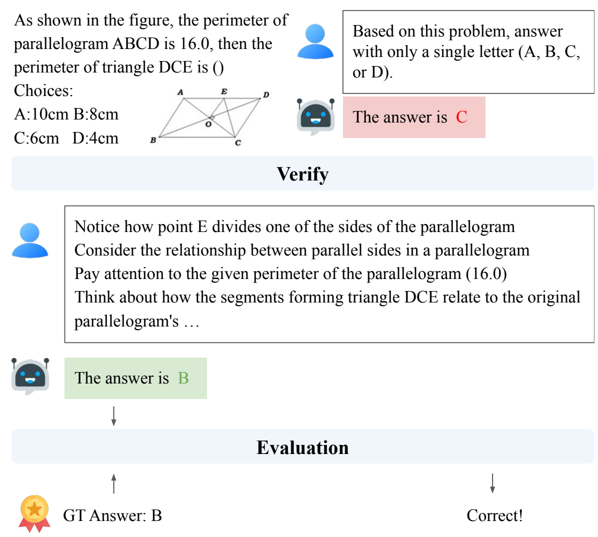

Figure 3: Overview of the proposed framework InterFeedback for assessing an LMM's ability to improve itself through feedback. The model interacts with humans to progressively solve a problem, and after each conversation round, we verify the correctness of the answer. If the answer is incorrect, an LMM-stimulated human will provide constructive feedback.

3.2.1 Data Sources

We gather the data with high difficulty and diversity across the domains: visual logic, mathematics, and coding. These were selected to probe the cognitive depth of the models, especially when confronted with complex, multi-step reasoning problems. The visual logic data we manually collected from publicly available resources. The emphasis on visual logic tasks reflects the growing demand for models to handle image-based reasoning challenges, such as pattern recognition (Wei et al., 2025) (e.g., determining the next shape in a sequence) and character-based logic (e.g., interpreting transformations between symbols). We also collect the multimodal mathematic data from the existing dataset MathVerse (Zhang et al., 2024) and the multimodal expert-level data from MMMU-Pro (Yue et al., 2024b). Additionally, we also involve the natural language task into InterFeedback-Human to analyze such capability in the NLP area.

3.2.2 Data Statistics

In summary, InterFeedback-Human encompasses a total of 120 tasks distributed across the five task types: 80 visual logic tasks, 10 mathematical logic tasks (sampled from NuminaMath (Li et al., 2024b)), 10 coding tasks (sampled from CodeComprehension (Imbue, 2024)), 10 MMMU-Pro tasks, and 10 MathVerse tasks.

3.2.3 Hierachical Feedback

We design a hierarchical feedback generation scheme to gradually increase the information intensity. Specifically, we ask the human to give the following three-level feedback:

- Level 1: Provide a basic and simple description that leads to the correct answer.
- Level 2: Provide an expanded explanation that leads to the correct answer.

- Level 3: The correct answer is GT Answer. Provide a comprehensive and detailed explanation that leads to the correct answer.

Since most of our questions have four options, giving more than three rounds of feedback might let the model guess the answer by elimination rather than by reasoning. For example, if the correct answer is A and the model already gave B, C, and D, a third round of feedback is unnecessary. Therefore, we directly provide the GT Answer in Level 3 feedback prompts to test the models' ability to explain their thinking process.

3.2.4 Evaluation Integration

To ensure fairness and consistency in our evaluation, we engaged only one experienced user. Since human-in-the-loop feedback is inherently subjective, involving multiple participants could introduce variability due to differences in background and expertise. This approach helps maintain the reliability of the relative performance comparisons across candidate LMMs.

4 Experiments

4.1 Experiment Setup

Evaluation Models. We evaluate the performance of foundation models served as the feedback receiver M_r across 10 representative LMMs: LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-1.6-7B (Liu et al., 2024b) (Mistral-7B), LLaVa-OneVision-7B (Li et al., 2024a) (Qwen2-7B (Yang et al., 2024)), Qwen2-VL-7B (Wang et al., 2024), GLM-4V-9B (Wang et al., 2023), InternVL2 (OpenGVLab, 2024), Molmo (Deitke et al., 2024), MiniCPM-V (Yao et al., 2024), Phi-3.5-Vision (Abdin et al., 2024), and Fuyu-8B (Bavishi et al., 2023). The feedback provider M_p includes the three best available models from three model families: OpenAI (`gpt-4o-2024-08-06`), Gemini (`Gemini-1.5-Pro`), and Claude (`Claude-3.5-Sonnet-2024-10-22`).

Evaluation Metrics. In addition to the **Accuracy** metric, we leverage the **Correction Rate**, defined as the percentage of corrected answers of all erroneous samples. Let N denote the total number of samples, N_e the number of erroneous samples, and N_c the number of samples that have been corrected. The Accuracy and Correction Rate metrics can be formulated as follows:

$$\text{Accuracy} = \frac{(1 - N_e)}{N} \times 100\%, \quad \text{Correction Rate} = \frac{(N_c)}{N_e} \times 100\%. \tag{1}$$

Implementation Details. We set the temperature to 0 for all tested models and API models. The image resolution of the Qwen2-VL model we restrict to 512 $\times$ 512 to avoid the memory exceeded error. All evaluations were conducted on two NVIDIA RTX A6000 GPUs. To ensure the reliability of results, we obtain the intersection set for both the feedback receiver and provider models that are able to output the correct answer format. Based on our preliminary experiments, we limited the interactive benchmarking to a single round. This decision is driven by two observations: most models fail to provide correct answers in subsequent rounds, and multiple rounds tend to lead to answer guessing, which undermines the reliability of quantitative evaluation.

Feeback Types. As introduced in Section 3.1, we employ closed-source LMMs to stimulate the human to provide pertinent feedback at each conversation round. Additionally, we propose a simplified feedback mechanism that only indicates correctness (i.e., correct or incorrect), without a detailed explanation. In summary, we evaluate the models using two feedback types: *Detail* and *Simple*. The *Detail* feedback comprises both *Simple* feedback and detailed LMM-generated feedback.

4.2 Experimental Analysis on Interactive Benchmarking

To thoroughly investigate the ability of LMMs to integrate feedback and improve their problem-solving performance, we present evaluation results for various models on two datasets—MathVerse (Zhang et al., 2024) in Table 1 and MMMU-Pro (Yue et al., 2024b) in Table 2, respectively. Below, we provide a detailed discussion of key findings.

Table 1: **Correction Rate Results of three Feedback Providers on MathVerse Dataset. Acc (%)**: The average accuracy of MathVerse's *testmini* set. The results are tested by ourselves. **# Neg**: The number of negative samples produced by the model. **# Test**: The total number of test samples evaluated. **Detail (%)**: correction rate of using LMM-generated feedback. **Simple (%)**: correction rate of using simple feedback (0 or 1).

Model			GPT-4o			Gemini-1.5-Flash			Claude-3.5-Sonnet		
	Acc (%)	# Neg	# Test	Detail (%)	Simple (%)	# Test	Detail (%)	Simple (%)	# Test	Detail (%)	Simple (%)
LLaVa-OneVision-7B	25.6	2933	373	36.2	18.0	428	29.0	15.7	2953	4.1	2.4
InternVL2-8B	38.1	2440	379	49.6	41.2	375	48.8	44.4	376	43.4	40.2
Molmo-7B	25.6	2931	452	55.1	52.0	507	36.5	38.9	597	37.4	40.0
MiniCPM-V	16.2	3301	552	28.4	20.3	741	16.6	25.4	772	18.7	27.1
GLM-4V-9B	20.2	3146	440	38.6	28.2	568	30.1	29.9	603	30.0	26.4
Phi3.5-Vision-4.2B	19.0	3192	534	36.1	33.7	579	31.3	33.7	616	26.8	29.1
LLaVa-1.5-7B	13.5	3409	763	23.2	14.3	678	18.0	14.7	816	8.3	11.2
LLaVa-1.6-Mistral-7B	14.8	3357	549	41.0	35.9	661	5.9	5.9	617	33.5	33.2
Fuyu-8B	21.8	3083	582	24.1	19.8	635	15.0	12.9	755	14.0	11.5
Qwen2-VL-7B	22.5	3052	295	66.8	72.2	470	41.9	44.9	505	50.5	52.7

Table 2: **Correction Rate Results of three Feedback Providers on MMMU-Pro Dataset.** We test models on a single image setting of MMMU-Pro.

Model			GPT-4o			Gemini-1.5-Flash			Claude-3.5-Sonnet		
	Acc (%)	# Neg	# Test	Detail (%)	Simple (%)	# Test	Detail (%)	Simple (%)	# Test	Detail (%)	Simple (%)
LLaVa-OneVision-7B	47.1	915	312	31.7	15.7	333	35.4	18.6	408	27.5	16.4
InternVL2-8B	45.7	939	343	50.1	41.4	329	57.1	50.2	437	50.1	41.2
Molmo-7B	43.8	973	362	51.7	48.9	383	41.5	43.1	436	29.8	27.5
MiniCPM-V	38.1	1071	410	27.3	23.7	503	21.5	21.7	540	24.4	23.3
GLM-4V-9B	46.0	935	327	38.8	30.0	359	38.7	31.5	441	34.9	27.9
Phi3.5-Vision-4.2B	43.2	983	366	44.3	42.3	396	40.9	39.6	484	39.9	38.0
LLaVa-1.5-7B	36.5	1099	506	31.9	12.3	470	20.0	16.0	595	13.9	13.4
LLaVa-1.6-Mistral-7B	38.8	1058	432	46.1	36.1	429	14.7	14.7	515	42.3	35.3
Fuyu-8B	34.1	1140	481	6.0	8.7	1140	3.7	3.5	612	9.5	6.9
Qwen2-VL-7B	48.1	898	268	50.4	44.8	322	39.4	37.6	389	42.9	37.3

Interactive process could improve the performance of most LMMs. As demonstrated in both tables, integrating our proposed framework InterFeedback enables most models to benefit from feedback provided by SOTA LMMs, such as GPT-4o and Claude-3.5-Sonnet. Notably, even the weaker model Fuyu-8B sees 24.1% of its erroneous samples corrected through GPT-4o's feedback.

Current LMMs struggle to enhance performance through feedback. As shown in the tables, most LMMs are unable to correct all erroneous samples, even when provided with feedback from state-of-the-art closed-source models such as Claude-3.5-Sonnet and GPT-4o. For example, consider the two leading open-source models, Qwen2-VL-7B and Molmo. Qwen2-VL-7B achieves a 66.8% correction rate on the MathVerse dataset with GPT-4o's feedback, but only a 50.4% correction rate on the MMMU-Pro dataset. Similarly, Molmo-7B attains correction rates of 55.1% and 51.7% on the MathVerse and MMMU-Pro datasets, respectively. Overall, the correction rates for the rest models remain below 50%. This suggests that even with constructive feedback from advanced LMMs, current models struggle to enhance performance through feedback generally.

Accuracy result may not truly reflect the model's capability. As shown in Table 1, although InternVL2-8B achieves a higher accuracy (38.1%), its correction rate is only 49.6%. In contrast, Qwen2-VL-7B, with a lower accuracy of 22.5%, attains the highest correction rate of 66.8% when using GPT-4o's feedback. Similarly, Molmo-7B surpasses InternVL2-8B in correction rate despite having lower accuracy. On the MMMU-Pro dataset (see Table 2), LLaVA-OneVision-7B records the second-best accuracy (i.e., 47.1%) but only a 31.7% correction rate, which is lower than that of several models who have inferior accuracy (e.g., InternVL2-8B, Molmo-7B, GLM-4v-9B, and Phi3.5-Vision-4.2B). This inconsistency between initial answering ability and self-improvement capability indicates that evaluating models solely on accuracy may not fully capture their true potential.

Simple feedback also enhances performance. In addition to using detailed LMM-generated feedback, we evaluated models with binary (0/1) feedback that simply indicates the correctness of their current response. Surprisingly, the results show that all models benefit from this simple feedback

Table 3: **Human Evaluation Results across LMMs on InterFeedback-Human.** $\text{Math}^{\text{Text}}$ and $\text{Coding}^{\text{Text}}$ represent two text-only task categories. The scores represent the average percentage of correct samples among all samples.

Model	Visual Logic	MMMU-Pro	MathVerse	$\text{Math}^{\text{Text}}$	$\text{Coding}^{\text{Text}}$	Average
Gemini-2.0	21.3	50.0	70.0	50.0	50.0	32.5
Claude-3.5	**37.5**	60.0	80.0	70.0	70.0	**48.3**
OpenAI-o1	28.8	60.0	**90.0**	**90.0**	**90.0**	46.7
GPT-4o	25.0	**70.0**	80.0	60.0	50.0	38.3

Table 4: **Correction Rate Results across various LMMs on InterFeedback-Human.** $\text{Math}^{\text{Text}}$ and $\text{Coding}^{\text{Text}}$ represent two text-only task categories. # Round denotes the number of interaction rounds. The correction rate is the percentage of corrected samples among all erroneous samples.

Model	# Round	Visual Logic	MMMU-Pro	MathVerse	$\text{Math}^{\text{Text}}$	$\text{Coding}^{\text{Text}}$	Average
Gemini-2.0	1	38.1	20.0	33.3	0.0	80.0	37.0
	2	20.6	0.0	33.3	20.0	20.0	19.8
	3	41.3	80.0	33.3	80.0	0.0	43.2
Claude-3.5	1	38.0	0.0	50.0	33.3	66.7	37.1
	2	32.0	25.0	50.0	33.3	66.7	30.6
	3	30.0	75.0	0.0	66.7	0.0	32.3
OpenAI-o1	1	38.6	0.0	100.0	11.1	100.0	39.1
	2	21.1	0.0	0.0	0.0	0.0	18.8
	3	40.4	100.0	0.0	0.0	0.0	42.2
GPT-4o	1	41.7	33.3	100.0	25.0	40.0	41.9
	2	31.7	0.0	0.0	0.0	0.0	25.7
	3	26.7	66.7	0.0	75.0	60.0	32.4

mechanism. This suggests that while LMMs have the inherent potential to generate correct answers, they may require additional prompting techniques to fully harness the problem-solving capabilities.

LMM-generated feedback is not always better than simple feedback. By comparing the results obtained using *Detail* feedback from GPT-4o with those using *Simple* binary feedback, we observe that most models perform better with detailed feedback. For example, on the MathVerse dataset, LLaVA-OneVision-7B achieves 36.2% with detailed feedback versus 18.0% with binary feedback; InternVL2-8B increases from 41.2% to 49.6%; and MiniCPM-V increases from 20.3% to 28.4%. The only exception is Qwen2-VL, which scores 66.8% with detailed feedback and 72.2% with simple feedback. Similarly, on the MMMU-Pro dataset, only Fuyu-8B performs worse with detailed feedback (6.0% vs. 8.7%).

The quality of feedback is crucial: low-quality feedback can degrade performance more than simply providing binary (0/1) feedback. We compare the feedback provided by GPT-4o and Gemini-1.5-Flash on the challenging MathVerse dataset, where most models achieve accuracies below 30%, highlighting the difficulty of its problem instances. We find that leveraging a suboptimal model (Gemini-1.5-Flash) to deliver simple binary feedback—merely indicating the correctness of the tested model's output—can outperform LMM-generated detailed feedback. Specifically, the correction rates using simple feedback exceed those with detailed feedback for several models: Molmo-7B (38.9% vs. 36.5%), MiniCPM-V (25.6% vs. 16.6%), Phi3.5-Vision-4.2B (33.7% vs. 31.3%), and Qwen2-VL-7B (44.9% vs. 41.9%).

4.3 Experimental Analysis on Human Benchmarking

In this section, we will introduce the human evaluation results of several well-known closed-source families: OpenAI (`GPT-4o`, `OpenAI-o1`), Claude (`Claude-3.5-Sonnet-20241022`), and Gemini (`Gemini-2.0-Flash-Exp`).

Overall Results. In Table 3: (1) The best scores for each subcategory in our InterFeedback-Human are 37.5% (Claude-3.5-Sonnet), 70.0% (GPT-4o), 90% (OpenAI-o1), and 90% (OpenAI-o1), respectively. (2) Overall, Claude-3.5 achieves the highest average accuracy at 48.3%.

Correction rate results analysis. Comparing the correction rates across rounds in Table 4 reveals that GPT-4o benefits the most from human feedback in the first round, correcting 41.9% of erroneous samples, while Claude-3.5 exhibits its strongest correction performance in the second round, with

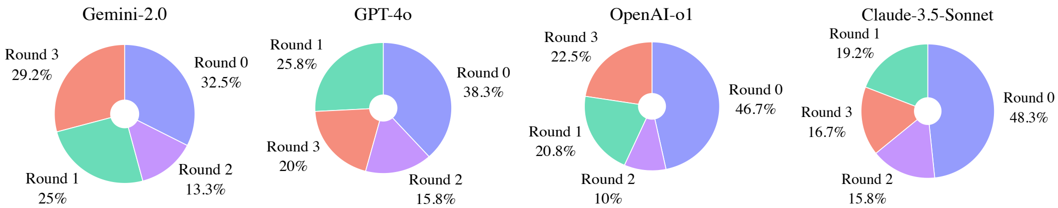

Figure 4: Distribution of samples being corrected in each round. We can observe that Claude-3.5-Sonnet archives the best performance in round 0.

30.6% of erroneous samples corrected. Given that the ground truth answer is provided in the third round, all LMMs are able to supply their reasoning steps for selecting the correct answer.

Distribution of Tasks Corrected Across Rounds. Figure 4 illustrates the distribution of tasks solved by each LMM across the interaction rounds. Round 0 represents the initial accuracy before beginning human-AI interactions. For example, GPT-4o solved 38.3% of instances in Round 0, 25.8% in Round 1, and 20% in Round 2. Additionally, during the first two rounds, both OpenAI-o1 and Claude-3.5-Sonnet solved the same number of samples, achieving a performance of 67.5%.

Distribution of corrected samples across various task categories. As shown in Figure 5, Visual logic tasks are mostly resolved within the first two rounds, whereas Math (Text-only) and MMMU-Pro tasks show little corrections in rounds 1 and 2. In contrast, Coding (Text-only) and MathVerse tasks exhibit corrections during rounds 1 and 2.

Summarization. The closed-source SOTA LMMs demonstrate enhanced problem-solving capabilities when provided with human feedback. Most models show improvement after the first round of feedback, with over 55% of samples being addressed.

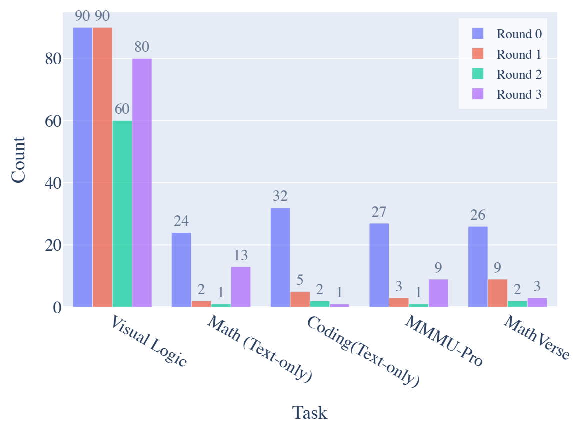

Figure 5: Distribution of corrected samples across various task categories. Visual logic tasks are mostly resolved within the first two rounds, whereas Math (Text-only) and MMMU-Pro tasks show little corrections in rounds 1 and 2. In contrast, Coding (Text-only) and MathVerse tasks exhibit corrections during rounds 1 and 2.

4.4 Limitations

Our method is not without limitations. First, as an initial attempt in the field, this work proposes a straightforward method to bootstrap the LMMs in an interactive way. We use the leading LMM to stimulate the humans mimicking the human-AI interaction process. Due to the difficulty of existing benchmarks, the leading LMMs may not fully provide all pertinent feedback though we propose two strategies: 1) select the intersection set for testing and 2) record the valid output only. Second, due to the testing limits of Deepseek-R1, we cannot test its interactive intelligence in this version.

5 Conclusion

In this work, we introduced InterFeedback-Bench, the first solution to concern the critical importance of evaluating the interactive intelligence of current LMMs. We build an interactive framework InterFeedback which can be applied to any LMM and dataset to bootstrap the testing in an interactive way. We conduct the comprehensive evaluations on 10 open-source LMMs by demonstrating with two representative datasets MathVerse and MMMU-Pro. Additionally, we present InterFeedback-Human, a new benchmark for manually testing the leading models such as OpenAI-o1 and Claude-3.5 with 120 curated samples. Our evaluation results show that even the SOTA LMM (like OpenAI-o1) can only correct their results through human feedback with less than 50%. Several findings point to the essential need for methods that improve the LMM's ability to receive feedback to improve themselves.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku, 2024. URL https://www.anthropic.com/news/3-5-models-and-computer-use.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Sanxing Chen, Sam Wiseman, and Bhuwan Dhingra. Chatshop: Interactive information seeking with language agents, 2024a. URL https://arxiv.org/abs/2404.09911.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Lakshita Dodeja, Pradyumna Tambwekar, Erin Hedlund-Botti, and Matthew Gombolay. Towards the design of user-centric strategy recommendation systems for collaborative human–ai tasks. *International Journal of Human-Computer Studies*, 184:103216, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

Gemini. Our next-generation model: Gemini 1.5, 2024. URL https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/.

Gemini. Gemini 2.0, 2025. URL https://deepmind.google/technologies/gemini/.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Imbue. Imbue code comprehension, 2024. URL https://huggingface.co/datasets/imbue/code-comprehension.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024b.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models, 2024c. URL https://arxiv.org/abs/2411.17451.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

OpenAI. Gpt-4o, 2023. URL https://openai.com/index/hello-gpt-4o.

OpenAI. Openai o1 system card, 2024. URL https://openai.com/index/openai-o1-system-card.

OpenGVLab. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, July 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Maria Virvou. The emerging era of human-ai interaction: Keynote address. In *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–10. IEEE, 2022.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, and Daxin Jiang. Slow perception: Let's perceive geometric figures step-by-step, 2025. URL https://arxiv.org/abs/2412.20631.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

John Yang, Akshara Prabhakar, Karthik R Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=fvKaLF1ns8.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1490–1504, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.83. URL https://aclanthology.org/2024.naacl-long.83/.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024b. URL https://arxiv.org/abs/2409.02813.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Jiehuang Zhang, Ying Shu, and Han Yu. Human-machine interaction for autonomous vehicles: A review. In *International Conference on Human-Computer Interaction*, pp. 190–201, 2021.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Liyang Zhou, Hainiu Xu, Li Zhang, Lara J Martin, Rotem Dror, Sha Li, et al. Human-in-the-loop schema induction. *arXiv:2302.13048*, 2023.

Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. LOVA3: Learning to visual question answering, asking and assessment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=vIOKLMl6wu.

A Additional Experimental Details

A.1 Model Sources.

For different LMMs, we select their latest models with sizes around 7B for evaluation. Table 5 presents the release time and model sources of LMMs used in InterFeedback-Bench.

Table 5: The release time and model source of LMMs used in our InterFeedback-Bench.

Model	Release Time	Source
		Closed-source Models
GPT-4o (OpenAI, 2023)	2024-08-26	https://openai.com/index/hello-gpt-4o/
OpenAI-o1 (OpenAI, 2024)	2024-12-17	https://openai.com/o1/
Gemini-1.5-Flash (Gemini, 2024)	2024-09-24	https://deepmind.google/technologies/gemini/
Gemini-2.0-Flash	2025-01-21	https://deepmind.google/technologies/gemini/
Claude-3.5-Sonnet	2024-10-22	https://www.anthropic.com/claude/sonnet
		Closed-source Models
LLaVA-One-Vision	2024-08-05	https://llava-vl.github.io/blog/2024-08-05-llava-onevision/
InterVL2-8B	2024-07-04	https://internvl.github.io/blog/2024-07-02-InternVL-2.0/
Molmo-7B	2024-09-24	https://huggingface.co/allenai/Molmo-7B-D-0924
MiniCPM-V	2024-08-03	https://huggingface.co/openbmb/MiniCPM-V
GLM-4V-9B	2024-11-01	https://huggingface.co/THUDM/glm-4v-9b
Pih3.5-Vision-4.2B	2024-08-20	https://huggingface.co/microsoft/Phi-3.5-vision-instruct
LLaVA-1.5-7B	2023-10-05	https://huggingface.co/liuhaotian/llava-v1.5-7b
LLaVA-1.6-Mistral-7B	2024-01-30	https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf
Fuyu-8B	2023-10-27	https://huggingface.co/adept/fuyu-8b
Qwen2-VL-7B	2024-08-30	https://huggingface.co/Qwen/Qwen2-VL-7B

B Qualitative Examples.

Interactive process could improve the performance of leading LMMs. In Figure 6, we provide the qualitative results of different models. For the same question, Claude-3.5-Sonnet gives the correct answer C without human feedback, Gemini-2.0-Flash uses two rounds while OpenAI-o1 uses three rounds. It indicates that 1) even the SOTA models like OpenAI-o1 can not fully address the visual logic problem which is worse than Claude-3.5-Sonnet, 2) the responses can be corrected by human feedback which shows that the models have the capability of interpreting and incorporating the feedback into their reasoning, 3) Different models shows a different level of this capability. Additionally, we provide another example in Figure 7.

LMMs may not truly reasoning-They guess answers by elimination. In Figure 8, we find that the model will guess the answer when we only have four options, the model tends to guess answers. For the same question, we conduct twice runs and find that OpenAI-o1 could not solve this problem at the beginning, but two different answers were given in these two runs. In the first run, the model outputs D at the beginning while in the second run, the model outputs the A at the beginning. In the following rounds, we provide the same prompts to ensure the fairness comparison, one can see that based on the same prompt, it outputs the same answer C in the second round. The left run in the figure shows the correct answer in the third round while the right run in the figure shows the incorrect answer D. We continue to give the third feedback for round 4, and the right run finally gives answer B. It is obvious that when a problem cannot solved by a model, it will 1) outcome answer randomly, and 2) outcome the answer through an elimination approach. These results may indicate that LMMs may not always truly reason they may give the answer by guessing. Additionally, we provide another example in Figure 9 to illustrate that LMMs may guess answers when they can not solve the challenging problems.

LMMs still fail when the GT answer is not provided in the level 3 feedback. As discussed in the main submission, we include the GT answer in the level 3 feedback prompt to examine whether the model can generate the correct reasoning procedure that leads to the correct answer. When we remove the GT answer as in Figure 10, the model still fails to produce the correct answer, indicating its limited capability in solving challenging problems even when detailed feedback is provided as guidance.

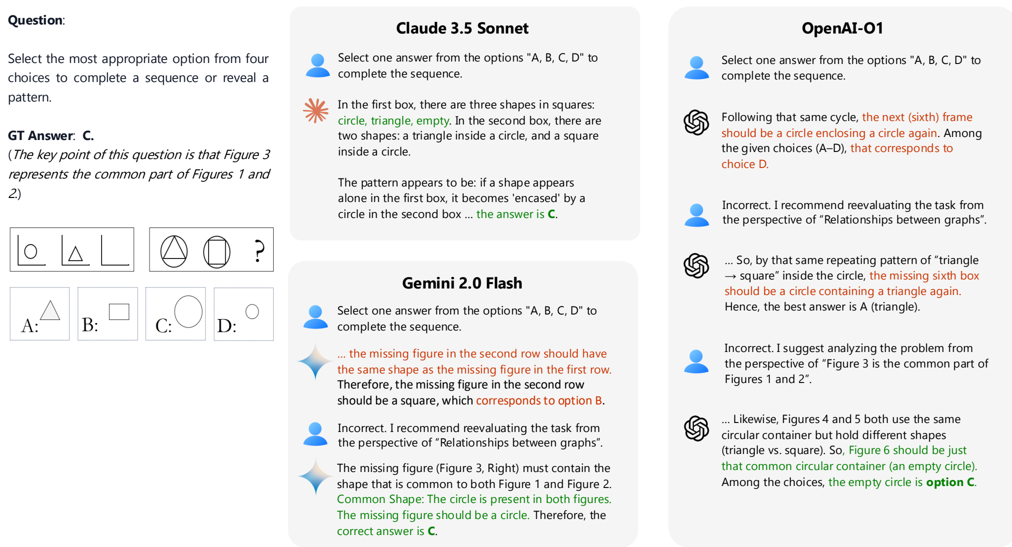

Figure 6: Qualitative results on different LMMs.

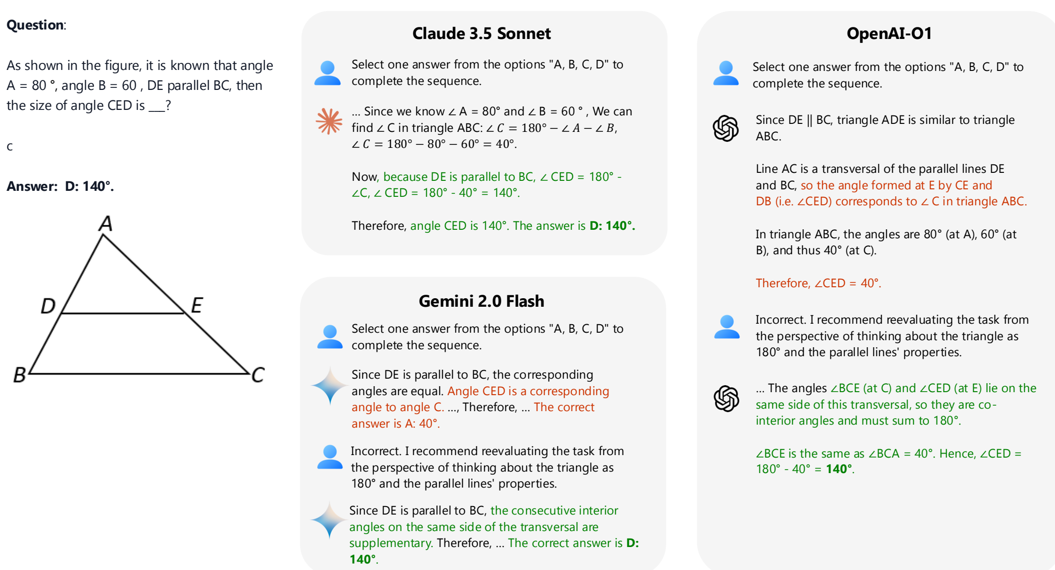

Figure 7: Qualitative results on different LMMs.

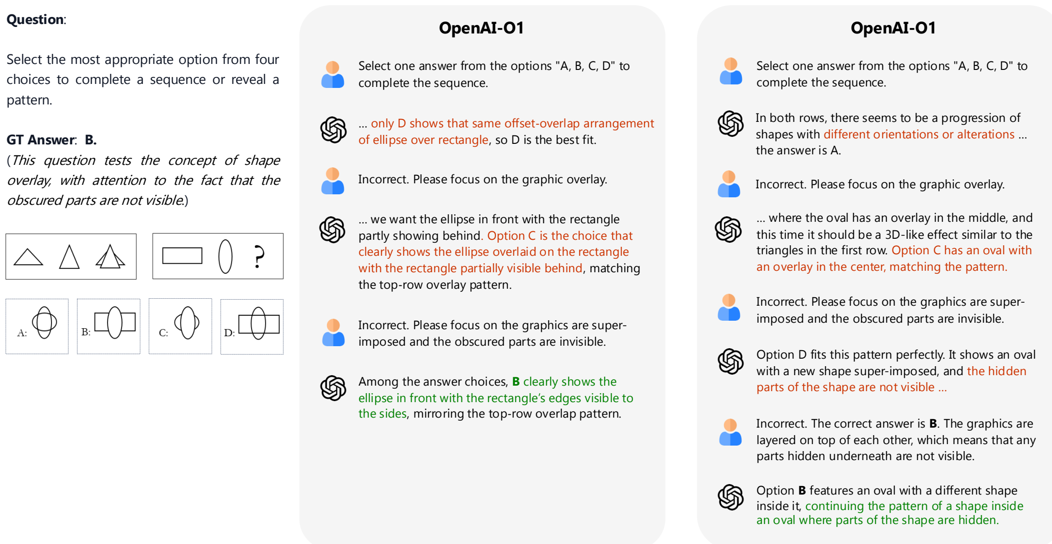

Figure 8: An example that model tends to guess answers.

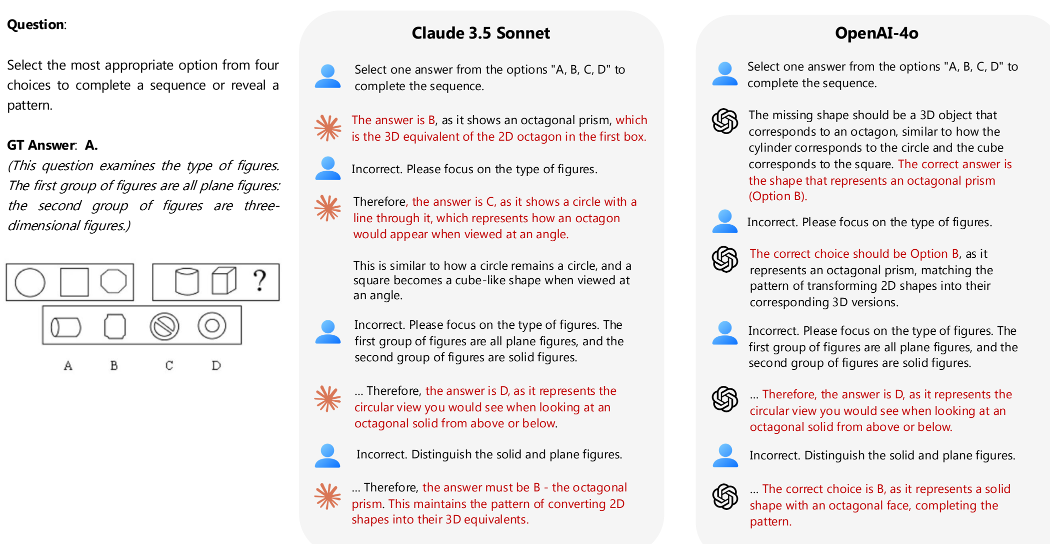

Figure 9: An example that model tends to guess answers.

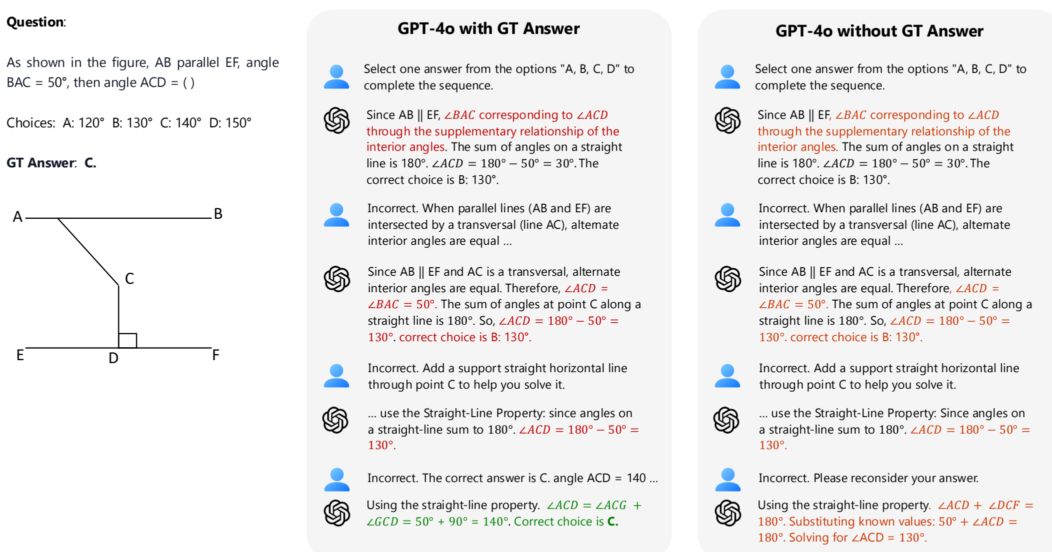

Figure 10: Qualitative results by removing GT answer in level 3 feedback.