MM-Soc: A Comprehensive Benchmark for Multimodal Large Language Models in Social Media Platforms

Anonymous ACL submission

Abstract

Social media platforms are hubs for multimodal information exchange, encompassing text, images, and videos, making it challenging for machines to comprehend the information or emotions associated with interactions in online spaces. Multimodal Large Language Models (MLLMs) have emerged as a promising solution to address these challenges, yet struggle with accurately interpreting human emotions and complex contents like misinformation. This paper introduces MM-Soc, a comprehensive benchmark designed to evaluate MLLMs' understanding of multimodal social media content. MM-Soc compiles prominent multimodal datasets and incorporates a novel large-scale YouTube tagging dataset, targeting a range of tasks from misinformation detection, hate speech detection, and social context generation. Through our exhaustive evaluation on ten size-variants of four open-source MLLMs, we have identified significant performance disparities, highlighting the need for advancements in models' social understanding capabilities. Our analysis reveals that, in a zero-shot setting, various types of MLLMs generally exhibit difficulties in handling social media tasks. However, MLLMs demonstrate performance improvements post fine-tuning, suggesting potential pathways for improvement. Our code and data are available at Anonymous GitHub[1]

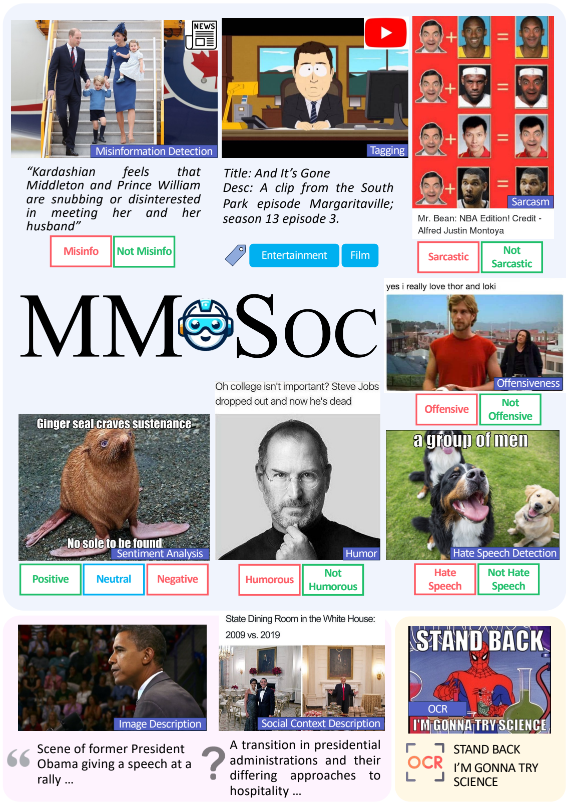

Figure 1: The MM-Soc benchmark includes 10 multimodal tasks, including 7 image-text classification tasks (misinformation detection, tagging, sarcasm, offensiveness, sentiment analysis, hate speech detection, and humor), 2 generative task (image description and social context description) and a text extraction task (OCR).

1 Introduction

Social media platforms have become the epicenter of multimodal information exchange, blending various formats of content such as text, images, and videos. These platforms not only serve as channels for sharing news and personal experiences but also for spreading rumors and shaping public opinions (Ferrara, 2020; Vosoughi et al., 2018). The inherent multimodality of social media content requires users to not only interpret individual modalities such as text or images but also to understand the interplay between them, pushing the boundaries of how machines comprehend human communication in online spaces.

Multimodal Large Language Models (MLLMs) have recently emerged as powerful tools for bridging the understanding of natural language and visual cues, showcasing their potential in a range of tasks ranging from image captioning to complex question answering (Ramos et al., 2023; Liu et al., 2023c,b). Despite these advancements, the complexity of tasks such as understanding hu-

[1] https://anonymous.4open.science/r/MLLMEval-875E

man emotions, memes, and verifying misinformation presents significant evaluation challenges to MLLMs. These tasks require not only combining signals extracted from both textual and visual domains, *but* also considering various social contexts upon making a decision regarding contextual appropriateness or correctness, which often require knowledge of cultural contexts and subjective interpretations (Ruch, 2010; Jacobi, 2014). For instance, the task of explaining visual memes requires not only proficiency in image recognition and language generation, but also capability of understanding the underlying situation of the image on why it should be considered humorous. Given that large language models struggle at solving tasks requiring social knowledge (Choi et al., 2023), we anticipate multimodal social tasks to prove an even harder challenge.

The complexity of multimodal tasks from social media demands a benchmark that can evaluate MLLMs on their understanding of the different data domains as well as the social context. Such a benchmark would not only highlight the current limitations of MLLMs, but also lead to future innovations aimed at bridging the gap between human and machine understanding of multimodal content.

This Work. This paper introduces MM-SOC, a novel multimodal benchmark to rigorously assess the capabilities of MLLMs across diverse tasks typical of social media environments. Along with existing prominent multimodal datasets, we add a large-scale, newly collected YouTube tagging dataset, resulting in ten tasks across five datasets. Our analysis primarily targets open-source MLLMs, recognizing their advantages in terms of rapid deployment, reduced operational costs, and superior capacity for maintaining data integrity compared to centralized proprietary models. Through MM-SOC, we conduct a thorough and systematic examination of MLLMs, exploring and validating new methodologies to augment MLLM efficacy in handling multimodal tasks. Finally, we provide a detailed discussion on the performances, shedding light on the implications of our findings for future MLLM development and deployment.

Contributions. Our contributions are summarized as follows. First, we introduce MM-SOC, a comprehensive benchmark to holistically evaluate MLLMs' capability in tackling multimodal tasks derived from online social networks. Second, we perform a comprehensive evaluation and benchmark 10 representative open-source MLLMs on MM-SOC, comparing their performances with fine-tuned LLM baselines. Third, we conduct two case studies on MM-SOC for testing the effectiveness of two methods: self-improvement and explanation-augmented finetuning. We find that, while zero-shot MLLMs often fall short in achieving comparable performances compared to fine-tuned models, their performances can be improved via specific fine-tuning strategies. We aim to facilitate ongoing research and development in the field by releasing all of our code, data, and tools upon the acceptance of this work.

2 The MM-SOC Benchmark

Overview. The deployment of Multimodal Large Language Models (MLLMs) as general-purpose assistants across social networks marks a significant shift from traditional, specialized models designed for singular tasks. This transition necessitates a comprehensive skill set enabling these models to navigate the multifaceted challenges presented by user-generated content.

Motivated by this, we design MM-SOC, which spans both natural language understanding and generation tasks. These tasks are designed to test the models' abilities to interact with user-generated content encountered online. The selection includes binary classification, multi-class classification, text extraction, and text generation tasks, aiming to cover a wide range of interactions MLLMs might encounter with online content. To ensure a comprehensive evaluation, we employ a variety of 10 tasks that mirror the complexity of real-world scenarios, from understanding online video contents, identifying misinformation to detecting hate speech in memes. The statistics of the dataset are in Table 1.

Tagging. In digital content management, the ability to accurately predict appropriate tags for online content is particularly significant given their diverse and multimodal nature, which includes textual narratives, visual features, and cultural contexts. Effective tagging enhances content discoverability, facilitates content moderation, and significantly improves the user experience.

To this end, we introduce *YouTube2M*, a novel dataset comprising 2 million YouTube videos, specifically curated to assess models' proficiency in predicting tags from a predefined set in Table 7 based on video titles, descriptions, and visual content. We retrieved the URLs of all YouTube videos shared on Reddit over 12 years spanning from 2011

Dataset	Domain	Modality	Size
PolitiFact GossipCop	misinformation	news content, online posts, images, user metadata	485 12,840
Hateful Memes	hate speech, OCR	images, embedded text	12,143
Memotion	sentiment, humor, OCR, offensiveness, sarcasm	images, embedded text	10,000
YouTube	tagging	images, text, channels	1,963,697

Table 1: Statistics of the MM-SOC benchmark.

to 2022. Subsequently, we used YouTube Data API [2] to collect comprehensive metadata of the YouTube videos, including their titles, descriptions, channels, publish timestamps, restrictions, default languages, topic categories, and embeddability status. Additionally, we compiled extensive statistics for each video, covering aspects such as duration, and the number of comments, likes, and views they garnered. To ensure the quality and relevance of the dataset, we filtered the dataset and retained only videos with valid tags and thumbnail images, resulting in a dataset with 1,963,697 videos.

Misinformation Detection. Misinformation detection represents a critical challenge as the proliferation of multimodal misinformation across online platforms can undermine trust in digital ecosystems and lead to real-world harm (Swire-Thompson et al., 2020; Yang et al., 2022; Jin et al., 2022; He et al., 2023). Here, we formulate misinformation detection as a binary classification problem and utilize the PolitiFact and GossipCop datasets (Shu et al., 2020). The task aims at evaluating a model's ability to accurately differentiate between true news and misinformation, leveraging both the textual content and the associated images of news articles.

Hate Speech Detection. The prevalence of hate speech in online platforms has several detrimental effects, both on the individual user-level and on the platform as a whole (Mondal et al., 2017; He et al., 2021). To support research targeted at curbing the spread of harmful content and abusive language, we incorporate the Hateful Memes (Kiela et al., 2020) dataset. This dataset evaluates the ability to recognize messages that attack or demean a group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. Such ability is essential for creating inclusive online environments, protecting users from harm, and complying with legal standards.

Emotion Analysis. The interactions among users in online social media platforms often contain rich and diverse exchanges of emotions. These emotions include not only sentiment but also humor, sarcasm, and offensiveness. Coupled with multimodal means of expressions such as memes, it can be challenging for MLLMs to accurately capture the true emotion conveyed through the message. Therefore, we include the Memotion (Sharma et al., 2020) dataset which focuses on sentiment and emotion analysis within online memes, presenting a multifaceted challenge that spans sentiment analysis and the detection of humor, sarcasm, and offensive contents.

OCR. Optical character recognition (OCR) refers to the task of extracting text within images into machine-encoded text. A model's OCR proficiency is directly related to its ability to access and interpret online information such as infographics, memes, and screenshots of textual conversations, which are prevalent forms of communication and information dissemination online (Zannettou et al., 2018). We use the Hateful Memes and Memotion datasets to evaluate OCR capabilities.

Image & Social Context Description. Image description assesses a model's ability to generate accurate, contextually relevant, and coherent natural language descriptions of images. The capability to accurately describe an image in natural language aids in the understanding of the visual content, which both provides an intermediary step in reasoning about the multimodal inputs and also aids human users in understanding their decisions in an interpretable way. Previous studies have demonstrated that commercial models such as GPT-4/3.5 possess extensive domain knowledge in various fields, including social sciences, and have shown promising results in data annotation, surpassing the performance of human annotators (Savelka et al., 2023; Gilardi et al., 2023; Zhu et al., 2023a). Thus, for each example in the dataset, we employed GPT-4V as a strong teacher to generate descriptions of images and their associated social contexts. For each example within the dataset, we instructed the model to provide a comprehensive description

[2] `https://developers.google.com/youtube/v3`

of the image, encompassing its foreground, background, major subjects, colors, and textures, as well as the social context for each example, such as cultural backgrounds, possible interpretations within various societal groups, and the potential target demographics. These examples served as references for evaluating MLLMs' capabilities to understand both the image contents and social knowledge.

3 Model Selection

We consider 10 prominent open-source models spanning four different distinct architectures: LLaVA-v1.5 (Liu et al., 2023b), BLIP2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), and LLaMA-Adapter-v2 (Zhang et al., 2023b). Details on model parameter volumes are in Table 10. The models are selected to cover diverse model sizes. We apply our prompts (Table 6) to test the performances of MLLMs in a zero-shot setting. For tasks in which ground-truth texts are available as inputs, we compare MLLMs' performances with five unimodal discriminative models in a full fine-tuning setting, including BERT (Kenton and Toutanova, 2019), RoBERTa-Base/Large (Liu et al., 2019), DeBERTa (He et al., 2020), and MiniLM (Wang et al., 2020). These text-only models have shown competitive performances in text classification. Implementation details can be found in Appendix B.2.

4 Benchmark Results

Table 2 shows the overall performances across 10 tasks. Here, we use a unified score for each task to facilitate a high-level performance comparison across diverse tasks. For text classification and extraction tasks, we use the macro-F1 score as the aggregated measure. For text generation tasks including image description (ID) and social context description (SCD), we use ROUGE-L (Lin, 2004). The results for misinformation detection are averaged across PolitiFact and GossipCop, and the results for OCR are averaged across Memotion and Hateful Memes. The complete evaluation results can be found in Appendix B.1.

Zero-shot MLLMs are on par with random guesses. Despite their large model sizes and extensive training corpus, all MLLMs demonstrate underwhelming performances in zero-shot settings, often paralleling and sometimes falling short of the random baseline. This trend is especially evident on the offensiveness detection task, where none of the 10 models surpass the random baseline, with an average macro F1 score of 0.402 compared to the baseline of 0.493. A similar pattern emerges in humor detection, with eight models underperforming the baseline. The tasks in our benchmark which simulate real-life interactions in social media are indeed challenging for most MLLMs.

Zero-shot MLLMs underperform fully fine-tuned models in most settings. We next focus on the misinformation detection task, which takes a binary classification form and can thus be evaluated using encoder-only LLMs such as BERT. Table 5 reveals a consistent underperformance of MLLMs compared to fully fine-tuned LLMs which *only* use textual information. To our surprise, DeBERTa emerges as the top-performing model with only 98 million parameters, whereas zero-shot MLLMs achieve significantly inferior performances.

The low performances of zero-shot MLLMs can be attributed primarily to two reasons: 1) **The divergence in training objectives.** Unlike discriminative models, which are explicitly fine-tuned to predict correct labels, MLLMs are oriented towards maximizing cross-modal alignment and instruction-following abilities. Their training regimes are designed to enhance text generation capabilities based on input images. Such an alignment does not cater to misinformation detection, which demands not only multimodal reasoning but also the ability to evaluate the reliability of sources and incorporate extensive external knowledge. 2) **Disparity in the training corpus content.** MLLMs are predominantly trained for tasks such as object detection, image captioning and visual question answering (VQA) (Dai et al., 2023; Liu et al., 2023c), which rarely encompass tasks in social knowledge reasoning. The lack of tasks requiring subjective reasoning may inherently limit the MLLMs' performance regarding these tasks, and is further supported by the fact that performing task-specific fine-tuning on even much smaller models that use only limited information significantly outperforms MLLMs.

LLaVA achieves highest performance among all MLLMs in most tasks. Among the tested MLLMs, LLaVA-v1.5-13b/7b achieve the best and second best overall performances with average scores of 0.402 / 0.368, a 18.9% / 8.9% improvement over InstructBLIP Vicuna 13B. The performance gap is most significant on the text generation tasks, including ID and SCD as shown in Table 2, where LLaVA-v1.5-13B reaches a performance improvement of 76.9% and 55.7% compared with the other models. This advantage could result from

Model	Misinfo	Hate	Humor	Sarc.	Off.	Sent.	Tag	OCR	ID	SCD	Avg.
llava-v1.5-7b	0.494	0.490	0.450	0.452	**0.484**	0.250	0.068	0.514	**0.260**	**0.218**	0.368
llava-v1.5-13b	**0.642**	0.578	0.534	0.436	0.451	0.291	0.071	0.542	0.259	0.216	**0.402**
instructblip-vicuna-7b	0.311	0.442	0.246	0.481	0.477	0.251	/	0.611	0.048	0.033	0.322
instructblip-vicuna-13b	0.435	0.528	0.435	0.437	0.417	0.262	0.050	0.701	0.097	0.020	0.338
instructblip-flan-t5-xl	0.455	0.470	0.282	0.274	0.464	0.185	0.057	0.652	0.041	0.046	0.293
instructblip-flan-t5-xxl	0.463	0.570	0.406	0.447	0.282	**0.335**	0.128	0.627	0.043	0.023	0.332
blip2-opt-2.7b	0.261	0.369	0.309	0.389	0.411	0.291	0.022	**0.723**	0.141	0.140	0.306
blip2-flan-t5-xl	0.467	0.400	0.183	0.497	0.282	0.245	0.157	0.718	0.147	0.137	0.323
blip2-flan-t5-xxl	0.373	**0.587**	0.200	**0.512**	0.282	0.295	**0.188**	0.676	0.133	0.113	0.336
llama-adapter-v2	0.553	0.524	**0.556**	0.453	0.471	0.268	0.021	0.111	0.098	0.139	0.319
random	0.459	0.500	0.467	0.460	0.493	0.286	/	/	/	/	/

Table 2: Performance comparison across all models on the tasks. Best and 2nd best performances among the MLLMs are highlighted in **bold** and underline, respectively. "ID" and "SCD" stand for the image description task and the social context description task, respectively. Note that instructblip-vicuna-7b fails to generate valid answers on the tagging task. A full comparison of all models on all metrics can be found in Appendix B.1.

Setting	Model	PolitiFact				GossipCop			
		$F1_{macro}$	Acc	AUC	SR%	$F1_{macro}$	Acc	AUC	SR%
zero-shot	llava-v1.5-7b	0.488	0.740	0.534	100.0	0.499	0.812	0.524	100.0
	llava-v1.5-13b	**0.749**	**0.827**	**0.721**	100.0	0.534	0.773	0.535	100.0
	instructblip-vicuna-7b	0.376	0.388	0.511	76.9	0.246	0.251	0.466	70.5
	instructblip-vicuna-13b	0.434	0.485	0.441	94.2	0.435	0.503	0.468	90.0
	instructblip-flan-t5-xl	0.418	0.718	0.500	99.0	0.492	0.811	0.521	98.1
	instructblip-flan-t5-xxl	0.519	0.543	0.537	100.0	0.406	0.429	0.497	100.0
	blip2-opt-2.7b	0.213	0.227	0.429	21.2	0.309	0.309	0.437	11.2
	blip2-flan-t5-xl	0.419	0.721	0.500	100.0	0.514	**0.819**	0.534	100.0
	blip2-flan-t5-xxl	0.545	0.548	0.634	100.0	0.200	0.215	0.481	100.0
	llama-adapter-v2	0.550	0.553	0.613	87.5	**0.556**	0.673	**0.581**	83.6
finetuned	bert-base-uncased	0.850	0.875	0.850	100.0	0.769	0.869	0.797	100.0
	roberta-base	0.894	0.923	0.894	100.0	0.812	0.879	**0.824**	100.0
	roberta-large	0.846	0.885	0.825	100.0	**0.820**	0.858	0.820	100.0
	MiniLM-v2	0.793	0.827	0.806	100.0	0.777	0.858	0.785	100.0
	deberta-v3-large	**0.952**	**0.962**	**0.952**	100.0	0.817	**0.895**	0.792	100.0
random	/	0.471	0.500	0.494	/	0.448	0.500	0.500	/

Table 3: Results of fine-tuning and zero-shot misinformation detection on PolitiFact and GossipCop (Shu et al., 2020). The best and 2nd best performances of each category is highlighted in **bold** and . We report the Macro F1-score (F1), Accuracy (Acc), Area Under the Curve (AUC), and Success Rate (SR). As the number of parameters in the model increases, the model is better at following instructions as seen from their increasing success rate.

both having a wider range of training data and pretraining objectives — multiturn conversation, detailed description, and complex reasoning. For example, the complex reasoning objective typically requires a step-by-step reasoning process by following rigorous logic. Figure 2 shows the performances of the strongest models under each model architecture. The scores are normalized in the 0-1 range. Interestingly, we found that no single model achieves the best performance across all tasks. LLaVA-v1.5-13B performs the best on text generation such as ID or SCD as well as tasks that require social reasoning like misinformation detection, but its ability in tagging is relatively poor. BLIP2 is best on OCR and discriminative tasks like sarcasm and hate speech detection, whereas its generative abilities are relatively poor.

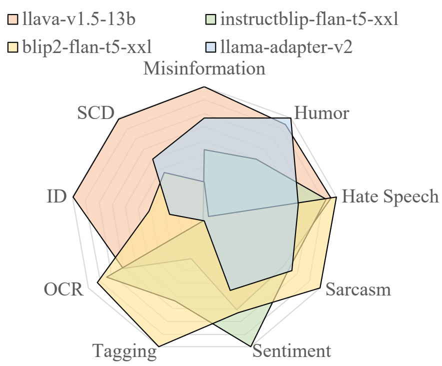

Figure 2: Performances of the 4 representative models on the MM-Soc benchmark.

Larger models exhibit better instruction-following abilities. To quantify an LLM's adherence to predefined content constraints, we leverage a success rate metric, defined as the percentage of responses from a model that aligns with the requested formats. We see a compelling positive

Model	Image Description					Social Context Description				
	M	R-1	R-2	R-L	Len	M	R-1	R-2	R-L	Len
instructblip-vicuna-7b	0.016	0.053	0.008	0.048	3.0	0.014	0.034	0.007	0.033	1.7
instructblip-vicuna-13b	0.040	0.113	0.020	0.097	6.6	0.010	0.021	0.002	0.020	1.9
instructblip-flan-t5-xl	0.014	0.044	0.005	0.041	2.7	0.022	0.050	0.006	0.046	3.0
instructblip-flan-t5-xxl	0.014	0.048	0.005	0.043	2.5	0.009	0.023	0.003	0.023	1.6
blip2-opt-2.7b	0.076	0.158	0.025	0.141	21.2	0.081	0.163	0.021	0.140	16.3
blip2-flan-t5-xl	0.065	0.172	0.026	0.147	9.8	0.069	0.156	0.024	0.137	9.5
blip2-flan-t5-xxl	0.058	0.151	0.025	0.133	9.7	0.066	0.132	0.014	0.113	10.4
llama-adapter-v2	0.041	0.110	0.019	0.098	9.1	0.113	0.152	0.020	0.139	128.5
llava-v1.5-7b	0.223	0.288	0.074	0.260	78.2	0.229	0.247	0.057	0.218	110.1
+ FT	0.217	0.285	0.074	0.253	85.9	0.217	0.249	0.052	0.215	101.1
+ FT w/ explanations	0.240	0.322	0.104	0.289	67.4	0.242	0.280	0.069	0.247	80.9
Improvement	7.7%	12.0%	40.5%	11.0%	-13.8%	5.6%	13.4%	20.9%	13.4%	-26.5%
llava-v1.5-13b	0.223	0.293	0.079	0.259	71.0	0.239	0.247	0.059	0.216	111.5
+ FT	0.207	0.282	0.068	0.252	68.7	0.213	0.246	0.050	0.218	97.3
+ FT w/ explanations	0.248	0.323	0.103	0.294	68.1	0.239	0.278	0.066	0.244	80.8
Improvement	11.0%	10.2%	30.5%	13.5%	-4.1%	0.0%	12.7%	11.6%	13.1%	-27.5%

Table 4: Results on the image description (ID) and social context description (SCD) tasks. We report METEOR (M), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), and the length of responses (Len), calculated as the number of words in the responses. "FT" represents fine-tuning with the ground-truth, and "FT w/ explanations" represents fine-tuning with both the ground-truth and the explanations. The Improvement row indicates performance gain for the FT w/ explanations setting w.r.t. zero-shot baselines. LLaVA-v1.5-7B/13B consistently achieve the best performances among all MLLMs, and exhibit improved performances after fine-tuning on explanations.

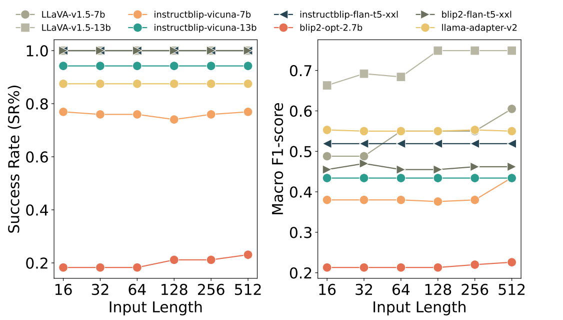

Figure 3: Success Rate (left) and macro-F1 scores (right) of varying input lengths on PolitiFact. The instruction following abilities of MLLMs remains stable across varying input lengths, and exhibit improvements as model size increases.

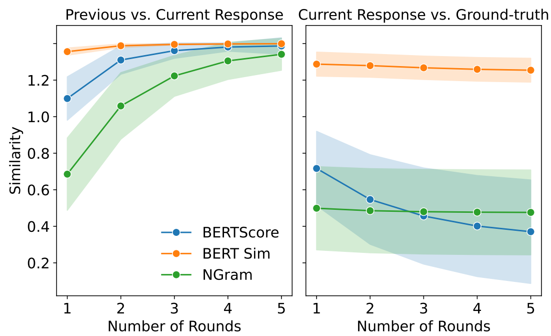

Figure 4: Left: Pairwise similarity between responses at adjacent rounds; right: similarity between response of each round and the ground-truth.

correlation between the parameter size of the text encoder and its ability to follow instructions and precisely classify news content. Table 5 shows that the macro F1-score on PolitiFact for Instruct-BLIP increases from 0.376 to 0.434 when the text encoder changes from Vicuna-7B to Vicuna-13B, and improves from 0.418 to 0.519 when changing from FlanT5-XL to FlanT5-XXL. This correlation indicates that models with larger parameter sizes are equipped with more complex reasoning abilities and a sophisticated understanding of social knowledge, which are essential components for accurately evaluating the veracity of news articles.

Online content ranges from concise and engaging social media posts and microblogs to detailed and extensive narratives found in news articles and in-depth blog posts. This diversity in content length poses a significant challenge for MLLMs, as it requires the models to maintain their generative capabilities over varying context sizes and a wide range of information densities (Peng et al., 2023; Peysakhovich and Lerer, 2023). To address these concerns, we vary the number of tokens used as input to detect misinformation on the PolitiFact dataset from 16 to 512 tokens. The results, as depicted in Figure 3, provide compelling evidence of the MLLMs' stable instruction-following abilities. Notably, we observed an increase in the macro-F1 score as the input length expanded, suggesting that MLLMs are able to leverage evidence from longer contexts for enhanced reasoning and performances.

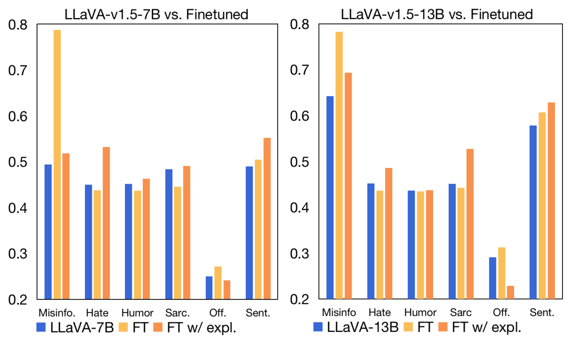

Figure 5: Results of finetuned LLaVA-v1.5-7/13B. Compared to the zero-shot baseline, finetuning with explanations (FT w/ Expl.) and standard finetuning (FT) improves performance across different sets of tasks.

5 Illustrative Uses of MM-Soc

The MM-Soc benchmark can be used to experiment with new methods for enhancing MLLMs in solving multimodal reasoning and generation tasks. We conduct two case studies, proposing new directions for strengthening MLLM capabilities.

5.1 Can MLLMs Self-improve Its Answers?

The ability of MLLMs to self-improve – enhancing their answers iteratively without external supervised signals – can help generate increasingly consistent and robust answers, diminishing the need for human oversight. Using our benchmark, we investigate the self-improvement capabilities of MLLMs. The initial phase involves the model generating an answer for each question. Subsequent iterations, starting from the second round, require the model to produce new answers conditioned on the multimodal inputs and its prior responses. The iterative process is performed for six rounds. To quantitatively assess the evolution of answers across these iterations, we employed three established similarity metrics: BERTScore (Zhang et al., 2019), sentence embeddings similarity (Reimers and Gurevych, 2019), and bigram similarity (Kondrak, 2005). These metrics enabled us to measure the consistency of answers from one round to the next, as well as their fidelity to the ground truth.

Figure 4 displays a notable trend towards convergence in the model's answers with each iteration. For instance, the average BERTScore between answers from consecutive rounds (first to second, and second to third) exhibited a significant increase, from 0.699 to 0.910. Meanwhile, over 55% of all answer pairs between the second and third rounds achieved a sentence embedding similarity score exceeding 0.99. Despite improvements in internal consistency, our analysis revealed a gradual divergence from the ground truth over successive iterations. This was evidenced by a decrease in sentence embedding similarity between MLLM-generated answers and the ground-truth (0.887 $\rightarrow$ 0.854), signaling a potential limitation in the model's ability to maintain factual accuracy in iterative generation.

5.2 Does finetuning MLLMs Improve Overall Performance?

We examine whether MLLMs can improve on MM-Soc via additional fine-tuning steps. Instead of fine-tuning models on separate tasks, we use the data across all different tasks at once for training and examine whether this setting still can contribute towards improvements for each task.

We employed two distinct strategies for fine-tuning. The first approach directly fine-tunes the model using the default input and output data, analogous to a standard fine-tuning setting. In the second approach, we leverage GPT-4(V) as a strong teacher to generate explanations after each ground truth answer for each sample. Along with the original input data, the GPT-generated explanations are augmented as additional training data.

Figure 5 shows the performances of fine-tuned LLaVA-7B and 13B models along with baselines; details can be found in Appendix B.3. With standard fine-tuning, we observe notable gains in detecting misinformation, offensiveness, and sentiment, but also drops in hate, humor, and sarcasm detection. Meanwhile, fine-tuning with explanations improved performance across a broader spectrum of tasks, e.g., increases of 18.2% in hate speech detection and 12.7% in sentiment analysis. Notably, text generation tasks such as image description and social context demonstrated greater gains.

Table 4 further reinforces the positive effects of finetuning with explanations for text generation tasks. Compared to the zero-shot baseline, both the 7B & 13B LLaVA models achieve higher ROUGE-2 scores on image description (40.5% for 7B and 30.5% for 13B). Similarly, for social context description, we observe improvements of 20.9% and 11.6% respectively. These improvements are accompanied by a reduction in response verbosity, highlighting the importance of explanations and rationales for improving multimodal text generation tasks. Interestingly, finetuning without explanations performs *worse* than the baseline, indicating that the standard finetuning approach may not be sufficient to learn the tasks in MM-Soc and signaling the need for refined finetuning strategies.

Setting	Model	PolitiFact $F1_{macro}$	PolitiFact Acc	PolitiFact AUC	PolitiFact SR%	GossipCop $F1_{macro}$	GossipCop Acc	GossipCop AUC	GossipCop SR%
zero-shot	llava-v1.5-7b	0.488	<u>0.740</u>	0.534	100.0	0.499	<u>0.812</u>	0.524	100.0
	llava-v1.5-13b	**0.749**	**0.827**	**0.721**	100.0	<u>0.534</u>	0.773	<u>0.535</u>	100.0
	instructblip-vicuna-7b	0.376	0.388	0.511	76.9	0.246	0.251	0.466	70.5
	instructblip-vicuna-13b	0.434	0.485	0.441	94.2	0.435	0.503	0.468	90.0
	instructblip-flan-t5-xl	0.418	0.718	0.500	99.0	0.492	0.811	0.521	98.1
	instructblip-flan-t5-xxl	0.519	0.543	0.537	100.0	0.406	0.429	0.497	100.0
	blip2-opt-2.7b	0.213	0.227	0.429	21.2	0.309	0.309	0.437	11.2
	blip2-flan-t5-xl	0.419	0.721	0.500	100.0	0.514	**0.819**	0.534	100.0
	blip2-flan-t5-xxl	0.545	0.548	<u>0.634</u>	100.0	0.200	0.215	0.481	100.0
	llama-adapter-v2	<u>0.550</u>	0.553	0.613	87.5	**0.556**	0.673	**0.581**	83.6
finetuned	bert-base-uncased	0.850	0.875	0.850	100.0	0.769	0.869	0.797	100.0
	roberta-base	<u>0.894</u>	<u>0.923</u>	<u>0.894</u>	100.0	0.812	<u>0.879</u>	**0.824**	100.0
	roberta-large	0.846	0.885	0.825	100.0	**0.820**	0.858	<u>0.820</u>	100.0
	MiniLM-v2	0.793	0.827	0.806	100.0	0.777	0.858	0.785	100.0
	deberta-v3-large	**0.952**	**0.962**	**0.952**	100.0	<u>0.817</u>	**0.895**	0.792	100.0
random	/	0.471	0.500	0.494	/	0.448	0.500	0.500	/

Table 5: Results of fine-tuning and zero-shot misinformation detection on PolitiFact and GossipCop (Shu et al., 2020). The best and 2nd best performances of each category is highlighted in **bold** and . We report the Macro F1-score (F1), Accuracy (Acc), Area Under the Curve (AUC), and Success Rate (SR). As the number of parameters in the model increases, the model is better at following instructions as seen from their increasing success rate.

6 Related Works

Multimodal Large Language Models: Multimodal Large Language Models (MLLMs) have demonstrated exceptional natural language understanding and generation abilities by integrating visual information with textual inputs (Awadalla et al., 2023; Yu et al., 2023; Liu et al., 2023a; Verma et al., 2023). Models such as LLaVA (Liu et al., 2023b,c), BLIP2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), and LLaMA-Adapter (Zhang et al., 2023b; Gao et al., 2023) have showcased their superior performance in a range of applications. The success of MLLMs suggests their potential for widespread use in scenarios requiring not only factual analysis and comprehension but also subjective judgment and decision-making based on a nuanced understanding of social contexts and human perceptions. Our study reveals that current MLLMs still fall short in fully grasping and responding to complex social scenarios with the required depth of understanding and sensitivity.

Benchmarking Large Language Models: The evaluation of LLMs is crucial for uncovering their capabilities and identifying potential risks associated with their deployment in sensitive domains (Wang et al., 2024; Liu et al., 2020; Zhang et al., 2023a; Zhao et al., 2023). Benchmarking efforts across various domains such as legal (Deroy et al., 2023), healthcare (Jin et al., 2023), finance (Zhou et al., 2023), psychology (Li et al., 2023a) have provided valuable insights into LLMs such as their reliability (Shu et al., 2023), robustness (Zhu et al., 2023b), and ethical implications (Sun et al., 2023). Despite these efforts, there remains a notable gap in the development of comprehensive multimodal benchmarks for social domains. In this work, we create a holistic multimodal benchmark that captures the broad spectrum of social language and interactions.

7 Conclusion

Our study presents a comprehensive evaluation of 4 leading MLLMs on 10 carefully constructed multimodal social media tasks from diverse domains such as misinformation, hate speech, memes, and a novel YouTube dataset, which comprises our proposed MM-SOC benchmark. Our evaluation of the current capabilities presents the following insights: (i) zero-shot capabilities of certain MLLMs are near-random and underperform drastically in comparison to smaller fully fine-tuned models, (ii) LLaVA-v1.5 is the most competitive open-source MLLM so far, and (iii) instruction following capabilities of MLLMs improve with their size. MM-SOC also enables quantitative case studies, two of which were illustrated in this work and revealed (a) the limitations of MLLMs in self-improving their accuracy and (b) the effectiveness of fine-tuning MLLMs with labeled data. As benchmarks highlight current limitations and guide future research, we intend to expand MM-SOC's coverage to more models and social media tasks to encourage reliable applicability of MLLMs in online spheres.

8 Limitations

We address some limitations of the current study settings, while discussing potential directions for future works.

8.1 Exclusion of Proprietary Models

We excluded models like GPT4V and Gemini from our study for specific reasons. First, this research aims to spotlight the constraints of *open-source MLLMs* in tackling multimodal tasks derived from social media contexts. This emphasis on open-source models is driven by our commitment to enhancing privacy protection. Unlike proprietary models that aggregate data of multiple platforms onto a central server, posing significant privacy risks and operational costs, open-source models are able to process data in a decentralized way (Fan et al., 2023; Zhang et al., 2023c). This distinction not only ensures better privacy safeguards but also resonates with our objective to spotlight and scrutinize the limitations inherent within open-source frameworks when deployed in complex, real-world scenarios like social media. By doing so, we hope that the research community can dedicate resources towards the development of more sophisticated open-source models that address these gaps, promoting the ethos of open science. Second, proprietary models like Gemini reject images containing people and prompts associated with misinformation and hate speech. These restrictions present significant barriers to a comprehensive analysis of MLLMs' performance in handling the diverse and often complex content found on social media platforms.

8.2 Scope of Datasets Included in Benchmark

Online platforms facilitate several well-being discussions and provide support to potentially vulnerable members of the community (Alghowinem et al., 2016; Sindoni, 2020). While our current datasets consider applications of MLLMs for some safety-critical tasks like misinformation and hate detection, extensions of MM-SOC should include datasets and tasks that cover applications that promote inclusivity and support-offering on online platforms. The current version of the benchmark is not "open-world, universal, and neutral," the likes of which have been contested to exist (Raji et al., 2021), but an evolving-effort to contextualize the progress in MLLMs with respect to widely-used social media tasks.

9 Ethical Considerations & Broader Impacts

MLLMs are recognized for exhibiting decision-making biases, a direct consequence of biases present within their training datasets. These include but are not limited to, biases in core sociodemographic categories such as gender, race, and religion (Janghorbani and De Melo, 2023; Ruggeri and Nozza, 2023). This can cause severe issues during downstream applications of MLLMs, particularly in contexts where decisions can significantly affect individual choices.

A significant portion of the biases in MLLMs may be attributed to the data it is trained on. The annotation of subjective tasks in NLP benchmarks also requires consideration, as highlighted in various studies (Aroyo and Welty, 2015; Waseem, 2016; Al Kuwatly et al., 2020). The interpretation of humor or offensive content can significantly vary across different cultural and societal backgrounds, and thus benchmarks should incorporate a broader spectrum of human viewpoints. This is also applicable to certain tasks within our benchmark, where the labels of our questions are reflective of the viewpoints of a hypothetical "average Twitter user." We recognize the importance of this diversity and inclusivity. Our hope is for subsequent research leveraging our benchmark to hopefully develop and include datasets that are more representative of social diversity and inclusiveness, thereby addressing these disparities.

One consistent theme throughout our empirical investigations is that the current performances of MLLMs in general are suboptimal. Notably, certain zero-shot MLLMs exhibit lower accuracy compared to both LLMs fine-tuned exclusively on textual data and even random scores. This underperformance is likely attributable to the insufficient training of MLLMs on tasks requiring subjective judgment and comprehension of social context. For MLLMs to achieve broader and more reliable applicability, future versions should be trained on more tasks that cover ethical, social, and cultural dimensions, thereby ensuring a more comprehensive understanding and interaction capability in diverse contexts.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 9(4):478–490.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv:2308.01390*.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *EMNLP 2023*.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv:2305.06500*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv:2306.01248*.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. 2023. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv:2310.10049*.

Emilio Ferrara. 2020. Dynamics of Attention and Public Opinion in Social Media. In *The Oxford Handbook of Networked Communication*. Oxford University Press.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv:2303.15056*.

Bing He, Yibo Hu, Yeon-Chang Lee, Soyoung Oh, Gaurav Verma, and Srijan Kumar. 2023. A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators. *arXiv:2310.02095*.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *ASONAM*, pages 90–94.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *ICLR*.

Lora L Jacobi. 2014. Perceptions of profanity: How race, gender, and expletive choice affect perceived offensiveness. *North American Journal of Psychology*, 16(2).

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. *arXiv e-prints*, pages arXiv–2310.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The significance of recall in automatic metrics for mt evaluation. In *AMTA*, pages 134–143. Springer.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. The good, the bad, and why: Unveiling emotions in generative ai. *arXiv:2312.11111*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv:2311.10774*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv:2010.03743*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv:2309.00071*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv:2310.01427*.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Willibald Ruch. 2010. *The sense of humor: Explorations of a personality characteristic*, volume 3. Walter de Gruyter.

Gabriele Ruggeri and Debora Nozza. 2023. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada. Association for Computational Linguistics.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv:2306.13906*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Dallas Card, and David Jurgens. 2023. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv:2311.09718*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Maria Grazia Sindoni. 2020. '# youcantalk': A multimodal discourse analysis of suicide prevention and peer support in the australian beyondblue platform. *Discourse & Communication*, 14(2):202–221.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv:2311.09730*.

Briony Swire-Thompson, David Lazer, et al. 2020. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health*, 41(1):433–451.

Gaurav Verma, Ryan A Rossi, Christopher Tensmeyer, Jiuxiang Gu, and Ani Nenkova. 2023. Learning the visualness of text using large vision-language models. In *EMNLP*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 33:5776–5788.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv:2401.10529*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2253–2262.

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

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA. Association for Computing Machinery.

Peiyan Zhang, Haoyang Liu, Chaozhuo Li, Xing Xie, Sunghun Kim, and Haohan Wang. 2023a. Foundation model-oriented robustness: Robust image model evaluation with pretrained models. *arXiv preprint arXiv:2308.10632*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023c. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9963–9977.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv:2310.17512*.

Peilin Zhou, Meng Cao, You-Liang Huang, Qichen Ye, Peiyan Zhang, Junling Liu, Yueqi Xie, Yining Hua, and Jaeboum Kim. 2023. Exploring recommendation capabilities of gpt-4v (ision): A preliminary case study. *arXiv:2311.04199*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023b. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv:2306.04528*.

A Details on Datasets

A.1 Tagging

The tagging task focuses on predicting appropriate "topic categories" for YouTube videos, chosen from a predefined set as listed in Table 7. These topics make it easier for users to find videos that match their interests but also enhance the overall content management strategy. This dataset exemplifies the necessity of multimodal understanding in categorizing online video content. The dataset is licensed under the Apache 2.0 License[3].

[3]https://opensource.org/license/mit/

Task	Prompt	Tags
Misinformation Detection	Is the following news misinformation? **[NEWS]**	misinformation, not misinformation
Hate Speech Detection	Is the following meme hateful?	hateful, not_hateful
Humor	Is the following meme humorous?	humorous, not_humorous
Sarcasm	Is the following meme sarcastic?	sarcastic, not_sarcastic
Offensiveness	Is the following meme offensive?	offensive, not_offensive
Sentiment Analysis	What is the overall sentiment expressed through this meme?	positive, neutral, negative
OCR	What is the text in the image?	/
Image Description	Describe the scene, such as its major subjects, colors, and texture.	/
Social Context Description	Describe the cultural and social context of the image. What particular groups is the image and text targeting at?	/
Tagging	Predict the tags of the following online video given its title, description, and thumbnail image. Different tags must be separated by commas. Title: **[TITLE]** Description: **[DESCRIPTION]**	(See Table 7 for the list of tags for YouTube videos)

Table 6: Prompts and possible values for each task.

YouTube Tags
action-adventure_game, action_game, american_football, association_football, baseball, basketball, boxing, business, casual_game, christian_music, classical_music, country_music, cricket, electronic_music, entertainment, fashion, film, food, golf, health, hip_hop_music, hobby, humour, ice_hockey, independent_music, jazz, knowledge, lifestyle, military, mixed_martial_arts, motorsport, music, music_of_asia, music_of_latin_america, music_video_game, performing_arts, pet, physical_attractiveness, physical_fitness, politics, pop_music, professional_wrestling, puzzle_video_game, racing_video_game, reggae, religion, rhythm_and_blues, rock_music, role-playing_video_game, simulation_video_game, society, soul_music, sport, sports_game, strategy_video_game, technology, television_program, tennis, tourism, vehicle, video_game_culture, volleyball

Table 7: Set of tags for YouTube videos

A.2 Misinformation datasets

We consider two datasets under the misinformation detection theme: PolitiFact and GossipCop. Both datasets were curated by Shu et al. (2020), distributed under the CC-BY-SA License, and are publicly available for download at https://github.com/KaiDMML/FakeNewsNet/.

A.2.1 PolitiFact

This dataset contains news content from the fact-checking website PolitiFact[4], which focuses on political discourse, and contains the title, body, images, and user metadata from news articles. The dataset contains 485 news articles. Each article is annotated into one of the two categories: ‘fake’ and ‘real.’

[4] https://www.politifact.com/

A.2.2 GossipCop

This dataset contains news content from GossipCop, which targets the realm of entertainment news, and includes the title, body, images, from the news articles. The article contains 12,840 new articles, each of which is categorized into one of the two categories: 'fake' and 'real.'

A.3 Hateful Memes

The Hateful Memes dataset contains 12,840 memes that were designed to include meme-like visuals with text laid over them. Since a unimodal classifier (i.e., text-only or image-only) would struggle to make an inference about the hateful nature of the memes without considering both the modalities, they present a unique opportunity to test the multimodal reasoning capabilities of MLLMs. The designed memes were manually annotated to be in one of the two categories: 'hateful' or 'benign.' The dataset is distributed under the MIT License.

A.4 Memotion

The Memotion dataset comprises 12,143 memes, each meticulously annotated with labels that categorize the memes according to their sentiment (positive, negative, neutral), the type of emotion they convey (sarcastic, funny, offensive, motivational), and the intensity of the expressed emotion. The emotion class and the overall sentiment were manually labeled by Amazon Mechanical Turk (AMT) workers. The dataset is distributed under the Community Free Resource License[5].

B Details on Experiments

B.1 Evaluation Metrics

Classification. For classification tasks, we employ metrics including macro precision, macro recall, macro F1-score, accuracy (Acc), and Area Under the Curve (AUC), reflecting the comprehensive assessment of the models' tagging proficiency.

Tagging. For the tagging task, we additionally leverage Hamming Loss and Jaccard index. Hamming loss ($\mathcal{L}_{\text{Hamming}}$) is used to measure the fraction of labels that are incorrectly predicted:

$$\mathcal{L}_{\text{Hamming}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|L|}\sum_{j=1}^{|L|}\text{XOR}(y_{ij}, \hat{y}_{ij}) \quad (1)$$

where $y_{ij} \in \{0, 1\}$ is a binary variable that indicates whether example i has label j. $\hat{y}_{ij} \in \{0, 1\}$ is the predicted binary variable. N is the number of examples in the dataset, and L is the set of labels.

Jaccard index is defined as the size of the intersection between the predicted labels and the ground-truth divided by the size of their union:

$$\text{Jaccard} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \quad (2)$$

where N is the total number of examples. $\hat{Y}_i$ and Y_i are the set of predicted and ground-truth labels for example i.

OCR. We use word error rate (WER), character error rate (CER), and BLEU scores (Papineni et al., 2002). The word error rate (WER) and character error rate (CER) are derived from the Levenshtein distance (Levenshtein et al., 1966), defined as:

$$\text{WER} = \frac{|W_S| + |W_D| + |W_I|}{|W|} \quad (3)$$

$$\text{CER} = \frac{|C_S| + |C_D| + |C_I|}{|C|} \quad (4)$$

where $|W|$ and $|C|$ are the number of words and characters in the ground-truth. $|W_S|$, $|W_D|$, and $|W_I|$ are the number of substitutions, deletions, and insertions at the word, and $|C_S|$, $|C_D|$, and $|C_I|$ are at the character level.

Text Generation. We use n-gram-based metrics including BLEU (Papineni et al., 2002) ROUGE (Lin, 2004), METEOR (Lavie et al., 2004), and n-gram similarity (Kondrak, 2005). These metrics evaluate the MLLMs by measuring the lexical overlap between the generated text and the reference text. Meanwhile, we use two established similarity metrics based on pretrained language models, including BERTScore (Zhang et al., 2019) and sentence embedding similarity (Reimers and Gurevych, 2019), to measure the high-level semantic overlap between two answers. Specifically, BERTScore leverages contextualized word embeddings to capture a token's usage in a sentence and encode sequence information. Sentence embedding similarity sim_{sent} is defined as the cosine similarity between the sentence embeddings of two answers:

$$\text{sim}_{\text{sent}}(\mathbf{s}_i, \mathbf{s}_j) = \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\|\|\mathbf{s}_j\|}, \quad (5)$$

where $\mathbf{s}_i$ is the embedding of the i-th response. Additionally, we calculate the length of response, defined as the number of words in a model-generate response.

[5] https://www.figma.com/legal/community-free-resource-license/

Memotion	P_{macro}	R_{macro}	$F1_{macro}$	WER	CER	BLEU1	BLEU2	BLEU3	BLEU4
llava-v1.5-7b	0.651	0.455	0.535	46.7	40.8	0.495	0.454	0.410	0.365
llava-v1.5-13b	0.665	0.470	0.551	45.0	39.2	0.521	0.481	0.437	0.396
instructblip-flan-t5-xl	0.850	0.482	0.615	46.3	42.2	0.490	0.449	0.405	0.363
instructblip-flan-t5-xxl	0.808	0.441	0.571	50.0	45.3	0.445	0.406	0.365	0.326
instructblip-vicuna-7b	0.853	0.558	0.675	38.7	35.1	0.569	0.534	0.497	0.459
instructblip-vicuna-13b	0.834	0.451	0.585	48.9	44.9	0.459	0.425	0.387	0.350
blip2-opt-2.7b	0.774	0.562	0.651	40.7	35.1	0.537	0.493	0.451	0.407
blip2-flan-t5-xl	0.825	0.593	0.690	37.8	31.3	0.606	0.546	0.488	0.432
blip2-flan-t5-xxl	0.791	0.623	0.697	36.3	27.8	0.632	0.569	0.507	0.448
llama-adapter-v2	0.183	0.084	0.115	94.5	82.2	0.059	0.036	0.027	0.021
Hateful Memes	P_{macro}	R_{macro}	$F1_{macro}$	WER	CER	BLEU1	BLEU2	BLEU3	BLEU4
LLaVA-v1.5-7b	0.560	0.441	0.493	42.3	34.1	0.535	0.500	0.468	0.412
LLaVA-v1.5-13b	0.619	0.469	0.534	40.3	32.8	0.568	0.536	0.506	0.450
instructblip-flan-t5-xl	0.839	0.584	0.689	34.6	27.3	0.618	0.572	0.524	0.467
instructblip-flan-t5-xxl	0.829	0.536	0.651	39.7	32.7	0.550	0.506	0.465	0.408
instructblip-vicuna-7b	0.835	0.644	0.727	29.7	22.5	0.670	0.629	0.587	0.529
instructblip-vicuna-13b	0.824	0.564	0.670	37.1	30.2	0.592	0.552	0.507	0.451
blip2-opt-2.7b	0.759	0.653	0.702	29.4	21.7	0.646	0.599	0.551	0.494
blip2-flan-t5-xl	0.810	0.690	0.745	26.4	17.0	0.726	0.661	0.596	0.527
blip2-flan-t5-xxl	0.777	0.721	0.748	26.0	14.6	0.734	0.662	0.597	0.521
llama-adapter-v2	0.118	0.099	0.108	94.5	78.5	0.075	0.042	0.031	0.024

Table 8: OCR results on Memotion and Hateful Memes. We report macro precision (P_{macro}), macro recall (R_{macro}), macro F1 ($F1_{macro}$), word error rate (WER), character error rate (CER), and BLEU-1/2/3/4 (Papineni et al., 2002).

Model	Pre	Rec	F1	Jaccard	$\mathcal{L}_{Hamming} \downarrow$
instructblip-flan-t5-xl	0.045	0.326	0.057	0.036	0.500
instructblip-flan-t5-xxl	0.092	**0.376**	0.128	0.078	0.161
instructblip-vicuna-13b	0.044	0.230	0.050	0.032	0.429
blip2-opt-2.7b	0.027	0.037	0.022	0.013	0.223
blip2-flan-t5-xl	**0.196**	0.191	0.157	0.112	0.092
blip2-flan-t5-xxl	0.176	0.350	**0.188**	**0.122**	0.085
llama-adapter-v2	0.028	0.029	0.021	0.012	**0.137**
llava-v1.5-7b	0.048	0.345	0.068	0.041	0.406
+ finetuning on ground-truth	0.162	0.373	0.209	0.148	0.063
+ finetuning on explanations	0.562	0.491	0.494	0.400	0.027
llava-v1.5-13b	0.052	0.361	0.071	0.043	0.342
+ finetuning on ground-truth	0.123	0.441	0.167	0.113	0.104
+ finetuning on explanations	0.533	0.473	0.474	0.387	0.027

Table 9: Results of tagging on the YouTube dataset. "FT-Labels" and "FT-Explanations" represent the models fine-tuned on the ground-truth labels and explanations, respectively. A "↓" next to the metric indicates that lower values represent better performances. instructblip-vicuna-7b fails to produce valid predictions in this context.

B.2 Details on Models

Table 10 contains the names and number of parameters of the language encoder and vision encoder for each of the models used in our study. Table 11 contains the accuracy scores of every classification task in our benchmark, examined across all of the zero-shot MLLMs.

B.3 Implementation Details

Benchmark Evaluation For inference, we use Nucleus Sampling (Holtzman et al., 2019) with a probability threshold of 0.9, a temperature of 1.0, and a maximum output length of 256 tokens. To account for the randomness in the generation process, we run each experiment with 3 random seeds and report the average results. All experiments were conducted on a server with 5 A100 80GB GPUs.

Model	Language Encoder	Vision Encoder
llava-v1.5-7b	LLaMA-2-7B-Chat	CLIP ViT-L/14 (0.43B)
llava-v1.5-13b	LLaMA-2-13B-Chat	CLIP ViT-L/14 (0.43B)
instructblip-vicuna-7b	Vicuna-7B	EVA-ViT-G (1.3B)
instructblip-vicuna-13b	Vicuna-13B	EVA-ViT-G (1.3B)
instructblip-flan-t5-xxl	Flan-T5-XXL (11.3B)	EVA-ViT-G (1.3B)
blip2-opt-2.7b	OPT-2.7B	EVA-ViT-G (1.3B)
blip2-flan-t5-xxl	Flan-T5-XXL (11.3B)	EVA-ViT-G (1.3B)
llama-adapter-v2	LLaMA-7B	CLIP ViT-L/14 (0.43B)

Table 10: Multimodal large language models (MLLMs) we evaluated in the experiment.

Model	Misinfo	Hate	Humor	Sarc.	Off.	Sent.	Avg.
llava-v1.5-7b	0.776	0.526	0.763	0.721	0.492	0.485	0.627
llava-v1.5-13b	**0.800**	0.580	0.767	0.775	0.591	0.327	**0.640**
instructblip-vicuna-7b	0.319	0.534	0.771	0.638	0.481	0.547	0.549
instructblip-vicuna-13b	0.494	0.550	**0.776**	0.775	**0.599**	0.443	0.606
instructblip-flan-t5-xl	0.765	0.508	0.226	0.560	0.393	0.387	0.473
instructblip-flan-t5-xxl	0.486	0.587	0.762	**0.777**	0.393	0.471	0.579
blip2-opt-2.7b	0.268	0.508	0.543	0.393	0.418	**0.637**	0.461
blip2-flan-t5-xl	0.770	0.500	0.224	0.597	0.393	0.373	0.476
blip2-flan-t5-xxl	0.775	**0.600**	0.767	0.674	0.393	0.420	0.605
llama-adapter-v2	0.613	0.548	0.721	0.770	0.473	0.455	0.597
random	0.500	0.500	0.510	0.502	0.499	0.326	0.473

Table 11: Accuracy of all models on the tasks. Best and 2nd best performances among the MLLMs are highlighted in **bold** and underline, respectively. "ID" and "SCD" stand for the image description task and the social context description task, respectively.

The models are implemented using the Transformers library (Wolf et al., 2020). We use the NLTK package (Loper and Bird, 2002) to calculate BLEU scores, the rouge[6] package to calculate ROUGE scores and the sentence-bert[7] package to calculate sentence embedding similarities, respectively.

Model Finetuning. We finetuned the models for 1 epoch using a batch size of 16, a warmup ratio of 0.03, a learning rate of 2e-4 and a cosine annealing learning rate scheduler.

[6]https://github.com/pltrdy/rouge

[7]https://github.com/UKPLab/sentence-transformers

GPT-4(V)

The split-frame image contrasts two significant periods in Elon Musk's life, symbolizing not only personal transformation but also financial success over time.

The left side of the image is set in 1999, with a backdrop of a modest, unadorned office space that reflects the early struggles and humble beginnings. Elon, a then young entrepreneur, was portrayed in a simple attire with thinning hair, embodying the initial phase of his journey under financial constraints. The colors in this part of the image are muted—grays, browns, and dull greens—highlighting the austerity of this period.

On the right side, the scene transitions to the present day, showcasing Elon in a modern, sleek office environment that signifies success, innovation, and achievement. The figure now displays a full head of hair, a testament to the personal and financial growth achieved over the years. This transformation is not just physical but also a symbol of the ability to afford high-end cosmetic procedures like hair transplant surgery, which represents the pinnacle of his financial success. The colors here are vibrant and rich—blues, silvers, and whites.

BLIP2

InstructBLIP FlanT5-XXL

Two men in black suits and hats

InstructBLIP FlanT5-XXL

The image shows a man in black and white suit.

LLaVA v1.5 13B

The scene features two men, one on the left and the other on the right, both smiling and posing for the camera. They are the main subjects of the image. The background consists of a black and yellow color scheme, which adds a contrasting element to the scene. The men are wearing ties, which further emphasizes their formal attire. The overall atmosphere of the image is positive and engaging, with both men appearing to be enjoying the moment.

LLaMA Adapter-v2

The scene features a man with a beard, who is wearing a suit and tie, standing next to a photo of himself from 20 years ago. The older photo is in black and white, while the current photo is in color. The man appears to be looking at the older photo, possibly reflecting on his life journey or comparing the differences between the two images. The overall atmosphere of the scene is contemplative and thought-provoking.

Figure 6: Example generation by GPT-4(V) and the four strongest MLLMs under each model architecture.