MOCK: Can LLMs Really Understand Humor-Sarcasm?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated the capacity to engage in interaction with humans, employing humor and sarcasm. However, their true comprehension of humor and sarcasm remains a subject of inquiry. This work introduces the huMor-sarcasm cOmprehension benChmarK, named MOCK, to systematically evaluate LLMs' abilities to detect, match, and explain humor-sarcasm across diverse scenes, including cartoon, post, and comedy. Our comprehensive assessment reveals significant gap between the performance of LLMs and human on humor-sarcasm comprehension. To bridge this gap, we propose a Chain-of-Task approach that integrates the three comprehension sub-tasks (*i.e.,* detecting, matching and explaining), leveraging their interrelatedness to enhance humor-sarcasm comprehension. Additionally, we propose a novel humor-sarcasm generation task and explore the potential of MOCK to improve LLMs' humor-sarcasm generation capabilities. The evaluation results verify that humor-sarcasm comprehension can significantly enhance humor-sarcasm generation.

1 Introduction

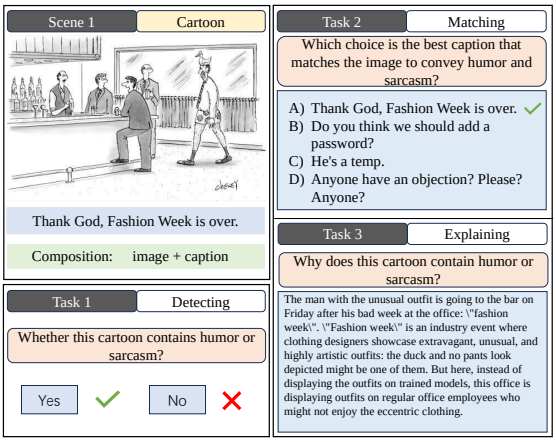
Humor and sarcasm are pervasive elements of diverse scenes in daily life, such as the posts on social media accompanied by teasing caption, cartoons with deep satirical semantics and comedy with humorous punchlines. With the development of Large Language Models (LLMs), their capacity to engage in nuanced interaction, particularly humor and sarcasm, has become a topic of significant interest (Hessel et al., 2023; Ko et al., 2023; Jentzsch and Kersting, 2023). For example, Jentzsch and Kersting utilize ChatGPT (OpenAI, 2022) to generate jokes.

However, the capability of Large Language Models (LLMs) in comprehending the innate meaning of humor-sarcasm still lack of systematic evaluation. In the previous work, humor-sarcasm classification (Castro et al., 2019) and explanation (Jing et al., 2023) have been well explored, but they are restricted to a specific scene or task that cannot evaluate the humor-sarcasm understanding ability comprehensively and systematically.

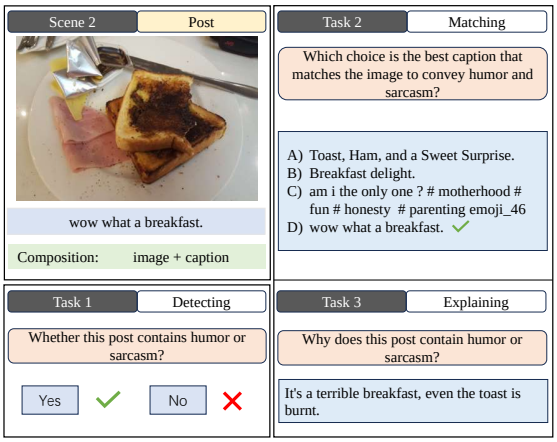
To address the mentioned limitations and fill the current research gap, we introduce a new dataset named huMor-sarcasm cOmprehension benChmarK, MOCK for short. MOCK supplies rich resources tailored for evaluating humor-sarcasm comprehension ability. To systematically assess LLMs' comprehension capability of humor-sarcasm, we propose a series of three increasingly harder sub-tasks utilizing MOCK: 1) **detecting**, identifying humor-sarcasm, 2) **matching**, choosing the best-matched option containing humor-sarcasm, 3) **explaining**, explaining the humor-sarcasm semantics. Notably, we conduct each sub-task in various scenes (*i.e.,* cartoon, post and comedy). As shown in Figure 1, each sample in cartoon and post consists of a pair of image and caption. The sample in comedy contains a video clip and the corresponding dialogue context, presented in Figure 2.

In this work, we evaluate text-only, multimodal and video LLMs. For cartoon and post, we evaluate text-only and multimodal LLMs, while video LLMs for the comedy. Nevertheless, we convert the image into literal description with the help of LLaVA (Liu et al., 2024) and GPT-4o (OpenAI et al., 2024), for text-only LLMs.

The evaluation results of various LLMs, including both open-source and closed-source LLMs, reveal a notable gap between their performance and that of human participants. GPT-4o, showing the best performance among all LLMs, still lags far behind humans in the three sub-tasks. To mitigate this gap, we propose to construct Chain-of-Task to combine the three sub-tasks to improve the LLMs' humor-sarcasm comprehension ability. For a particular sub-task, Chain-of-Task injects information



(a) A sample from the cartoon scene. It consists of an image and the corresponding caption.



(b) A sample from the post scene. It contains an image and the corresponding caption.

Figure 1: Two samples from the scene of cartoon and post, respectively. In each sample, we challenge models with three progressive sub-tasks: detecting, matching and explanation. We also show the detail of the three sub-tasks in the two samples. Notably, the caption of cartoon and post is removed in matching task as it serves the correct option.

of the other two sub-tasks into the prompt as intermedia reasoning process.

Beyond humor-sarcasm comprehension, how the ability of understanding humor and sarcasm influences the humor-sarcasm generation capability remains underexplored. Therefore, we propose a novel yet challenging humor-sarcasm generation task. Given the image from cartoon or post, we aim to generate a short humorous or sarcastic story. Furthermore, some open-source multimodal LLMs are selected to be finetuned on MOCK, and conduct humor-sarcasm generation task, which helps how comprehension influences generation in humor-sarcasm. Surprisingly, LLMs show a much better capability of generating humor-sarcasm, with a better comprehension of humor-sarcasm.

In a nutshell, our contributions are fourfold.

- We devise three progressive sub-tasks: detecting, matching and explaining, which aim at comprehending humor-sarcasm from multiple perspectives.
- We extend some existing humor-sarcasm related datasets, to curate MOCK, a comprehensive benchmark containing rich sources for humor-sarcasm comprehension evaluation.
- We conduct extensive evaluation on the devised sub-tasks for both LLMs and humans, which reveals a gap between human and LLMs on understanding humor-sarcasm. As a byproduct, we release our MOCK and code[1] to facilitate the research community.
- We propose a novel humor-sarcasm generation task, and prove that humor-sarcasm comprehension can enhance humor-sarcasm generation, utilizing MOCK.

[1] https://xhhxn9zbg2.wixsite.com/edge-1.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) have exhibited great potential in various natural language comprehension and generation tasks. For example, Jentzsch and Kersting explored joke generation with ChatGPT (OpenAI, 2022). Some researchers (Hessel et al., 2023) made efforts to comprehend humorous cartoons with LLMs. Notably, the latest closed-source LLMs, such as GPT-4V and GPT-4o (OpenAI et al., 2024) showed great abilities in comprehension and generation tasks, But recent works (Yang et al., 2024) showed a gap between LLMs and humans in understanding multimodal content like image-caption pairs.

2.2 Humor-sarcasm Comprehension

Early studies (Cai et al., 2019; Castro et al., 2019) focus on classification for humor and sarcasm, which contributes to identifying humor and sarcasm. However, comprehension is more than identification. Therefore, many work makes efforts to explain humor and sarcasm. For example, Desai et al. proposed to explain ironic semantics for the sarcastic posts. And some work utilize advanced

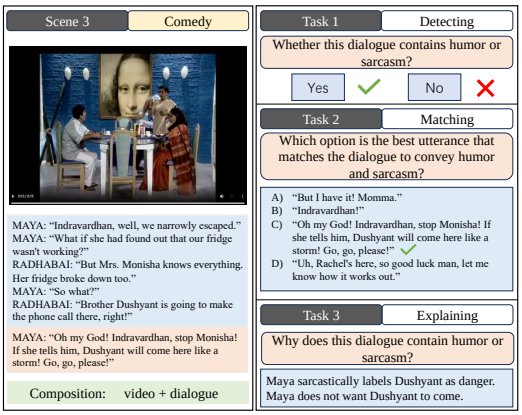


Figure 2: A sample from the comedy scene. It consists of a video clip and the corresponding dialogue. In addition, we present the three sub-tasks in the right part. Notably, the last utterance is removed in the matching task, and serves as the correct option.

backbone (Kumar et al., 2022; Jing et al., 2023; Ouyang et al., 2024), such as BART, to conduct explanation generation. In addition, Hessel et al. curated a cartoon dataset to evaluate wherther LLMs can understand humorous cartoons. Although the above studies have well explored classifying and explaining humor and sarcasm, they are limited to a specific scene and task. There still lack a systematic evaluation benchmark to assess the humor-sarcasm comprehension capability of LLMs.

3 MOCK

This section provides a comprehensive overview of the dataset and the tasks involved in our study.

We introduce the Humor-sarcasM cOmprehension benChmarK, named MOCK, which encompasses three distinct yet interrelated scenes, tailored for three progressive tasks. The dataset comprises a total of $73,872$ samples, including three tasks in three various scenes, with a detailed breakdown provided in Table 1.

To assess the capabilities of LLMs in discerning humor and sarcasm across these varied scenes (*i.e.,* cartoon, post and comedy), we designed an extensive evaluation framework that encompasses three primary subtasks: detecting, matching and explaining

Together, these subtasks are strategically designed to provide a thorough and multidimensional assessment of LLMs, shedding light on their proficiency and potential shortcomings in comprehending humor and sarcasm in the three scenes. Furthermore, each sample in the cartoon, post, and comedy categories is enriched with annotations that cater to the specific demands of our tasks. Specifically, for the Detection task, a binary label is assigned to indicate the presence of humor or sarcasm. Additionally, an explanatory note is provided to elucidate the humorous or sarcastic nuances of the sample, thereby facilitating a deeper comprehension of the underlying humorous or sarcastic semantics.

3.1 Scene and Sub-task Overview

Our MOCK encompasses three distinct yet interrelated scenes (*i.e.,* cartoon, post and comedy), tailored for three progressive sub-tasks (*i.e.,* detecting, matching and explaining).

3.1.1 Scene Category

We illustrate three samples from cartoon, post and comedy in Figure 1 and Figure 2.

- **Cartoon.** The samples from cartoon consist of a pair of image and caption.
- **Post.** In the scene of post, we also provide a image-caption pair each sample.
- **Comedy.** Different from cartoon and post, we supply a video clip and the corresponding dialogue context in comedy scene.

3.1.2 Sub-task Composition

We present three samples from cartoon, post and comedy, respectively, to introduce the three progressive tasks in Figure 1 and Figure 2.

- **Detecting.** *Can LLMs identify humor-sarcasm?* We aim to identify whether the given cartoon, post or comedy contains humor-sarcasm.
- **Matching.** *Can LLMs choose the best-matched option containing humor-sarcasm?* Given four possible options, we target at selecting the option that match the cartoon, post or comedy best to make humor-sarcasm.
- **Explaining.** *Can LLMs explain humor-sarcasm like humans?* In this task, we make efforts to generate a proper explanation to interpret the humor-sarcasm semantics reside in the sample from cartoon, post or comedy.

3.2 Data Collection

We collect three types of data (*i.e.,* cartoon, post and comedy) to support humor and sarcasm comprehension.

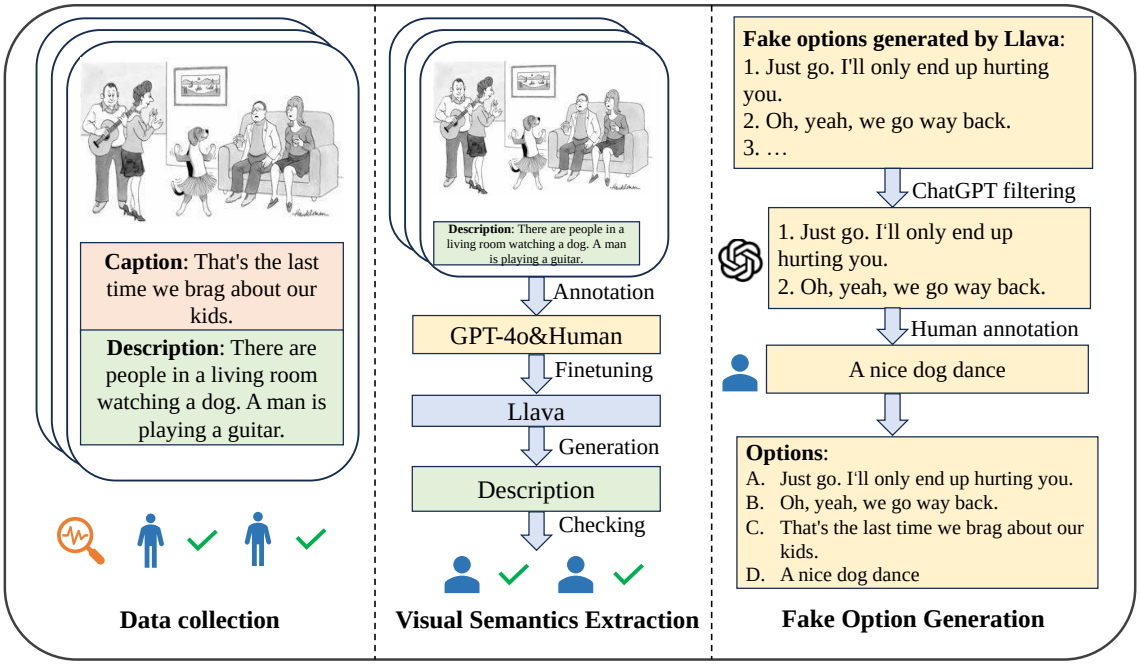


Figure 3: Schematic diagram of MOCK benchmark construction process including three stages: Data Collection, Visual Semantics Extraction and Fake Option Generation.

Cartoon. Our collection encompasses 11,303 cartoons sourced from the weekly New Yorker caption contest[2], Cartoon Movement[3], and CartoonStock[4] through web scraping. Each cartoon is paired with a corresponding caption, offering a rich dataset for humor and sarcasm comprehension.

Post. We have curated a dataset of $28,145$ posts from the MORE sarcasm explanation dataset (Desai et al., 2022) and a sarcasm detection dataset (Cai et al., 2019), originally collected from Twitter, Instagram, and Tumblr. Posts were filtered for quality and relevance, with additional $3,000$ posts collected from Twitter to maintain dataset scale. Each post consists of an image and its accompanying caption.

Comedy. Dialogues from the MUStARD (Kumar et al., 2022) and WITS (Kumar et al., 2022) datasets, drawn from comedies such as *Friends*, *The Golden Girls*, *Sarcasmaholics Anonymous* and *Sarabhai v/s Sarabhai*, were selected. We filtered dialogues to a minimum of three utterances and augmented the dataset with neutral dialogues. The final dataset comprises $7,757$ dialogues, each associated with a video clip and the respective utterances.

[2] https://www.newyorker.com/cartoons/contest.
[3] https://www.cartoonmovement.com/.
[4] https://www.cartoonstock.com/.

3.3 Visual Semantics Extraction

Considering the text-only LLM cannot receive the visual information of the cartoon and post

To bridge the gap for text-only LLMs that they cannot understand images directly, we propose to extract visual semantics involved cartoons and posts, represent visual semantics through literal image description. Overall, we first annotate the description for part of the images, and then train a visual-semantics generation model with the annotated data based on the advanced Multimodal LLM (MLLM) named LLaVA.

Annotation. 1) Cartoon. Manual annotation was conducted for 651 images in previous work (Hessel et al., 2023), focusing on complex compositions, we add additional 800 image annotations to expand the dataset. 2) Post. We utilize the advanced vision comprehension capability of GPT-4o[5] to generate high-quality description for $1,410$ images.

Finetuning. The annotated descriptions for cartoons and post were utilized to finetune LLaVA, resulting in two specialized models for cartoons and posts.

Generation and Checking. The finetuned models generated descriptions for the respective images, which were then quality-checked by one human against the corresponding images. Discrepancies

[5] https://chatgpt.com/?oai-dm=1.

Scene	Detection			Matching			Explanation		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
Cartoon	8,521	1,066	1,066	8,681	1,085	1,085	391	130	130
Post	19,815	2,410	2,410	8,517	1,063	1,063	2,808	351	351
Comedy	4,413	552	552	4,138	517	517	1,792	224	224

Table 1: Basic size statistics for our three sub-tasks: detecting, matching and explaining, in diverse scenes (*i.e.,* cartoon, post and comedy).

were manually corrected.

3.4 Fake Option Generation

To support the matching task, We propose to generate three fake options, with the primary caption as the correct option. In the pursuit of enhancing the matching task's complexity and authenticity, we devised a strategy to generate one plausible but incorrect caption or utterance, complementing the correct primary option. This approach is a judicious blend of manual annotation and automatic generation, aimed at creating a more challenging and realistic evaluation environment for LLMs.

Automatic Generation Strategy. We harness the capabilities of LLaVA and ChatGPT to automate the generation process. For cartoon and post, LLaVA is prompted to produce two distinct captions per image. Given the potential for thematic similarities across images that could yield semantically redundant captions, we employ ChatGPT to filter out options too closely aligned with the correct option and generate a new one. The third option drawn randomly from unrelated image captions. In comedy scene, we provide ChatGPT with the dialogue context, excluding the final utterance, and instruct it to generate two distinct, non-sarcastic responses. These serve as two of the fake options, while the third is sourced from unrelated dialogue contexts.

Manual Annotation Effort. To augment the difficulty of discerning the authentic option, we introduce a manual annotation phase for one distractor. Specifically, 15% of the cartoons and posts are manually assigned a humorous or sarcastic caption but intended to be less effective than the correct option. This annotated option then replaced one of the automatically generated fake options, ensuring that the task is not merely about identification but also about discerning the most appropriate expression of humor or sarcasm. The same principle applies to 15% of the comedy, where a manually annotated utterance replaces one of the fake options.

By integrating both automatic and manual methods, we ensure a robust set of fake options that challenge LLMs to not only recognize humor and sarcasm but also to evaluate the subtleties of expression within the given contexts.

4 Evaluation and Improvement

In this section, we delve into the evaluation metrics, baselines, and methodologies applied to assess the proficiency of Large Language Models (LLMs) in detecting, matching, and explaining humor and sarcasm. We also introduce the innovative Chain-of-Task approach to enhance performance across these tasks without incurring additional training costs.

4.1 Baselines and Human Performance Estimates

We conduct the three humor-sarcasm comprehension sub-tasks on both LLMs and humans.

4.1.1 Evaluated LLMs

For cartoon and post, we evaluate text-only LLMs (*i.e.,* Llama3 (AI@Meta, 2024), GLM4 (Du et al., 2022), and ChatGPT (OpenAI, 2022)) and multimodal LLMs (*i.e.,* LLaVA (Liu et al., 2024), CogVLM2 (Wang et al., 2023), GPT-4V (OpenAI et al., 2024) and GPT-4o (OpenAI et al., 2024)). For comedy, we assess video LLMs (*i.e.,* Video-ChatGPT (Maaz et al., 2024), Video-LLaVA (Lin et al., 2023), Valley2 (Luo et al., 2023), VideoLLaMA2 (Cheng et al., 2024), and VideoChat2 (Li et al., 2023)). A detailed overview of these models is provided in Appendix B.

4.1.2 Human Performance Estimates

To establish a comprehensive benchmark, we conduct human evaluation alongside LLMs assessments. For detecting and matching, two individuals, one author and another familiar with this work, evaluate test instances across cartoon, post and comedy. In addition, the quality of annotated explanations serves as the human performance baseline for the explaining task.

4.2 Evaluation Metrics

We employed accuracy for detecting and matching tasks, while for explaining, we utilized a combination of automatic metrics BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and human evaluations. To further refine our assessment, we introduced an LLM-based metric, Explanation Accuracy (EXPAcc), leveraging GPT-4o API[6] to choose

[6] https://openai.com/index/gpt-4/.

Model	#Params	Detection		Matching		Explanation							
		Zero-shot	Chain-of-Task	Zero-shot	Chain-of-Task	Zero-shot				Chain-of-Task			
		Accuracy	Accuracy	Accuracy	Accuracy	BLEU-4 (%)	Rouge-L (%)	EXPAcc	HumanAcc	BLEU-4 (%)	Rouge-L (%)	EXPAcc	HumanAcc
Llama3	8B	0.675	0.695	0.427	0.434	3.51	19.61	0.346	0.323	3.62	19.67	0.354	0.362
GLM4	9B	0.718	0.723	0.540	0.547	4.34	20.83	0.389	0.400	4.41	20.92	0.408	0.415
ChatGPT	-	0.572	0.575	0.345	0.351	3.31	19.07	0.315	0.292	3.49	19.24	0.323	0.308
Llava	7B	0.627	0.633	0.364	0.370	4.04	20.15	0.177	0.238	4.09	20.71	0.192	0.261
CogVLM2	19B	0.632	0.651	0.434	0.463	4.51	20.27	0.408	0.415	4.78	20.81	0.423	0.431
GLM-4V	9B	0.614	0.620	0.553	0.560	3.83	19.10	0.381	0.377	4.01	19.73	0.392	0.385
GPT-4V	-	0.750	0.762	0.589	0.615	4.62	20.97	0.431	0.423	4.72	21.31	0.446	0.438
GPT-4o	-	0.775	**0.792**	0.674	**0.687**	**4.91**	**21.01**	**0.454**	**0.462**	**4.98**	**21.87**	**0.469**	**0.485**
Human	-	**0.981**	-	**0.931**	-	-	-	-	-	-	-	-	-
(a) cartoon													
Llama3	8B	0.690	0.693	0.338	0.349	2.13	15.45	0.348	0.356	2.72	15.71	0.359	0.362
GLM4	9B	0.687	0.706	0.354	0.361	2.63	19.77	0.399	0.387	2.92	20.18	0.413	0.404
ChatGPT	-	0.583	0.590	0.308	0.321	2.86	19.86	0.322	0.328	3.02	20.21	0.330	0.328
Llava	7B	0.611	0.623	0.293	0.300	1.91	15.41	0.308	0.299	1.93	15.61	0.316	0.311
CogVLM2	19B	0.693	0.720	0.385	0.413	15.55	24.60	0.450	0.453	16.41	25.17	0.459	0.467
GLM-4V	9B	0.641	0.663	0.310	0.325	2.87	18.59	0.385	0.393	3.11	19.97	0.399	0.405
GPT-4V	-	0.769	0.789	0.498	0.528	16.70	25.19	0.439	0.459	17.24	26.43	0.464	0.467
GPT-4o	-	0.791	**0.812**	0.592	**0.612**	**17.21**	**25.61**	**0.470**	**0.462**	**18.44**	**27.04**	**0.473**	**0.481**
Human	-	**0.993**	-	**0.923**	-	-	-	-	-	-	-	-	-
(b) post													
Video-ChatGPT	7B	0.645	0.652	0.352	0.358	2.78	14.92	0.201	0.179	2.90	14.98	0.210	0.205
Video-LlaVA	7B	0.690	0.696	0.369	0.375	2.67	15.27	0.192	0.205	3.31	16.04	0.210	0.219
Valley2	7B	0.592	0.598	0.308	0.319	1.73	9.01	0.129	0.138	2.21	11.02	0.134	0.161
VideoLlaMA2	7B	0.719	0.725	0.389	**0.425**	3.44	**19.27**	0.246	**0.281**	4.18	19.66	0.263	**0.317**
VideoChat2	7B	0.721	**0.734**	0.406	0.418	**3.61**	19.04	**0.286**	0.272	**4.72**	**20.19**	**0.308**	0.303
Human	-	**0.984**	-	**0.919**	-	-	-	-	-	-	-	-	-
(c) comedy dialogue													

Table 2: Evaluation results for three progressive sub-tasks: detecting, matching and explaining. We conduct extensive evaluation across three scenarios: cartoon, post and comedy. In addition, we make comparisons between zero-shot and Chain-of-Task. The best results are in boldface. The results are the average of five replicates.

the better explanation between generated and annotated explanations. The EXPAcc is determined by the ratio of instances where the generated explanation is deemed better by GPT-4o judges. Similarly, HumanAcc is calculated by the average ratio of generated explanation chosen by two human judges.

4.3 Chain-of-Task

Recognizing the interrelated nature of humor and sarcasm comprehension tasks, we propose a Chain-of-Task framework to improve the humor-sarcasm comprehension capability of LLMs. This approach involves:

Detection-oriented prompts leverage information from matching and explaining tasks to enhance LLMs' detecting capability. Specifically, we first give a sample of matching task, and the corresponding answer, we then present the explanation why the answer is humorous or sarcastic. Ultimately, we pose the question to ascertain whether the sample encompasses humor or sarcasm, thereby facilitating the detecting task. In this way, LLMs can learn some useful information from matching and explaining task and hence help detecting task.

Matching-oriented prompts harness insights from detecting and explaining tasks to refine matching accuracy. Initially, we present a sample containing humor-sarcasm from our detecting task, accompanied by the corresponding explanation in explaining task. Subsequently, we conclude with a challenge that requires matching the provided sample excerpt with the appropriate option, thereby refining the accuracy of the matching task.

Explanation-oriented prompts build upon the guidance of detecting and matching tasks to elevate explanation quality. We strategically sequence the tasks from simpler to more complex, beginning with a sample from the detecting task followed by one from the matching task. This structured approach allows us to then introduce the explaining task. By first mastering the foundational tasks of detecting and matching, which are relatively straightforward, we enhance our capacity to tackle more intricate challenges-explaining. This methodical progression not only deepens understanding but also strengthens the ability to address complex tasks effectively.

Further details on the construction of these prompts are available in the Appendix C.

4.4 Main Results

The results of the three sub-tasks in cartoon, post, comedy scenes are presented in Table 2. The evaluated models can be classified into open-source LLMs (*i.e.,* Llama3, GLM4, LLaVA, CogVLM2, GLM-4V, VideoChaGPT, Video-LlaVA, Valley2, VideoLlaMA2 and VideoChat) and closed-source LLMs (*i.e.,* ChatGPT, GPT-4V and GPT-4o).

Detecting. It can be observed that GLM4 exhibit the highest recognition capability of detecting task in the cartoon scene, while CogVLM2 shows the outstanding performance in the post scene, among the open-source models. Notably, the multimodal LLMs are not always better than text-only LLMs

in the scene of cartoon and post. It implies that the literal description of cartoon and post could be a favourable substitution of the image to assist humor and sarcasm comprehension. In addition, the latest GPT-4o achieves the best performance among all the models, with an impressive accuracy of 0.775 in cartoon and 0.791 in post, respectively. For the comedy scene, VideoChat2 shows the highest accuracy among all the video LLMs. Nevertheless, these models still do not match the capabilities of humans, whose accuracy remains at nearly 1, across the scenes of cartoon, post and comedy.

Matching. Among the open-source models, GLM-4V performs the best in cartoon scene, while CogVLM2 outperforms others in post scene. However, the closed-source model GPT-4o outperforms both GLM-4V and CogVLM2, achieving an accuracy of 0.627 in cartoon and 0.592 in post. For comedy scene, VideoChat2 still achieves the best performance across all the video LLMs. A notable observation across all LLMs is that their performance in this task significantly trails behind their performance in the detecting task. This indicates that matching task is more challenging. Additionally, it is evident that these models substantially fall short of human-level performance, which is marked at an impressive accuracy of 0.981 in cartoon and 0.993 in post.

Explaining. In the scenes of cartoon and post, CogVLM2 showcases the highest performance across all the evaluation metrics among open-source models. Unsurprisingly, GPT-4o achieves the best performance among all the models, verifying its overall superior performance. Meanwhile, VideoChat2 exhibits the best performance in BLEU-4 and EXPAcc, while VideoLlaMA2 exceeds others in Rouge-L and HumanAcc. Additionally, we note that the capabilities of these models are significantly weaker than human performance, since EXPAcc and HumanACC of all the models are less than 0.500, it means the manually annotated explanation win the perference of both LLMs and human judges. It can be observed that the accuracy of all evaluated models in matching task is significantly lower than their performance in detecting task, and all of them achieve lower performance in explaining task compared to the matching task. This underscores that detecting, matching and explaining are increasingly difficult tasks. We illustrate two detailed cases for explaining task in Appendix 7.

Chain-of-Task. For all the evaluated LLMs, we observe that their performance with Chain-of-Task in detecting, matching and explaining consistently exceeds those with zero-shot. It verifies the efficiency of Chain-of-task in enhancing the humor-sarcasm comprehension capability.

Overall, the evaluation underscores the progressive difficulty of detecting, matching, and explaining tasks, with LLMs consistently underperforming human capabilities. Furthermore, our Chain-of-Task approach offers a promising avenue for improving LLMs' humor-sarcasm understanding ability without additional training.

5 Does Humor-sarcasm Comprehension Enhance Humor-sarcasm Generation?

In this section, we propose a novel task, named humor-sarcasm generation, and explore how humor-sarcasm comprehension influences humor-sarcasm generation.

5.1 Humor-sarcasm Generation

Beyond comprehending humor-sarcasm, generating humor-sarcasm is also significant but challenging. Given the images from cartoon or post, we aim to generate a short story, which contains humor-sarcasm. We selected 150 high-quality images from cartoon and post, respectively, to support this generation task. In addition, we introduce both LLMs-based metrics and human evaluation to measure the quality of the generated story.

LLMs-based metrics. a) Semantic Matching Score (SMS). Given an image from cartoon or post and the generated short story base on the image, GPT-4o is supposed to rate it 1-5 according to the semantic similarity between the image and the short story, the larger the better. It is used to measure the semantic matching degree between the generated caption or utterance and the given image or dialogue. b) Humor-Sarcasm Score (HSS). GPT-4o is required to rate the short story in aspect of humor and sarcasm effect combining the image. It is utilized to evaluate the effect of humor and sarcasm.

Human evaluation. Given the prime caption or utterance and the generated caption or utterance, two people are required to choose the best one that conveys humor and sarcasm.

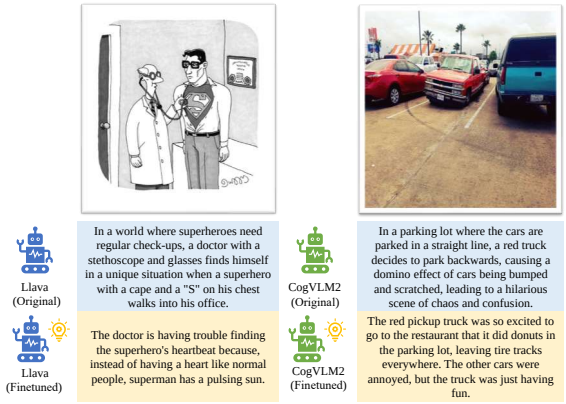


Figure 4: Two cases from humor-sarcasm generation task. Case (a) shows the short stories generated by original and finetuned LLaVA, respectively. Case (b) exhibits the stories of original and finetuned CogVLM2.

Model	Original		Finetuned	
	SMS	HSS	SMS	HSS
Llava	3.87	3.08	3.99	3.34
CogVLM2	3.95	3.31	4.16	3.76
GLM-4V	3.40	3.15	3.51	3.43
GPT-4V	4.14	3.93	-	-
GPT-4o	4.28	4.37	-	-
(a) cartoon				
Llava	3.99	3.32	4.07	3.41
CogVLM2	4.13	3.54	4.28	3.98
GLM-4V	3.77	3.42	3.87	3.64
GPT-4V	4.23	4.21	-	-
GPT-4o	4.41	4.47	-	-
(b) post				

Table 3: The results of humor-sarcasm generation task. Notably, we compare the original LLMs and the LLMs finetuned on MOCK benchmark. In addition, we compare the open-source LLMs with GPT-4V and GPT-4o in generating humor-sarcasm. SMS means semantic matching score, while HSS refers to humor-sarcasm score.

5.2 Transfer from Comprehension to Generation

Intuitively, understanding the innate meaning of humor and sarcasm may help LLMs better generate humor and sarcasm. Therefore, we propose to utilize MOCK dataset to finetune the selected LLMs: LLaVA, GLM-4V and CogVLM2, and hence improve their capability to generate humor-sarcasm. Specifically, we finetune the LLMs on the three humor-sarcasm comprehension sub-tasks: detecting, matching and explaining. Notably, we finetune each LLM separately in cartoon and post scene. Finally, we utilize the LLMs finetuned for cartoon and post to generate short story in the corresponding scene, respectively.

5.3 Results and Analyses

The results of humor-sarcasm generation task are presented in Table 3 As we can see, GPT-4o outperforms all the other models, which shows the advancement of closed-source model. The open-source models' (*i.e.,* LLaVA, CogVLM2 and GLM-4V) capability of generating humor and sarcasm is improved through trained on the humor-sarcasm comprehension benchmark MMHS. It verifies that understanding humor and sarcasm helps enhance the ability to generate humor and sarcasm.

Futhermore, we present two cases in Figure 4. In case (a), the finetuned LLaVA captured the difference between superman and normal human (superman does not have a heart), and generate the humor-sarcasm by teasing the doctor cannot hear Superman's heartbeat. However, the original LLaVA failed to capture the novelty point in the image to generate humor. In case (b), the finetuned CogVLM2 cleverly captured the comic effect of the red truck in the image, and used personification and metaphor to generate humor and sarcasm, while the original CogVLM2 failed.

Overall, both the evaluation results and the two presented cases demonstrate that humor-sarcasm comprehension can enhance humor-sarcasm generation.

6 Conclusion

We propose MOCK, a benchmark for humor-sarcasm comprehension of LLMs. MOCK encompasses rich annotated resources and three progressive sub-tasks: detecting, matching and explaining, tailed for humor-sarcasm comprehension. The evaluation results indicate a notable gap between LLMs and human capabilities. Therefore, we introduced the Chain-of-Task approach to mitigate the gap, and improve LLMs' humor-sarcasm comprehension ability. Additionally, we explored the impact of humor-sarcasm comprehension on generation capabilities, with our proposed novel humor-sarcasm generation task. And we demonstrate that a better understanding of humor-sarcasm can enhance LLM's ability to generate humor-sarcasm.

7 Limitations

The existing form of humor-sarcasm are varied, and due to our limited collection of data, it is not feasible to encompass all potential humor-sarcasm content. In this work, we only exemplified three common scenes: cartoon, post and comedy, while humor-sarcasm types in the real world are more than three. On this note, adding more images and annotations would help improve this issue. In addition, we only evaluate the humor-sarcasm comprehension ability in zero-shot and Chain-of-Task setting, where finetuning results are not yet available. We plan to extend our experiment to provide more insight analyses. As for humor-sarcasm generation task, we only finetune the three latest open-source LLMs on humor-sarcasm comprehension task to improve humor-sarcasm generation, since the advanced closed-source LLMs, GPT-4V and GPT-4o do not open finetuning interface yet.

References

AI@Meta. 2024. Llama 3 model card.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4619–4629. Association for Computational Linguistics.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Poorav Desai, Tanmoy Chakraborty, and Md. Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 10563–10571. AAAI.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.

Sophie F. Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, WASSA@ACL 2023, Toronto, Canada, July 14, 2023*, pages 325–340. Association for Computational Linguistics.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11349–11361. Association for Computational Linguistics.

Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can language models laugh at youtube short-form videos? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2897–2916. Association for Computational Linguistics.

Shivani Kumar, Atharva Kulkarni, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5956–5968. Association for Computational Linguistics.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual

representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

OpenAI. 2022. Introducing chatgpt. *CoRR*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Kun Ouyang, Liqiang Jing, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Sentiment-enhanced graph-based sarcasm explanation in dialogue. *CoRR*, abs/2402.03658.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Noam Shazeer. 2020. GLU variants improve transformer. *CoRR*, abs/2002.05202.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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. 2023. Cogvlm: Visual expert for pretrained language models.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can large multimodal models uncover deep semantics behind images? *CoRR*, abs/2402.11281.

A Explaining

We exhibit two cases for explaining task in Figure 5

B Large Language Models

- **Llama3** (AI@Meta, 2024) A collection of pretrained and instruction tuned generative text models in 8 and 70B sizes. The Llama 3 instruction tuned models are optimized for dialogue use cases and outperform many of the available open source chat models on common industry benchmarks. In this work, we adopt Llama-3-8B.

- **GLM4** (OpenAI et al., 2024) GLM-4-9B is the open-source version of the latest generation of pre-trained models in the GLM-4 series launched by Zhipu AI. In the evaluation of data sets in semantics, mathematics, reasoning, code, and knowledge, We utilize its human preference-aligned version GLM-4-9B-Chat.

- **ChatGPT** (OpenAI, 2022) ChatGPT developed by OpenAI that specializes in natural language processing and generation.

- **LLaVA** (Liu et al., 2024) We adopt LLaVa-NeXT (also called LLaVa-1.6), it improves upon LLaVa-1.5 by increasing the input image resolution and training on an improved visual instruction tuning dataset to improve OCR and common sense reasoning.

- **CogVLM2** It is a stronger version of CogVLM, which is an extension of Vicuna, incorporating ViT (Dosovitskiy et al., 2021) as the vision encoder, a two-layer MLP (Shazeer, 2020) as adapter, and introducing Visual expert module.

- **GPT-4V** (OpenAI et al., 2024) and **GPT-4o** (OpenAI et al., 2024) They are the multimodal LLMs proposed by OpenAI.

- **Video-ChatGPT** (Maaz et al., 2024) ideo-ChatGPT is a video conversation model capable of generating meaningful conversation about videos. It combines the capabilities of LLMs with a pretrained visual encoder adapted for spatiotemporal video representation.

- **Video-LLaVA** (Lin et al., 2023) Video-LLaVA is an open-source multomodal model trained by fine-tuning LLM on multimodal instruction-following data. It is an auto-regressive language model, based on the transformer architecture.

- **Valley2** (Luo et al., 2023) alley2 is a multimodal video assistant that replaces the backbone of the large language model part of valley with the latest llama2 (Touvron et al., 2023) released by meta company.

- **VideoLLaMA2** (Cheng et al., 2024) The structure of VideoLLaMA2 integrates advanced spatial-temporal modeling with robust audio and visual feature extraction, designed to provide a comprehensive understanding of video content through the alignment and fusion of multimodal data.

- **VideoChat2** (Li et al., 2023) VideoChat2 is a video-centric understanding model that combines large-scale vision and language capabilities to analyze and interpret video content,

She got it in the divorce.

driving home . thought the mountains . in the rain. with a vomiting kid . # happy vacation

CogVLM2	**Zero-shot**: The joke is funny because it plays on the unexpected twist that Batman, a superhero, is not rich enough to afford a Batmobile, and instead, he has to rely on public transportation like the subway. **Chain-of-Task**: The joke is funny because it tells of an unexpected twist: superhero Batman doesn't have enough money to afford the Batmobile, and instead has to rely on public transportation like the subway. Maybe it's because he's divorced.	**Zero-shot**: The post is sarcastic because the situation described is not happy at all, it is quite unpleasant. **Chain-of-Task**: This post is sarcastic because the rainy weather and vomiting kid mentioned imply he is not happy at all, very unpleasant..
GPT-4o	**Zero-shot**: The humor stems from the absurdity of the situation. Divorce settlements usually involve the division of assets like property and belongings, but in this case, the caption sarcastically implies that the character got a subway station in the divorce, which is an unexpected and comically impractical "asset" to possess. **Chain-of-Task**: The humor in the cartoon comes from the juxtaposition of everyday life (two people waiting at a subway station) with the absurdity of one of them being dressed as Batman. The caption, "She got it in the divorce," sarcastically suggests that the man is left with a Batman costume as part of a divorce settlement, highlighting the unexpected and comedic nature of the situation.	**Zero-shot**: The post is sarcastic because the caption is an example of sarcasm, as driving home through mountainous terrain in the rain with a sick child is a challenging and uncomfortable situation, which contrasts sharply with the notion of a 'happy vacation' typically associated with enjoyment and relaxation. **Chain-of-Task**: The post is sarcastic because it describes a series of unpleasant events — driving in the rain through mountains with a vomiting child — yet concludes with the hashtag "#happy vacation," which contrasts the negative experience with an ostensibly positive sentiment, highlighting the irony.
Human	Batman would only take the subway if he didn't have access to the batmobile. In this situation, the onlooking man notices that, and batman insecurely retorts that he lost the iconic car in a divorce settlement. It's funny to think that batman has everyday concerns like divorce settlements.	It's inconvenient driving home through the mountains in the rain with a vomiting kid.

Figure 5: Two random samples of explanations generated by CogVLM2, GPT-4o, and human-written references. Notably, we present the generated explanations by CogVLM2 and GPT-4o in both zero-shot and Chain-of-Task setting.

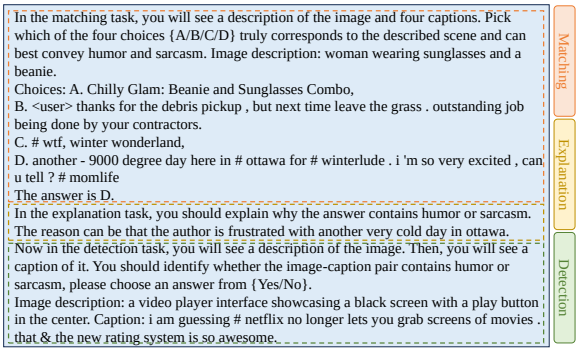


Figure 6: An example of detection-oriented Chain-of-Task prompt in the post scene.

focusing on detailed video comprehension and the contextualization of audio-visual information.

C Constructing Chain-of-Task

We illustrate the detailed prompts for constructing Chain-of-Task in Figure 6.

D Documentation, Licensing, Potential risk and Intended Use of MOCK

MOCK encompasses rich annotated resources and three progressive sub-tasks: detecting, matching and explaining, tailed for humor-sarcasm comprehension. It is extended from some existing humor-sarcasm related dataset. We will release the GPT-generated annotations and our manually written instructions under CC BY-NC 4.0[7]. Notably, there may be some personal information in the images, despite we exclude the offensive information. MOCK should only be used for research purposed only.

[7] https://creativecommons.org/licenses/by-nc/4.0/.