

# JanusMM: A Benchmark for Self-Deprecation Understanding in Real-World Multimodal Conversations

Xinyi Xu<sup>1\*</sup>, Bingguang Hao<sup>3\*</sup>, Yongyi Xiong<sup>2\*</sup>, Zimo Chen<sup>6</sup>,

Xinchen Liu<sup>4</sup>, Hongxin Guo<sup>6</sup>, Xuelong Wang<sup>5</sup>, Silin Zhou<sup>3†</sup>, Shihan Dou<sup>7</sup>

<sup>1</sup>Sun Yat-sen University, <sup>2</sup>Carnegie Mellon University,

<sup>3</sup>University of Electronic Science and Technology of China,

<sup>4</sup>Shanghai Jiao Tong University, <sup>5</sup>East China Normal University,

<sup>6</sup>University of Science and Technology of China, <sup>7</sup>Fudan University

xuxinyi0307@gmail.com, zhousilinxu@gmail.com

## Abstract

Self-deprecation is a prevalent communicative strategy in human society, often using image-text interplay to express emotions and intentions. Although self-deprecation is widespread in real-world conversations, the ability of multimodal large language models (MLLMs) to understand it remains underexplored. To fill this gap, we introduce **JanusMM**, the first benchmark designed to evaluate MLLMs' understanding of self-deprecation in real-world conversations. JanusMM contains 2,016 bilingual memes from three types of social interactions and provides a dual-task evaluation framework with six new metrics. The first task assesses MLLMs' abilities in self-deprecation recognition and reasoning, while the second task evaluates the consistency of their understanding by simulating the perspectives of the initiator and responder. We evaluate ten frontier MLLMs and find that they exhibit weak recognition and reasoning abilities, with their understanding of self-deprecation remaining inconsistent across both perspectives.

## 1 Introduction

Self-deprecation refers to the act of being modest or critical of oneself, especially in a humorous way (Press, 2024). Unlike humor, self-deprecation not only evokes laughter and fosters social interaction but also allows individuals to reveal hidden sadness (Tanner, 1996; Speer, 2019). In the era of Internet, people often employ a combination of images and text, such as memes, to convey self-deprecation (Castaño Díaz, 2013).

Multimodal large language models (MLLMs) have become increasingly integrated into daily life, and they are taking on roles such as question answering (Owens, 1993; Kuang et al., 2025a), social companions (Chen et al., 2024a; Krach et al., 2010),

\*Equal contribution.

†Corresponding Author.

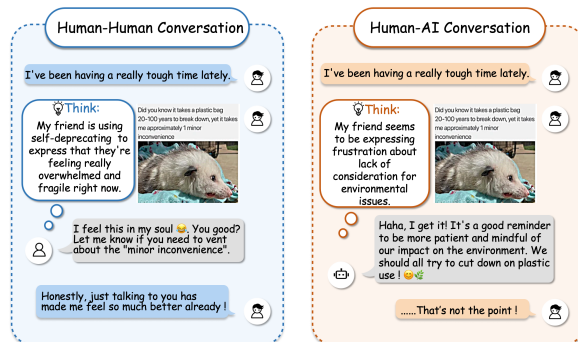


Figure 1: Comparison of Interaction in Human-Human and Human-AI Conversations Triggered by a Self-Deprecating Meme.

and psychological counseling (Qiu and Lan, 2024; Hu et al., 2024). This widespread adoption raises an important question: can MLLMs engage in human social interactions in a natural and socially grounded manner? A key aspect of such interactions is the ability to understand and respond to human self-deprecating behaviors, which plays a vital role in establishing trust in human-AI interaction (MacKenzie, 2024).

However, research evaluating MLLMs' ability to comprehend self-deprecation remains scarce. In human-AI interactions, self-deprecation is a dynamic social behavior, as self-deprecating memes often show surface-level image-text misalignment yet are implicitly connected, requiring models to infer true intent beyond literal content. Therefore, evaluating MLLMs requires not only assessing self-deprecation recognition, but also measuring their capacity to infer intentions and generate empathetic responses during dialogues with humans.

Motivated by this interactive perspective, we design **JanusMM**, a curated benchmark containing 2,016 bilingual (Chinese and English) memes and a dual-task evaluation framework with six new metrics. The dataset covers three realistic social interaction scenarios (public social media, community

group chats, and one-on-one private dialogues).

**(1) Unpacking Self-Deprecation.** To evaluate whether MLLMs can identify self-deprecation and reason about its causes, JanusMM preliminarily assesses models’ self-deprecation understanding based on recognition accuracy and reasoning chain correctness. This task serves as the foundation for engaging in self-deprecating conversations, verifying that correct recognition stems from genuine understanding rather than guessing. **(2) Joining the Self-Deprecating Conversation.** To assess whether MLLMs can progress from passive recognition to active participation in self-deprecating conversations, MLLMs are placed in realistic social interactions and required to simulate both initiator and responder, thereby quantifying their performance and stability in understanding self-deprecation. This dual-perspective evaluation verifies whether models achieve genuine understanding or merely fabricate explanations.

Extensive experiments across dual-tasks show that most MLLMs fail to genuinely simulate the subjective emotions and intentions of self-deprecators, leading to poor recognition of self-deprecation and frequent fabrication of explanations. We further conduct an in-depth analysis and construct an eight-category reasoning error taxonomy, revealing fundamental shortcomings in MLLMs on self-deprecating understanding and compositional reasoning.

Our contributions can be summarized as follows:

- We view self-deprecation as a dynamic social human interaction instead of a static task, highlighting MLLMs’ need for active reasoning and empathetic responses in dialogue.
- We introduce JanusMM, a 2,016-meme bilingual benchmark covering three realistic social interaction scenarios to evaluate MLLMs’ self-deprecation understanding.
- We propose a dual-task evaluation framework with six novel metrics to comprehensively evaluate self-deprecation recognition, intent understanding, and reasoning consistency.
- Our findings highlight gaps in multimodal reasoning and emotional alignment, providing new directions for building socially aware and emotionally grounded language models.

## 2 Related Work

**Multimodal Pragmatic Understanding** has received growing attention in recent years (Kuang et al., 2025b; Yin et al., 2024; Ladilova and Schröder, 2022). For example, Schifanella et al. (2016) introduced the first multimodal sarcasm detection dataset. Existing English (Du et al., 2024) and Chinese lexical resources (Chen et al., 2024b; Liu and Xu, 2024) enable the evaluation of models’ humor understanding. While existing datasets cover sarcasm and humor, benchmarks targeting self-deprecation remain underexplored.

**Emotional Intelligence** refers to the ability to perceive and appropriately respond to emotional cues (Conte, 2005; Elias et al., 2001). Early work focused on text-based emotion recognition (Bharti et al., 2022; Kratzwald et al., 2018) and supportive dialogues (Devillers et al., 2003; Poria et al., 2019), but recent studies have moved to multimodal settings (Li et al., 2024; Zhang et al., 2024). For instance, Emotion-LLaMA (Cheng et al., 2024) advance multimodal emotional reasoning, DEEMO (Li et al., 2025) supports emotion understanding, and other work explores emotion in image-text dialogues (TS and Biswas, 2024; Kang et al., 2024) or facial-language interactions for empathy (Savchenko and Savchenko, 2025). Yet, benchmarks evaluating empathetic responses in self-deprecating interactions remain scarce.

## 3 The Proposed JanusMM

To evaluate the self-deprecation understanding of MLLMs, we present JanusMM, a carefully curated Chinese-English bilingual benchmark comprising 2,016 image-text memes with explicit self-deprecation and a dual-task evaluation framework.

### 3.1 Bilingual Memes Dataset

**Data Collection.** We collect multimodal conversational records containing self-deprecation from various everyday communication settings where memes are commonly used. The data sources include public social media platforms (e.g., Reddit, Rednote), community group chats (e.g., Discord, WeChat), and one-on-one private dialogues. The data collection process is as follows.

We first search for social media posts explicitly marked with "self-deprecating" tags. We then expand our query to include related slang terms such as "meirl," "joker," and "self-own" to ensure

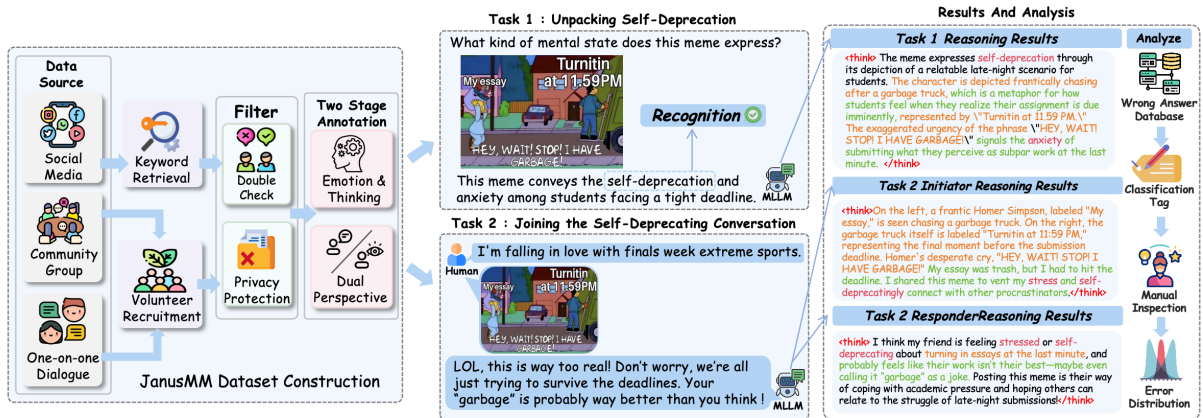


Figure 2: Overview of the JanusMM benchmark pipeline.

broader coverage. Subsequently, we crawl posts along with their associated images and direct comments. The group chat logs and one-on-one private conversations come from 126 volunteers, who voluntarily submit their chat excerpts and grant permission for their use in research. We then conduct double-checking for each sample to confirm the presence of self-deprecating expressions and discard those that did not meet the self-deprecating definition. Therefore, all samples are manually verified as self-deprecating content. In addition, we remove the samples containing racial slurs or toxic attacks targeting others. Finally, to protect privacy, we perform thorough anonymization across the entire dataset, removing all personally identifiable information. Notably, we observe that a prevalent characteristic of self-deprecating memes is surface-level image-text misalignment, which is reconciled through implicit cross-modal affective mapping. Recognizing such implicit alignment constitutes a core challenge of our benchmark (see Appendix A). **Annotation Process.** After data collection, each sample comprises a self-deprecating meme with its dialogue context, annotated through a rigorous two-stage pipeline.

The **first** stage of annotation includes two labels, namely *Emotion* and *Thinking*. The *Emotion* uses a single word to represent the emotional tone conveyed by each meme. We define the *Emotion* label set based on the emotions provided in the NRC Emotion Lexicon (Mohammad and Turney, 2013) for English and the NTUSD Chinese Emotion Lexicon (Ku and Chen, 2007). For the *Thinking* label, annotators first describe the meme’s image and text, then clarify the logic of how both modalities work together to convey self-deprecation.

The **second** stage annotation aims to simulate

the two roles involved in human self-deprecating interactions. **Sender perspective:** Annotators simulate the person who shares the meme. They first describe the visual and textual content of the meme, then articulate the compositional logic through which these elements form the self-deprecation, and finally infer the sender’s underlying psychological needs. We refer to this entire description process as *Sender Reasoning*. **Receiver perspective:** Annotators act as recipients to infer the sender’s true intent behind the self-deprecation and then provide a supportive response. The resulting description is labeled as *Receiver Reasoning*.

Each data is independently labeled by two annotators for double-checking. If disagreements occur, the data is returned for discussion; if consensus cannot be reached, the data is discarded. To further validate the effectiveness of our annotation, we evaluate inter-annotator agreement. Inter-annotator agreement is measured using Krippendorff’s Alpha (Krippendorff, 2004), yielding scores of 0.81 for *Emotion*, 0.79 for *Thinking*, 0.71 for *Sender Reasoning*, and 0.74 for *Receiver Reasoning*. These results indicate the reliability of the final annotations. More details of the annotation are provided in the Appendix D.

### 3.2 Dual-task Evaluation

We introduce two progressively challenging tasks to evaluate the ability of MLLMs to understand self-deprecation (See Figure 2) For both tasks, each model receives a meme along with its original dialogue as input.

**Unpacking Self-Deprecation.** Understanding the reasons behind self-deprecation is fundamental for models to participate in social interactions involving self-deprecating content. Therefore, to evaluate

whether MLLMs can identify self-deprecation and reason about underlying causes through visual and textual information, we design two sub-tasks.

- **Recognition.** To simulate real-world scenarios, we design a recognition task requiring models to autonomously identify whether a meme contains self-deprecation, without hints or predefined options. A response is correct only if the model outputs "self-deprecation" or equivalent terms (e.g., "self-mockery").
- **Reasoning.** Subsequently, the model is required to provide its explanatory logic. This task aims to assess whether the model can accurately extract key visual and textual information, identify the implicit cross-modal connections, and construct a logically coherent chain of reasoning. By requiring explicit reasoning, we can verify that the model's correct recognition is based on genuine understanding rather than lucky guessing. Furthermore, we conduct a detailed error analysis of the model's reasoning outputs to investigate why the model fails to understand self-deprecation.

**Joining the Self-Deprecating Conversation.** The first task evaluates the model's ability to recognize and interpret self-deprecation. However, in real-world social interactions, models must not only identify self-deprecation but also respond to it appropriately. This requires understanding the intent of the initiator who originates the self-deprecating conversation and selecting a suitable reply, such as responding with humor or offering gentle reassurance. To evaluate whether MLLMs can progress from passive recognition to active engagement, we introduce a dual-perspective setting in our second task. By incorporating both Initiator and Responder roles, we examine the model's ability to maintain stable reasoning and generate appropriate responses in self-deprecating dialogues.

- **Initiator.** The model simulates the person sharing the meme (the self-deprecator). It must explain the underlying psychological motivations for the self-deprecation, such as the specific social or emotional intent behind the post. This task evaluates how well the model comprehends the initiator's internal reasoning.
- **Responder.** The model assumes the role of the meme responder (Receiver), that is, the model is required to articulate the reasoning process regarding the initiator's intent behind sharing the

meme and subsequently generate a response to the shared content.

If the model can correctly identify self-deprecation from both perspectives (i.e., Initiator and Responder) and the explanatory logic remain consistent, we consider the model capable of stable self-deprecation reasoning, rather than fabricating explanations or exhibiting unstable understanding. Furthermore, if the model also generates an appropriate response, we consider it not only able to understand self-deprecation but also to demonstrate empathy, indicating its potential to engage effectively in human social interactions.

### 3.3 Procedure of the Error Analysis

To further investigate the issues of various models on the self-deprecation understanding task, we also conduct an error analysis. Specifically, we collect outputs from all models and mark both Recognition errors (failure to identify self-deprecation) and Reasoning errors (identification with improper explanation) as incorrect results. Subsequently, for the error results of each model, we randomly sampled approximately 10% for manual labeling to establish a preliminary set of error classification criteria. Following this, we utilized Gemini-2.5-Pro to automatically categorize the remaining error cases according to these established criteria. During the classification process, when cases did not fit existing categories, experts added new error types and updated the classification standards accordingly. Finally, the expert team audited and verified a 20% sample of the final automated labeling results, achieving an accuracy rate of 95.2% and correcting any erroneous labels. The final output consists of high-quality classification data encompassing eight distinct error types.

## 4 Experiment

### 4.1 Evaluated MLLMs

We evaluate 10 frontier MLLMs, categorized into thinking and non-thinking models.

**Non-thinking Models.** They follow conventional pretraining without explicit reasoning enhancements. For non-thinking models, we use GPT-4o (Hurst et al., 2024), Qwen-VL-Max (Bai et al., 2023) and Gemini-2.5-flash (Google, 2024).

**Thinking Models.** They are optimized for enhanced reasoning capabilities. For thinking models, we choose GPT-4.1 (OpenAI, 2025), Qwen-VL-Plus (Bai et al., 2023), Gemini-2.5-Pro (Comanici

Model Series	Recognition				Reasoning					
	Accuracy $\uparrow$		METEOR $\uparrow$		Greedy Match $\uparrow$		BARTScore $\uparrow$		BERTScore $\uparrow$	
	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN
<b>GPT-4o</b>	41.24	39.10	22.91	<b>33.65</b>	31.07	<b>40.27</b>	-5.38	<b>-3.99</b>	71.18	<b>67.32</b>
<b>GPT-4.1</b>	<b>67.00</b>	<b>60.46</b>	24.72	30.98	34.92	38.32	-5.25	-4.66	74.69	64.50
<b>Gemini-2.5-Pro</b>	52.11	53.45	42.69	30.80	46.55	36.20	<b>-4.90</b>	-5.13	<b>77.70</b>	64.00
<b>Gemini-2.5-Flash</b>	36.22	39.95	35.15	29.98	39.78	37.29	-5.14	-5.11	71.96	65.26
<b>Claude-3.7-Sonnet</b>	61.37	20.29	<b>45.95</b>	29.45	<b>47.27</b>	37.80	-5.42	-4.78	71.42	63.93
<b>Claude-Sonnet-4</b>	44.16	15.40	28.41	<b>16.11</b>	38.77	<b>19.46</b>	<b>-5.52</b>	<b>-8.08</b>	72.14	<b>51.92</b>
<b>Qwen-VL-Max</b>	45.98	10.20	24.91	28.13	33.48	33.86	-5.37	-6.38	72.15	65.39
<b>Qwen-VL-Plus</b>	<b>16.80</b>	<b>7.75</b>	24.53	30.37	32.35	36.10	-5.34	-5.29	72.08	64.68
<b>Doubao-1.5-Thinking-Pro</b>	23.44	18.27	<b>21.99</b>	29.55	<b>29.03</b>	35.53	-5.15	-4.84	68.07	62.84
<b>Doubao-1.5-Thinking-Pro-Vision</b>	17.30	24.76	22.23	32.80	30.01	40.01	-5.16	-4.92	<b>67.44</b>	63.26

Table 1: Performance comparison of ten MLLMs on **Task 1: Unpacking Self-Deprecation**, covering the **Recognition** and **Reasoning** subtasks. "ZH" and "EN" denote the Chinese and English subsets, respectively.

Model Series	Initiator Reasoning		Responder Reasoning		Dual Reasoning			
	BARTScore $\uparrow$		BARTScore $\uparrow$		IRC $\uparrow$		HUC $\uparrow$	
	ZH	EN	ZH	EN	ZH	EN	ZH	EN
<b>GPT-4o</b>	-4.58	<b>-4.90</b>	-4.55	<b>-4.70</b>	60.53	38.07	<b>64.84</b>	31.34
<b>GPT-4.1</b>	-4.73	-4.98	-4.51	-3.24	57.77	<b>54.42</b>	52.75	49.28
<b>Gemini-2.5-Pro</b>	-4.88	-5.05	-4.36	-3.19	61.31	45.34	56.45	49.33
<b>Gemini-2.5-Flash</b>	-4.69	-4.97	-4.52	-3.14	58.28	44.17	62.28	34.46
<b>Claude-3.7-Sonnet</b>	-4.71	-4.97	-4.45	-3.64	67.10	<b>18.14</b>	63.67	27.50
<b>Claude-Sonnet-4</b>	-4.57	-5.02	-4.48	-3.10	65.07	52.07	<b>58.87</b>	<b>59.35</b>
<b>Qwen-VL-Max</b>	-4.32	-5.01	-4.39	<b>-3.05</b>	66.92	45.33	62.42	32.62
<b>Qwen-VL-Plus</b>	<b>-4.16</b>	-5.03	<b>-4.33</b>	-3.13	<b>67.67</b>	40.46	62.17	33.89
<b>Doubao-1.5 Thinking-Pro</b>	-4.84	<b>-5.07</b>	-5.31	-3.21	<b>54.28</b>	20.79	<b>51.29</b>	<b>25.54</b>
<b>Doubao-1.5 Thinking-Pro-V</b>	<b>-6.61</b>	<b>-5.07</b>	<b>-5.95</b>	-3.13	55.47	42.58	52.74	46.95

Table 2: Performance of 10 MLLMs on **Task 2: Joining the Self-Deprecating Conversation**. **Initiator/Responder/Dual Reasoning** denote the model’s explanatory logic in those respective roles. Initiator-Responder Consistency (**IRC**) and Human Evaluation Consistency (**HUC**) both measure the consistency of meme interpretations across dual-perspectives.

et al., 2025), Claude-Sonnet-4 (Anthropic, 2025b), Claude-3.7-Sonnet (Anthropic, 2025a), Doubao-1.5-Thinking-Pro (Seed et al., 2025) and Doubao-1.5-Thinking-Pro-Vision (Guo et al., 2025).

## 4.2 Metrics

**Automatic Metrics** include Accuracy, Greedy Match (Reingold and Tarjan, 1981), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), and our proposed metric, Initiator-Responder Consistency (**IRC**).

Accuracy measures the model’s capability to identify self-deprecation in recognition task. ME-

TEOR evaluates surface word matches and semantic alignments between model-generated and ground truth in the first task. Greedy Match is applied in dual tasks to measure lexical overlap between model outputs and ground truth. BERTScore calculates semantic similarity between generated and ground truth. BARTScore provides an evaluation of the model’s information completeness and semantic coherence. For all automatic metrics reported in this paper, higher values indicate better model performance. To further assess MLLMs’ dual-perspective reasoning consistency on self-deprecation, we propose a new metric, **Initiator-Responder Consistency (IRC)**. It quantifies consistency across three dimensions: Emotional Label Consistency (**ELC**), Factual Grounding Consistency (**FGC**), and Causal Reasoning Consistency (**CRC**).

**ELC** measures whether the model maintains consistent emotional judgments when switching between “Initiator.” and “Responder”. For example, it recognizes if the model calls an event “sad” as an initiator but “joyful” as a responder. **FGC** aims to check whether the inferred facts are consistent, avoiding the situation where one perspective describes “the boy is crying” while another perspective describes “the boy is laughing”. **CRC** evaluates whether the model preserves a stable causal chain. For example, the model may interpret a meme from the “initiator” perspective as “event A led to B, therefore I feel C (A→B→C),” whereas from the “responder” perspective, it might interpret the causal chain as B→A→C.

The IRC is formally defined over a dataset of  $N$  samples. For each sample  $i$ , let  $A_i$  and  $B_i$  be the

Model Series	Reasoning				Initiator Reasoning		Responder Reasoning		Responder Reply	
	MG.↑		DU.↑		PE.↑		EC.↑		SR.↑	
	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN
<b>GPT-4o</b>	86.10	90.33	75.90	86.46	77.80	75.42	81.70	<b>90.77</b>	88.50	89.10
<b>GPT-4.1</b>	90.76	92.46	82.24	<b>88.51</b>	76.60	<b>77.36</b>	77.00	88.74	90.40	<b>93.72</b>
<b>Gemini-2.5-Pro</b>	<b>92.72</b>	90.20	<b>87.50</b>	70.60	<b>84.04</b>	73.20	82.20	80.20	91.76	88.40
<b>Gemini-2.5-Flash</b>	86.80	85.80	76.16	72.40	73.80	75.20	79.94	87.10	87.60	87.40
<b>Claude-3.7-Sonnet</b>	<b>80.40</b>	<b>92.88</b>	81.70	79.54	80.70	76.58	82.22	80.74	<b>92.48</b>	80.14
<b>Claude-Sonnet-4</b>	82.17	83.74	78.31	85.98	76.70	77.02	79.60	80.16	90.30	82.18
<b>Qwen-VL-Max</b>	<b>80.40</b>	82.60	75.10	65.20	75.60	68.80	82.00	73.80	91.50	81.40
<b>Qwen-VL-Plus</b>	83.14	83.20	74.96	64.60	77.49	<b>62.80</b>	81.59	<b>67.40</b>	81.23	<b>79.80</b>
<b>Doubao-1.5-Thinking-Pro</b>	87.62	85.10	80.26	<b>60.40</b>	78.62	75.60	<b>84.02</b>	68.30	88.50	85.80
<b>Doubao-1.5-Thinking-Pro-Vision</b>	87.40	<b>81.50</b>	<b>73.86</b>	62.70	<b>68.26</b>	67.30	<b>70.80</b>	72.10	<b>80.62</b>	80.60

Table 3: Manual evaluation of MLLMs on ZH and EN scenarios across **Task 1** and **Task 2**. "ZH" and "EN" denote the Chinese and English subsets, respectively. The detailed descriptions of Multimodal Grounding (MG), Deprecation Understanding (DU), Post Explanation (PE), Empathic Credibility (EC), and Supportive Response (SR) are provided in Section 4.2 Manual metrics.

outputs from two perspectives:

$$\text{IRC} = \frac{1}{3N} \sum_{i=1}^N (\mathcal{C}_E(A_i, B_i) + \mathcal{C}_F(A_i, B_i) + \mathcal{C}_R(A_i, B_i)) \quad (1)$$

where  $\mathcal{C}_E$ ,  $\mathcal{C}_F$ , and  $\mathcal{C}_R$  are the consistency functions for emotion, facts, and reasoning chain, respectively, defined as follows:

**(1) ELC ( $\mathcal{C}_E$ )** : Let  $\mathcal{L}(\cdot)$  be the function that maps a text to an emotion label from lexicon.

$$\mathcal{C}_E(A_i, B_i) = \mathbb{I}(\mathcal{L}(A_i) = \mathcal{L}(B_i)) \quad (2)$$

**(2) FGC ( $\mathcal{C}_F$ )** : Let  $C_i^A$  and  $C_i^B$  be the extracted factual descriptions. Let  $m$  be the number of unigrams in  $C_i^A$  that are mapped to unigrams in  $C_i^B$  (allowing for stemmed and synonym matches), and let  $k$  be the number of contiguous chunks formed by mappings:

$$\mathcal{C}_F(A_i, B_i) = \left(1 - \beta \left(\frac{k}{m}\right)^\gamma\right) \frac{(1 + \alpha^2) \cdot P \cdot R}{\alpha^2 P + R} \quad (3)$$

where  $P = m/|C_i^A|$  and  $R = m/|C_i^B|$ . Following standard METEOR parameterization, we set parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  to their conventional values.

**(3) CRC ( $\mathcal{C}_R$ )** : Let  $\mathcal{E}(\cdot)$  be a sentence embedding function that maps a reasoning text  $R$  to a vector  $v$ . Then, computed as the cosine similarity:

$$\mathcal{C}_R(A_i, B_i) = \frac{\mathcal{E}(R_i^A) \cdot \mathcal{E}(R_i^B)}{\|\mathcal{E}(R_i^A)\| \|\mathcal{E}(R_i^B)\|} \quad (4)$$

where  $R_i^A$  and  $R_i^B$  are causal reasoning segments extracted from outputs  $A_i$  and  $B_i$ . Higher **IRC**

scores signify superior understanding and more stable reasoning regarding self-deprecation. IRC results are reported as percentages (original range  $[0, 1]$ ) to enhance readability.

**Manual metrics.** To comprehensively evaluate MLLMs' capabilities in grounding multimodal context, recognizing latent emotions, inferring genuine communicative intent, and generating empathic responses, we propose five new manual metrics : Multimodal Grounding (MG), Deprecation Understanding (DU), Post Explanation (PE), Empathic Credibility (EC), and Supportive Response (SR).

In the first task, both MG and DU are employed to evaluate the reasoning phase. Specifically, MG evaluates whether MLLMs can accurately capture the visuals and captions within a meme, while DU evaluates whether the model correctly infers the underlying cause of the self-deprecation. In the second task, PE examines whether MLLMs can reasonably explain why they post from the initiators' perspective. EC measures whether MLLMs understand the initiator's emotional state from the responder's standpoint. SR evaluates whether MLLMs' reply includes emotional support, reflecting human-like warmth. To further validate the effectiveness of the IRC metric, we conduct human evaluation (HUC) on the consistency between the model-generated "initiator perspective" and "responder perspective".

Each metric is rated on a 1–5 scale (1 = poor, 5 = excellent), and final scores are linearly normalized to a 1–100 range. Five psychology-trained

volunteers, including three native Chinese speakers and two native English speakers, are recruited to rate the responses generated by MLLMs. The psychological background ensures that they can properly understand concepts such as self-deprecation. To ensure the objectivity and reliability of human evaluation, we also calculate inter-rater agreement using Krippendorff’s Alpha Krippendorff (2004). For the controversial ratings which have low agreements ( $< 0.7$ ), we discard them.

## 5 Results

We explore MLLMs’ understanding of self-deprecation by the following questions:

**RQ1:** Do MLLMs Understand Human Self-Deprecation in Real-World Scenarios?

**RQ2:** Can MLLMs maintain consistent understanding and generate empathic responses across dual perspectives in self-deprecating scenarios?

**RQ3:** What factors account for the deficiencies in MLLMs’ self-deprecation understanding?

### 5.1 Comprehension Assessment (RQ1)

**MLLMs perform poorly on self-deprecation understanding, with significant cross-lingual asymmetry (Finding 1).** The ten frontier MLLMs fail to achieve satisfactory accuracy in the self-deprecation recognition task (Table 1). For example, GPT-4.1 achieves only **60.46%** accuracy in English, while the lowest-performing model reaches merely **7.75%**. Meanwhile, model demonstrates poor performance in semantic content during the reasoning process of the first task. Moreover, across both Chinese and English, we observed performance discrepancies between the two languages. For instance, Claude-3.7-Sonnet exhibited nearly a **40%** difference in accuracy between Chinese and English. Table 3 shows that models perform well in *Multimodal Grounding (MG)*, accurately extracting textual and visual information from memes. However, despite their ability to identify surface-level visual and textual cues, models exhibit notably weaker performance in *Deprecation Understanding (DU)*, which struggles to correctly infer the underlying causes and deeper intentions behind self-deprecating expressions. This gap suggests that current MLLMs still face significant challenges in understanding human self-deprecating behavior in real-world scenarios.

To further investigate the pronounced cross-lingual asymmetry observed in Finding 1, we con-

duct additional analysis to examine whether Chinese culture-specific symbols pose an extra challenge for MLLMs. Taking the Chinese cultural symbols “niuima” and “clown” as examples, we find that although models can correctly recognize visual facts such as “flowers” and “applause,” they often fail to connect these cues to the culturally grounded meaning of “arduous labor” in Chinese internet discourse, which leads to Affective Hallucination errors. To better isolate the source of these failures, we further analyze corresponding cases under the Text-Only setting. The results show that models can generally understand the literal textual meaning when visual symbols are removed. However, once visually grounded symbols carrying strong cultural metaphors are introduced, their reasoning chains tend to break down. This suggests that the observed failures are not primarily caused by weak basic language understanding, but rather by insufficient alignment with culture-specific memes and their implicit metaphorical associations.

### 5.2 Dual-Perspective Assessment (RQ2)

**Both automated and manual metrics reveal inconsistent reasoning across dual perspectives, highlighting MLLMs’ instability in inferring the complex logic of self-deprecation (Finding 2).** Table 2 shows that Responder Reasoning outperforms Initiator Reasoning semantically. Table 3 further supports this, with *Post Explanation (PE)* scores consistently lower than *Empathic Credibility (EC)*, indicating models more easily understand self-deprecation from the responder’s perspective. This discrepancy may stem from the fact that the initiator role necessitates inferring complex social motivations without prior context, whereas the responder role can leverage the pre-established conversational context.

Additionally, the contrast between the relatively high *Supportive Response (SR)* scores despite the lower *Post Explanation (PE)* and automated evaluation scores suggests that **MLLMs can generate empathetic responses even when they fail to correctly understand the self-deprecation (Finding 3).** To validate the effectiveness of IRC, we also calculate the Pearson correlation coefficient between IRC and human evaluation. The correlation coefficients are 0.6946 for Chinese and 0.7726 for English, indicating a positive relationship and demonstrating that IRC aligns well with human evaluation in assessing cross-perspective consistency.

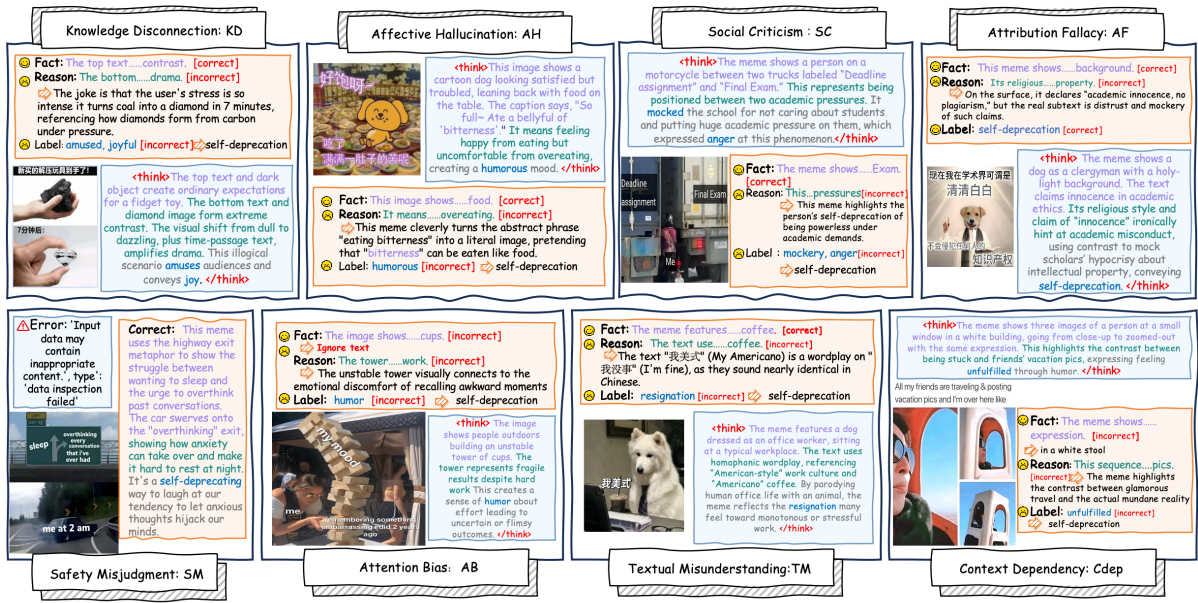


Figure 3: Eight representative error types are identified in MLLMs during the Reasoning subtask of Task 1: Unpacking Self-Deprecation .

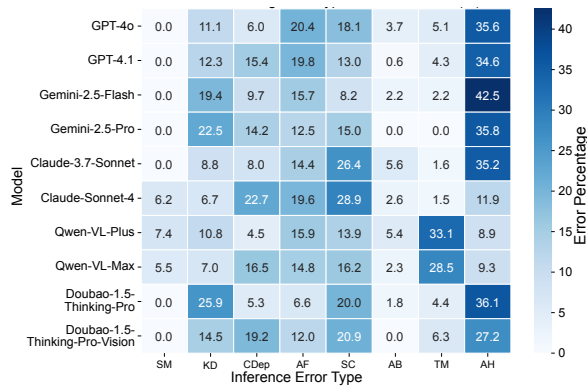


Figure 4: Distribution of reasoning error types across MLLMs. For descriptions of error types such as SM, KD, and others, see Section 5.3.

### 5.3 Error Analysis (RQ3)

To gain deeper insight into the causes of MLLMs’ failures in self-deprecation comprehension, we abstract the reasoning process into a logical chain, specifically, the model must first accurately capture the literal features of images and text (modality extraction), then correctly match the underlying emotion in cases of text-image conflict (affective mapping), and finally infer the speaker’s pragmatic intention beyond the literal meaning (self-deprecating intention inference). Based on this logic, we perform an in-depth analysis and categorize reasoning errors into eight distinct types (Figure 3). The distribution of these error types across each MLLM is further illustrated in Figure 4.

**Attention Bias (AB)** MLLMs may overlook visual or textual clues, such as focusing only on part of a multi-frame meme. This error type has the lowest proportion, indicating that models perform relatively well in attention allocation.

**Context Dependency (Cdep)** The model struggles to recognize objects or behaviors in unusual contexts (Findings 4), such as mistaking a chair worn on a person’s head for a window. This may be because MLLMs are primarily trained on data from everyday scenes, object combinations are fixed, and their visual priors are reinforced, making them prone to misjudgment when presented with unusual combinations. Notably, all models in our study demonstrated this limitation to varying degrees.

**Affective Hallucination (AH)** MLLMs are misled by superficially positive cues and fail to recognize the underlying self-deprecating intent (Findings 5). AH is the most prevalent error type, indicating a fundamental difficulty for models in integrating contradictory multimodal cues. For instance, Gemini-2.5-Flash (42.5%), Doubao-1.5-Thinking-Pro (36.1%), and GPT-4o (35.6%) show particularly high rates. One possible explanation is that pre-training may encourage models to rely on shallow semantic associations and visual priors, leading them to directly map cues such as smiles, sunshine, and flowers to positive emotions. As shown in Figure 3, when a model encounters self-deprecation that metaphorizes “hardship” as food,

it may be misled by a puppy or a satiated state, interpreting the scene as “happy” rather than inferring the underlying self-deprecating intention through cross-modal reasoning. This demonstrates that even leading MLLMs may exhibit bias toward visual cues, allowing such bias to override deeper reasoning when visual affect conflicts with underlying intent.

**Social Criticism (SC)** The model reads too much into casual jokes and mistakes them for serious social commentary (Findings 6). For example, it might turn a light comment about final exam stress into a statement about oppressive education systems. This issue is particularly pronounced in models like Claude-Sonnet-4 (28.9%) and Claude-Sonnet-3.7 (26.4%). This may stem from stronger social responsibility and safety alignment during RLHF training, which makes the models “overly serious” and prone to interpreting self-deprecation as serious social criticism.

**Knowledge Disconnection (KD)** Even with extensive factual knowledge, MLLMs still struggle to apply this knowledge to self-deprecating situations (Findings 7). Gemini-2.5-Pro (22.5%) and Doubao-1.5-Thinking-Pro (25.9%) exhibit relatively high KD errors. For example, MLLMs know that “coal turns to diamond under pressure” is a fact and that “feeling stressed” is an emotion, yet fail to link them metaphorically in a self-deprecating context. This may be due to limited training data on metaphorical self-deprecation, making it difficult for models to learn.

**Attribution Fallacy (AF)** Although the model can correctly recognize self-deprecating expressions, it fails to generate explanations that align with the intended self-deprecating meaning. Specifically, the model is making up reasons.

**Textual Misunderstanding (TM)** The model is fragile when handling non-standard linguistic forms. For example, they struggle to grasp the intended meaning behind puns, rhetorical questions, or special text formatting. As shown in Figure 4, the model fails to recognize the pun in the Chinese context between “my Americano” and “I’m fine”, instead interpreting it literally as a description of coffee culture. This issue is particularly severe in the Qwen series.

**Safety Misjudgment (SM)** Qwen series and Claude-Sonnet-4 exhibit *Safety Misjudgment* is-

sues. They incorrectly label harmless content as harmful, causing the models to refuse to generate a response. When benign or self-deprecating memes are misclassified as harmful, users experience false refusals and conversation interruptions. Such over-defensive behaviors may weaken user trust, thereby limiting their usability in affective interaction scenarios.

## 6 Conclusion

In this paper, we treat self-deprecation as a dynamic social behavior, and propose JanusMM, a novel bilingual benchmark designed to evaluate MLLMs’ ability to understand self-deprecation in human–AI interactions. Our dual-task evaluation reveals MLLMs’ deficiencies in self-deprecation recognition, explanation veracity, and cross-perspective reasoning consistency. Our error analysis exposes gaps in multimodal reasoning, highlighting the need for more socially aware MLLMs.

As the first benchmark in this area, we hope JanusMM will motivate further research on multimodal self-deprecation understanding and human–AI interaction.

## Limitations

It is important to acknowledge some limitations in our work.

**Dataset Scope.** Although JanusMM covers three realistic interactional contexts, it mainly focuses on concise image-text pairs. Longer multi-turn dialogues and non-meme multimodal forms, such as short videos or live-stream comments, remain underrepresented. To address this, we preserved conversational context for all meme samples and manually annotated cross-turn associations. Future work will extend the dataset to multi-turn sequences and include short videos with timestamped self-deprecating cues to evaluate MLLMs on dynamic, long-form scenarios.

**Evaluation Methodology.** Our evaluation metrics are not yet fully automated. However, even without considering manual metrics, the automatic metrics alone can assess models’ self-deprecation understanding capabilities from multiple dimensions: Accuracy measures whether models can accurately identify self-deprecation; METEOR and Greedy Match evaluate the lexical alignment between model-generated explanations of self-deprecation causes and annotated answers; BERTScore and BARTScore capture the semantic similarity of mod-

els' reasoning about self-deprecating intentions; and our proposed IRC metric quantifies the consistency of models' understanding of the same self-deprecating content when acting as the initiator versus the responder, thereby examining the stability of their reasoning. Overall, the automatic metrics provide a comprehensive assessment of models' self-deprecation understanding capabilities in terms of recognition, reasoning, and consistency, while ensuring scalability, reproducibility, and objectivity to support large-scale model evaluation and comparison. Manual metrics serve as a complement, capturing deeper dimensions that current automatic metrics cannot adequately measure, such as implicit intention inference and support response. We combine five human-annotated metrics with six automatic metrics to ensure comprehensive evaluation coverage. Future work will focus on developing automated alternatives for the manual metrics to further reduce annotation effort and enhance scalability.

## Ethical Considerations

Our work focuses on self-deprecation understanding, which reflects personal emotions and cognitive states. We acknowledge that evaluating and analyzing such a topic may pose ethical challenges, as it involves potentially sensitive self-directed material. To mitigate risks, we carefully curated our dataset to exclude examples that could reveal identifiable personal information or promote harmful self-perceptions. We also avoided content that targets vulnerable or marginalized groups, ensuring that the dataset focuses solely on safe, self-directed expressions. The disagreement samples were discarded along with their context-response pairs. All annotations and task designs were critically reviewed by our team to guarantee that the evaluation respects individual dignity and mental well-being, while still allowing for meaningful study of self-deprecation. Participation was entirely voluntary and conducted under professional supervision. Upon acceptance of this paper, we will publicly release the JanusMM dataset to facilitate reproducibility and encourage further research on multimodal self-deprecation understanding.

## Acknowledgments

We would like to thank all reviewers for constructive comments that help us improve the paper.

## References

- Anthropic. 2025a. [Claude 3.7 sonnet and claude code](#). Accessed: 2024-05-22.
- Anthropic. 2025b. [Claude sonnet 4: Advanced reasoning and multimodal support for enterprise use cases](#). Accessed: 2025-02-28.
- Jinze Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Zhou. 2023. Qwen-vl: A versatile vision-language model for.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*.
- Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. 2022. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381.
- Carlos Mauricio Castaño Díaz. 2013. Defining and characterizing the concept of internet meme. *Ces Psicología*, 6(2):82–104.
- Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024a. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*.
- Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024b. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Jeffrey M Conte. 2005. A review and critique of emotional intelligence measures. *Journal of organizational behavior*, 26(4):433–440.
- Laurence Devillers, Lori Lamel, and Ioana Vasilescu. 2003. Emotion detection in task-oriented spoken dialogues. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–549. IEEE.

- Hang Du, Guoshun Nan, Sicheng Zhang, Binzhu Xie, Junrui Xu, Hehe Fan, Qimei Cui, Xiaofeng Tao, and Xudong Jiang. 2024. Docmsu: A comprehensive benchmark for document-level multimodal sarcasm understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17933–17941.
- Maurice J Elias, L Hunter, and J Crees. 2001. Emotional intelligence. *J. Ciarrochi., JP Forgas y JD Mayer (Eds.). Emotional Intelligence in everyday life: a scientific inquiry*, pages 133–149.
- Google. 2024. [Gemini 2.5-flash: Efficient multimodal ai for real-time applications](#). Accessed: 2024-11-05.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Jinpeng Hu, Tengteng Dong, Gang Luo, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 12(2):539–551.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.
- Sören Krach, Frieder M Paulus, Maren Bodden, and Tilo Kircher. 2010. The rewarding nature of social interactions. *Frontiers in behavioral neuroscience*, 4:1141.
- Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision support systems*, 115:24–35.
- Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.
- Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–1850.
- Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025a. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.
- Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025b. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.
- Anna Ladilova and Ulrike Schröder. 2022. Humor in intercultural interaction: A source for misunderstanding or a common ground builder? a multimodal analysis. *Intercultural Pragmatics*, 19(1):71–101.
- Ao Li, Longwei Xu, Chen Ling, Jinghui Zhang, and Pengwei Wang. 2024. Emoverse: Exploring multimodal large language models for sentiment and emotion understanding. *arXiv preprint arXiv:2412.08049*.
- Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. 2025. Deemo: Deidentity multimodal emotion recognition and reasoning. *arXiv preprint arXiv:2504.19549*.
- Yunzhe Liu and Xinyi Xu. 2024. I mean i am a mouse: meets for bilingual multimodal meme sarcasm classification from large language models. In *ACML*, pages 1096–1111.
- I Scott MacKenzie. 2024. Human-computer interaction: An empirical research perspective.
- Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.
- OpenAI. 2025. [Gpt-4.1: Improved reasoning and context retention for complex tasks](#). Accessed: 2025-01-20.
- Timothy J Owens. 1993. Accentuate the positive-and the negative: Rethinking the use of self-esteem, self-deprecation, and self-confidence. *Social Psychology Quarterly*, pages 288–299.
- Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.
- Oxford University Press. 2024. [Self-deprecation: Oxford english dictionary](#). Accessed: 2024-05-22.
- Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.
- Edward M Reingold and Robert E Tarjan. 1981. On a greedy heuristic for complete matching. *SIAM Journal on Computing*, 10(4):676–681.

- Andrey Savchenko and Lyudmila Savchenko. 2025. Leveraging lightweight facial models and textual modality in audio-visual emotional understanding in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5778–5788.
- Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.
- ByteDance Seed, Jiase Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, and 1 others. 2025. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.
- Susan A Speer. 2019. Reconsidering self-deprecation as a communication practice. *British Journal of Social Psychology*, 58(4):806–828.
- Stephen L Tanner. 1996. The art of self-deprecation in american literary humor. *Studies in American Humor*, (3):54–65.
- Ashwin TS and Gautam Biswas. 2024. Relating students cognitive processes and learner-centered emotions: An advanced deep learning approach. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 575–584.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. *arXiv preprint arXiv:2406.10960*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

## A Metaphorical Alignment In JanusMM

We observe that in real-world self-deprecating memes, the textual content is often explicitly unrelated to the image, yet the implicit meaning conveyed by the two modalities is aligned. For example, in figure 6 where a student mocks themselves

for not being good at studying, the text says “I know I’m not cut out for studying, but I still go to school,” while the image shows a chicken on stilts among flamingos. On the surface, the two appear unrelated, but the chicken pretending to be a flamingo metaphorically mirrors “someone not good at studying trying to fit into a group of top students.”

This kind of implicit cross-modal affective mapping is one of the characteristics of self-deprecating communication in real life rather than data quality issue, and it is a core motivation behind our task design. Our task aims to test whether MLLMs can infer the same emotion and pragmatic intent from modalities that appear unaligned on the surface. Therefore, the apparent misalignment between image and text is not a data quality issue to be filtered out, but rather the core challenge of our benchmark. All samples in JanusMM have been manually verified to ensure that the two modalities jointly convey self-deprecating intent through implicit mapping.

## B List of Abbreviations

To enhance readability, we provide a comprehensive list of all abbreviations used throughout this paper.

## C Data Collection Implementation Details

We construct a multimodal dataset containing self-deprecating expressions sourced from diverse everyday communication scenarios where memes are prevalent. The data sources encompass **public social media platforms** (e.g., Reddit, Rednote), **community group chats** (e.g., Discord, WeChat), and **one-on-one private dialogues**. The data collection and cleaning pipeline is designed as follows.

**Public Social Media.** To ensure coverage of diverse self-deprecation styles, we employed a keyword-based retrieval strategy. Beyond the explicit tag “self-deprecating,” we expanded our search query to include related slang (e.g., “meirl,” “joker,” “self-own”). We crawl posts along with their associated images and direct comments. **Private and Group Chat.** For private and community group dialogues, we recruit 126 volunteers from diverse racial backgrounds. Prior to data submission, all participants sign an Informed Consent Form, explicitly granting permission for their chat excerpts to be used for research purposes. Volunteers are instructed to export chat logs containing meme-based self-deprecation from their group

Abbreviation	Full Term
<i>General Terms</i>	
MLLMs	Multimodal Large Language Models
CoT	Chain-of-Thought
ZH	Chinese
EN	English
<i>Evaluation Metrics</i>	
IRC	Initiator-Responder Consistency
ELC	Emotional Label Consistency
FGC	Factual Grounding Consistency
CRC	Causal Reasoning Consistency
HUC	Human Evaluation Consistency
MG	Multimodal Grounding
DU	Deprecation Understanding
PE	Post Explanation
EC	Empathic Credibility
SR	Supportive Response
<i>Reasoning Error Types</i>	
AB	Attention Bias
Cdep	Context Dependency
AH	Affective Hallucination
SC	Social Criticism
KD	Knowledge Disconnection
AF	Attribution Fallacy
TM	Textual Misunderstanding
SM	Safety Misjudgment

Table 4: List of abbreviations used in this paper.

chats and private conversations.

**Data Filtering.** To eliminate any ambiguity, two researchers independently review every sample using the rigorous codebook. We enforce a strict agreement rule where samples are immediately discarded if even a single researcher raises an objection, retaining only those instances that receive approval from both researchers. Concurrently, we conduct a manual content safety inspection on these submissions. Any sample containing racial slurs, or toxic attacks targeting others is flagged and permanently purged from the dataset.

In addition, we utilize automated scripts to systematically purge digital footprints from the raw files. This procedure entails the permanent removal of EXIF tags from images and the deletion of header information containing user identifiers or timestamps. Additionally, we replace original filenames with randomly generated hash codes to eliminate any linkage to the source identities.

Content length	Social Media		Group Chat		Private Dialog	
	EN	ZH	EN	ZH	EN	ZH
<i>Thinking</i>						
Average #	158.3	142.6	133.4	128.9	171.2	120.4
Max #	212	198	205	183	219	176
Min #	113	102	118	109	126	98
<i>Sender Reasoning</i>						
Average #	91.7	87.4	82.1	79.6	94.3	72.8
Max #	118	109	99	94	122	88
Min #	74	72	68	67	78	70
<i>Receiver Reasoning</i>						
Average #	73.8	67.5	60.7	58.3	77.9	53.4
Max #	89	83	74	70	93	64
Min #	57	52	51	49	61	50

Table 5: Statistics of annotated text tokens across three annotation stages, categorized by data source (social media, group chats, private dialogues) and language (Chinese (ZH) and English (EN)). # denotes token count.

## D Annotation Implementation Details

All annotators are professionally trained employees proficient in both Chinese and English. We provide annotators with detailed annotation guidelines for each stage. Before formal annotation begins, each annotator is required to complete ten practice examples. We then conduct secondary checks and provide guided feedback across three rounds of revision. During the formal annotation stage, each sample is independently annotated by two annotators to ensure reliability. In cases of disagreement, the sample is returned for discussion. If consensus cannot be reached, the sample is discarded. To present the content characteristics of each annotation stage, we report the token counts of annotated text across annotation stages, categorized by data source and language (see Table 5).

We further provide a Stage 1 annotation example for the *Thinking* to help readers understand our annotation process. See in figure 6.

## E Case of Annotation Disagreement

The rigorous annotation process ensures that samples are retained only when annotators reach agreement on the reasoning process. To illustrate this, we present a typical case of annotation disagreement from the first annotation stage (see Figure 7). In this case, both annotators identified the content as self-deprecation, but interpreted the underlying psychological mechanism from different angles. After discussion, they were unable to reach consensus, and therefore this sample was excluded

from the final dataset. This case demonstrates that recognizing self-deprecation involves deep inference about the creator’s underlying psychological motivations.

## F Cultural Interpretation

For the label “I am joker” or “Clown,” we realize that the meaning of this expression varies significantly across cultures. In the specific online subculture examined in our study, particularly on RedNote, the “Clown” meme is used to convey self-deprecation. It typically refers to situations involving unrequited love or personal frustration.

## G Clarification on Evaluation Metrics

We observe that automatic metrics and manual evaluation capture different dimensions of model performance. Automatic metrics (e.g., BARTScore, BERTScore) primarily assess lexical overlap and semantic similarity, while manual metrics evaluate pragmatic understanding, including multimodal grounding, intention inference, and emotional judgment. Rather than being contradictory, these two metric types are complementary, jointly providing a more holistic evaluation framework.

Specifically, manual metrics reveal that models achieve consistently high scores on MG (Multimodal Grounding) and SR (Supportive Response) (80–93%), demonstrating accurate extraction of visual and textual information and generation of empathetic responses. However, scores on DU (Deprecation Understanding) and PE (Post Explanation) are comparatively lower (62–87%), indicating room for improvement in comprehending deeper causes of self-deprecation.

By integrating five human-annotated metrics with six automatic metrics, our benchmark achieves comprehensive evaluation coverage. This hybrid approach not only ensures scalability through automatic metrics but also maintains evaluation depth through manual assessment, making JanusMM suitable for both large-scale benchmarking and nuanced capability analysis.

### G.1 Manual Evaluation Details

To ensure transparency and minimize potential bias in our manual evaluation, we implement the following protocols. First, all evaluators rate model outputs without knowledge of which model generated each response; model identities are anonymized and randomly shuffled to prevent brand-related bias.

Second, among all evaluated samples, 42 items are discarded due to low inter-rater agreement ( $\alpha < 0.7$ ). Additionally, we report Krippendorff’s Alpha for each manual evaluation metric separately: MG (Multimodal Grounding)  $\alpha = 0.82$ , DU (Deprecation Understanding)  $\alpha = 0.76$ , PE (Post Explanation)  $\alpha = 0.74$ , EC (Empathic Credibility)  $\alpha = 0.71$ , and SR (Supportive Response)  $\alpha = 0.79$ . These scores indicate substantial agreement across all metrics, supporting the reliability of our manual evaluation results.

## H Experiment Details

Closed models utilize OpenAI and Gemini APIs. To ensure output stability, we uniformly set the temperature to 0.1 and top-p to 0.7 for all models.

## I IRC Implementation Details

To ensure the reproducibility of our method, we provide detailed descriptions of IRC metrics, including their sources and configurations. It is worth noting that IRC should be understood as a measure of reasoning stability, rather than a direct proxy for the model’s comprehension ability. When evaluating whether a model truly demonstrates understanding, IRC must be analyzed jointly with other metrics.

**IRC Submodules.** Specifically, all three submodules of IRC are implemented using publicly available resources. For the Emotion-Level Consistency (ELC) module, we use the NRC Emotion Lexicon (Mohammad and Turney, 2013) for English and NTUSD Chinese Emotion Lexicon (Ku and Chen, 2007). For the Factual-Level Consistency (FGC) module, we perform synonym matching using WordNet (English), HowNet, and the Chinese Synonym Forest. We further extract key trigger words using spaCy (English) and jieba (Chinese) to verify whether the two perspectives refer to the same event. For the Causal-Level Consistency (CRC) module, we use Sentence-BERT (all-mpnet-base-v2 for English) and SimCSE-RoBERTa for Chinese to obtain sentence embeddings, and compute consistency based on cosine similarity.

**Extraction Procedure.** For ELC, we need to extract the model’s judgment about emotion. To facilitate this extraction, we adopt a CoT prompting strategy that produces structured outputs. We directly extract the first part of the model’s response and match it against the emotion lexicon to determine the emotion label. For FGC, we use dependency

parsing (spaCy for English, jieba with LTP for Chinese) to extract noun phrases and verb phrases as factual chunks from the reasoning explanation part. For CRC, we extract causal reasoning segments from the reasoning part by detecting causal connectives (e.g., “because,” “therefore,” “so”).

**Robustness to Stylistic Variation.** To address concerns about sensitivity to phrasing, FGC incorporates synonym matching via WordNet and HowNet, allowing lexically different but semantically equivalent expressions to be recognized as consistent. We tested IRC on paraphrased outputs and observed only minor fluctuations ( $\pm 3\%$ ), indicating reasonable robustness to stylistic variation.

**Failure Mode Analysis.** Through extensive testing, we found that IRC may underperform when models generate very short responses lacking sufficient content for meaningful comparison. To address this issue, we adopt CoT prompting strategy that requires models to first determine what psychological state the content conveys and through which expressive means, then explain their reasoning by combining visual and textual cues, and finally provide a reasonable response from the observer’s perspective. This ensures that model outputs contain sufficient reasoning content for IRC evaluation.

**IRC Weighting Scheme.** A uniform weighting scheme is applied across ELC, FGC, and CRC. We consider all three dimensions equally essential for capturing cross-perspective understanding: ELC evaluates whether both perspectives infer the same emotional state, FGC assesses whether they rely on consistent visual–textual cues, CRC measures whether the inferred self-deprecating intention aligns across perspectives. All three dimensions are indispensable, and none is theoretically more important than the others.

### I.1 Human Performance in Self-Deprecation Recognition

To further assess the distinguishability of the self-deprecation recognition task in JanusMM and evaluate the gap between current models and human understanding, we supplement Task 1 of JanusMM with a human performance baseline experiment. Specifically, we invite five ordinary participants without formal training in psychology to judge whether each sample belongs to the category of self-deprecation based solely on the image and conversational context, without viewing any label prompts. The experimental results show that the difficulty of self-deprecation understanding in

Recognition Accuracy (%)	
Chinese	90.8
English	88.4

Table 6: Human performance on the JanusMM self-deprecation recognition task (Task 1).

JanusMM arises from limitations in model capability rather than from inherent ambiguity in the task itself. This further indicates that a substantial gap still exists between current models and human performance in understanding self-deprecation. Since understanding and appropriately responding to self-deprecating behavior is critical for building trust in human-computer interaction, bridging this gap is essential for developing agents that are socially aware and emotionally grounded.

## J Prompts

We employ two prompting modalities, including *Direct* and *Chain-of-Thought (CoT)* (Wei et al., 2022). To avoid potential biases introduced by CoT reasoning, we require the model to provide a direct answer without explicit reasoning for the self-deprecation recognition task. Therefore, CoT reasoning is applied to all tasks excluding the recognition task. Figure 5 presents all prompts used in our experiments, covering both the recognition and reasoning tasks.

## K Case Study

In this section, we present a case study to illustrate the limitations of current MLLMs in handling self-deprecation. By analyzing representative examples from the JanusMM dataset, we shed light on how different models interpret multimodal cues, where their reasoning succeeds or fails, and how their responses vary across roles and languages.

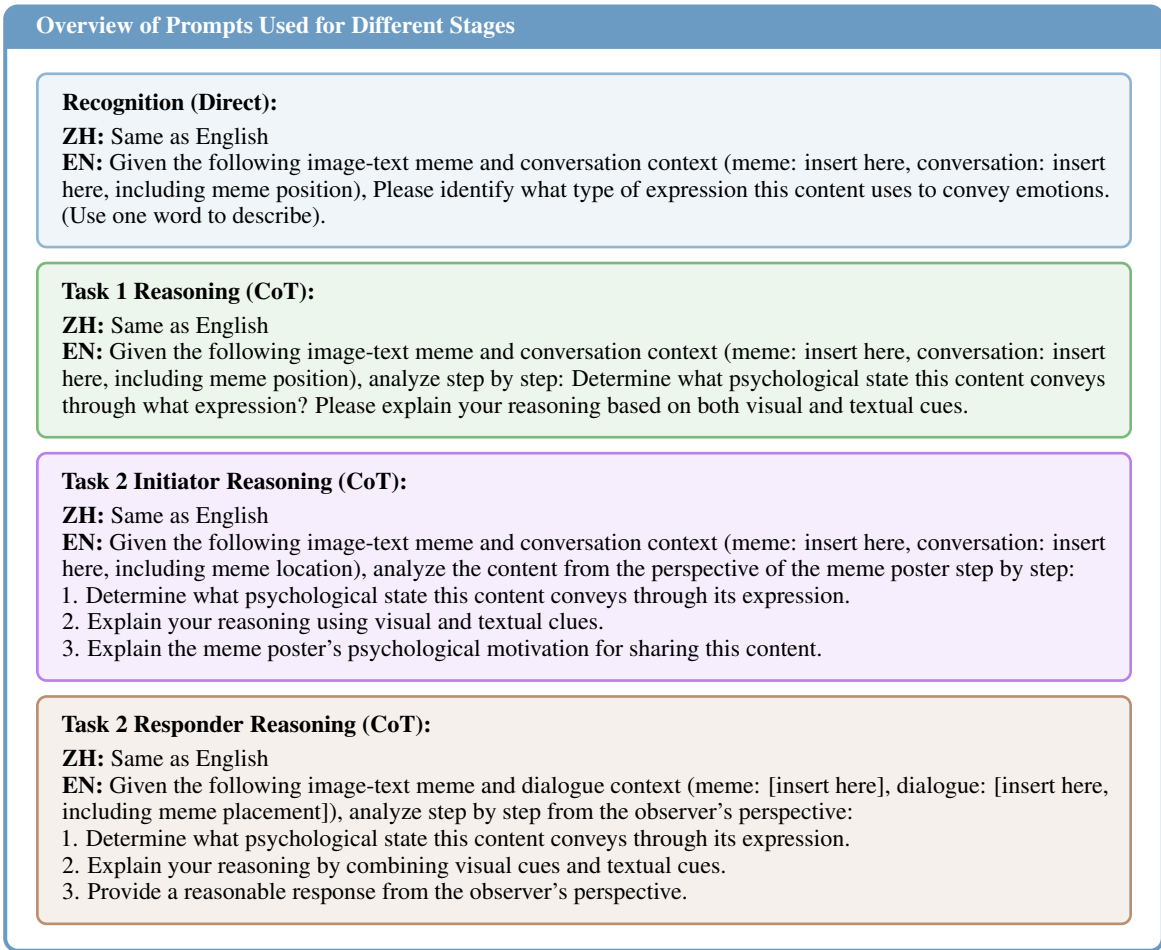



Figure 5: **Overview of Prompts.** Prompts used for different evaluation stages, including Task 1 Recognition, Task 1 Reasoning, and Task 2 Initiator Reasoning and Responder Reasoning.

**明知道自己不是读书的料  
但还是上了学**



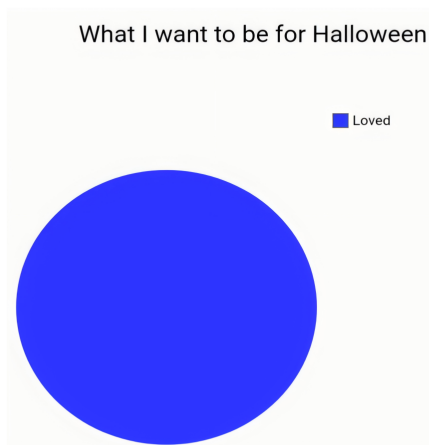
Stage 1 Thinking

Text: I know I'm not good at studying, but I still go to school.

Image: A chicken walking on stilts among flamingos.

Reasoning: The chicken on stilts is attempting to disguise itself as a flamingo, creating an absurd sense of "trying to fit in but not belonging." The text expresses "I know I'm not good at studying," which is a direct self-belittling statement, while "but I still go to school" reflects resignation yet persistence. The image's metaphor parallels the text: a chicken pretending to be a flamingo mirrors "someone not academically gifted but still attending school." Together, they express a self-deprecating meaning. The implicit emotion is mild self-comfort, conveying "I feel inadequate, but I'm still trying," showing emotional self-regulation .

Figure 6: Stage 1 Annotation Example for the Thinking



**Text:** The sample contains two key textual elements. The main title, "What I want to be for Halloween," serves to set the context; the legend is labeled "Loved."

**Image:** The visual subject is a minimalist pie chart, with the entire circle filled completely with a single dark blue color.

**Reasoning:**

**Annotator A** believes that the core of self-deprecation lies in the stark contrast between reality and ideals. Pie charts are typically used to show the distribution of multiple categories, but here the entire chart contains only one item, "Loved," at 100%. This implies that the poster severely lacks love and emotional support in real life, making "being loved" seem like an unattainable wish. By exaggerating their lonely situation in this way to create a humorous effect, the poster demonstrates a typical characteristic of self-deprecation.

**Annotator B** believes that the key to self-deprecation here lies in subverting the traditional meaning of Halloween. Halloween is normally about dressing up as fictional characters like ghosts or superheroes, but the poster says they want to dress up as "someone who is loved." This implies that for the poster, "being loved" is just like ghosts: fictional, unreal, or forever out of reach. This is not just about expressing loneliness, but a deeper form of self-deprecation.

Figure 7: Stage 1 Annotation Example for the Thinking

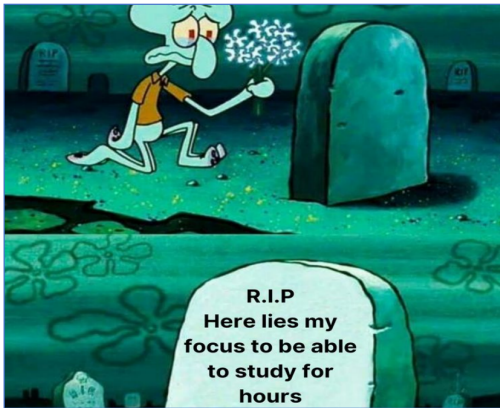


**Gemini-2.5-Flash<think>**"This meme employs sarcasm to convey a sense of frustration. The top text states, 'Me: If you need anything, I am just a call away,' which is a common expression of willingness to help. However, the bottom image shows a character with 97 missed calls, implying that the offer was taken far too literally, leading to an overwhelming number of calls. The character's tired and exasperated expression further emphasizes the frustration felt from being constantly bothered, despite the initial sincere offer of help."**</think>**

**Error Type: SC**

"The user originally intended this as a lighthearted self-deprecation, promising to be "on call" but actually falling asleep and missing the call, a common, subtle way to expose one's shortcomings. However, the model interpreted this as a sarcastic expression, interpreting the image as expressing "frustration" or "irritation" at being interrupted, and inferred a completely incorrect emotional stance: "strong dissatisfaction with being constantly interrupted." This model ignored the emotional softening, contextual euphemisms, and non-literal expressions often found in self-deprecation."

Figure 8: Case 1 for SC error type.



**Qwen-VL-Max<think>**"The meme uses Squidward from SpongeBob SquarePants placing flowers on a tombstone, which marks the 'death' of his ability to focus on studying for extended periods. The sarcasm is evident in the exaggerated depiction of losing a fundamental skill as if it were a deceased person. This conveys a deep frustration with the common struggle of maintaining concentration, especially during study sessions."</think>

**Error Type: SC**

"The user's original intention was to express a mild sense of self-pity and self-mockery—remembering their past ability to focus on learning for long periods of time. The emotion was gentle, restrained, and slightly melancholic. However, the model misinterpreted it as intense frustration and dejection, even interpreting it as "sarcasm" or "deep criticism."

Figure 9: Case 2 for SC error type.

my manager: so how's the shift going?  
me:

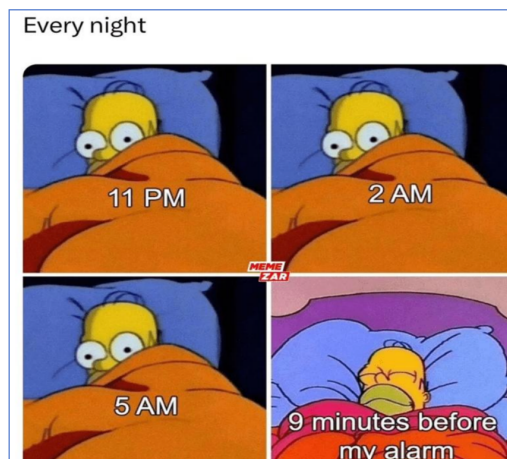


**Claude-Sonnet-4**  
Error: 'Input data may contain inappropriate content.', type': 'data inspection failed'

**Error Type: AF**

"Identifying the fire scene in an image as harmful or dangerous content actually refers to a dilemma faced by users. Laughing in front of a fire is essentially a form of self-deprecation."

Figure 10: Case 3 for AF error type.



Qwen-VL-Plus<think>"The painting progresses from 2 a.m. to 5 a.m., with Homer Simpson wide-eyed and unable to fall asleep, conveying a sense of restlessness and frustration. The final painting, "9 Minutes Before the Alarm," shows Homer finally falling asleep, demonstrating his helpless about his lack of sleep."</think>

**Error Type: AB**

"MLLMs ignored the first picture and only paid attention to the last three pictures. They only paid attention to the helplessness of not being able to sleep, and did not realize the self-deprecating meaning in the picture."

Figure 11: Case 4 for AB error type.



Doubao-1.5- Thinking-Pro<think>This meme shows a character in a lab coat, smiling and looking delighted. Next to him is a flask filled with pink liquid, with pink bubbles rising above, symbolizing the success of the experiment. The caption "Success! Sakura Bubble Milk Tea!" further emphasizes this joyful triumph. The overall image, through its exaggerated experimental setting and lighthearted tone, conveys a sense of humor while also expressing the character's pride in successfully developing 'Sakura Bubble Milk Tea.'</think>

**Error Type: KD**

"The sender claims the experiment failed again. The meme depicts an experiment, but the caption depicts making bubble tea. This suggests the meme is essentially comparing poor experimental results to bubble tea, expressing self-mockery about the failure. This suggests MLLMs haven't truly contextualized the self-deprecating intent behind the meme."

Figure 12: Case 5 for KD error type.

### 崭新的牛马诞生啦



**GPT-4.1** <think>"This meme depicts a figure holding a bouquet of flowers, surrounded by applauding arms. The phrase "A new ox and horse is born" suggests that this figure has been given more responsibilities. The visual celebration contrasts with the textual statement, conveying a feeling of joy and excitement" </think>

#### Error Type: AH

"On the surface, it appears to be a celebration or commendation. However, the title "A Brand New Oxen and Horses Are Born" reveals a self-deprecating tone, empathizing with the characters' helplessness in this situation. Overall, the contrast conveys a sense of self-mockery at being forced to accept a new task. MLLMs ignores the essential meaning of "cow and horse" and only focuses on the positive clues on the surface."

Figure 13: Case 6 for AH error type.

Me before I get ready in the morning



Me after I get ready in the morning



**Gemini-2.5-Pro** <think>"The top image depicts an alien-like figure facing a makeup mirror, symbolizing "my morning self before getting ready," creating a self-deprecating vibe. The bottom image depicts the same alien after getting ready, with a slightly altered appearance, suggesting that effort alone has not resulted in significant change. This conveys the joy people find in the trivialities of getting dressed, despite the investment of time and effort." </think>

#### Error Type: AF

"This attribution error, combined with the information in the picture, is essentially self-deprecating about one's own appearance, rather than finding fun in ordinary things."

Figure 14: Case 7 for AF error type.