CMBAN: Cartoon-Driven Meme Contextual Classification Dataset for Bangla

Newaz Ben Alam[1,*], AKM Moshiur Rahman Mazumder[1,*], Mir Sazzat Hossain[1,*], Mysha Samiha[1], Md Alvi Noor Hossain[1], Ahaj Mahhin Faiak[1], Md Fahim[1,2,†], Amin Ahsan Ali[1], Ashraful Islam[1], M Ashraful Amin[1], AKM Mahbubur Rahman[1]

[1]*Center for Computational & Data Sciences, Independent University, Bangladesh*
[2]*Penta Global Limited*

[*]Equal Contribution [†]Project Lead

Correspondence: fahimcse381@gmail.com

Abstract

Social networks extensively feature memes, particularly cartoon images, as a prevalent form of communication often conveying complex sentiments or harmful content. Detecting such content, particularly when it involves Bengali and English text, remains a multimodal challenge. This paper introduces **CMBAN**, a novel and culturally relevant dataset of 2,641 annotated cartoon memes. It addresses meme classification based on their sentiment across five key categories: Humor, Sarcasm, Offensiveness, Motivational Content, and Overall Sentiment, incorporating both image and text features. Our curated dataset specifically aids in detecting nuanced offensive content and navigating complexities of pure Bengali, English, or code-mixed Bengali-English languages. Through rigorous experimentation involving over 12 multimodal models, including monolingual, multilingual, and proprietary architectures, and utilizing prompting methods like Chain-Of-Thought (CoT), findings suggest this cartoon-based, code-mixed meme content poses substantial understanding challenges. Experimental results demonstrate that closed models excel over open models. While the LoRA fine-tuning strategy equalizes performance across model architectures and improves classification of challenging aspects in multilingual meme contexts, this work advances meme classification by providing effective solution for detecting harmful content in multilingual meme contexts.

1 Introduction

In today's digital landscape, memes, particularly cartoon-based ones, have emerged as a pervasive and influential form of communication, profoundly shaping online culture and serving as potent vehicles for conveying diverse ideas, emotions, and societal commentary. While frequently used for comedic effect or social critique, a growing concern lies in their exploitation for spreading hate speech and offensive content. Memes inherently harbor

Figure 1: A Pair of Images with Same Cartoon Template. Though the template is same but due to different content, the memes are labeled differently.

complexities, subtly intertwining visual and textual elements to mask sentiments ranging from sarcasm to deeply problematic narratives. Analyzing such multimodal content poses a unique interpretative challenge, distinct from traditional text-only posts, as harmful messages are often masked by the visual-textual interplay (Lin et al., 2023; Das and Mukherjee, 2023). Moreover, (Kiela et al., 2020) identify unique moderation challenges posed by memes' multimodal nature, where harmful messaging emerges from the complex interaction between seemingly innocuous text and imagery. Standard text analysis methods frequently prove inadequate, given memes' reliance on visual-textual synergy for nuanced interpretation.

The analytical complexity further escalates with cartoon-based memes, which hold exceptional popularity within Bengali online communities. These memes leverage exaggerated visual elements that can playfully or provocatively distort textual nar-

ratives, making their comprehension highly nuanced and culturally embedded (Das and Mukherjee, 2023). This challenge becomes particularly intricate when linguistic complexities, such as code-mixing in languages like Bangla and English, are involved. The multimodal and code-mixed nature of these memes creates distinctive interpretive obstacles; harmful messages can be intricately masked by the confluence of images and local linguistic patterns. Existing automated systems, typically trained on monolingual English data, often struggle to process this linguistic blend, thereby hindering their classification performance.

Despite increasing scholarly attention to meme analysis, a critical research gap remains concerning the systematic study of cartoon-driven, code-mixed Bengali memes. Previous works, such as the MemoSen dataset (Hossain et al., 2022a) for Bengali sentiment analysis or the BanglaAbuseMeme dataset (Das and Mukherjee, 2023) for abusive content, have not specifically addressed the unique traits and widespread prevalence of cartoon-based memes combined with code-mixing.

To bridge this gap, we introduce CMBAN, a novel dataset of 2,641 cartoon-based Bangla memes collected from various social media platforms. Each meme is annotated across five key attributes: humor, sarcasm, offensiveness, motivation, and overall sentiment, with special attention to code-mixed content (Bengali-English). Designed for multimodal analysis, the dataset enables deeper insights into how visual and textual elements interact in memes. Figure 1 illustrates this complexity, where identical visuals convey different meanings depending on subtle textual cues. Such fine-grained annotations are essential for training robust, context-aware models and improving content moderation systems. We further benchmark 12 vision-language models, including monolingual, multilingual, and proprietary architectures, under different prompting strategies (e.g., VQA-style, Chain-of-Thought) and with LoRA-based fine-tuning. Despite these methods, the models still struggle with nuanced categories like humor and sentiment, highlighting persistent challenges in understanding culturally embedded, code-mixed meme content.
Our contributions are as follows:

- We introduce CMBAN, the first cartoon-focused, code-mixed Bangla meme dataset, annotated for five nuanced categories to support fine-grained multimodal analysis.
- We conduct a comprehensive evaluation of 12 vision-language models using multiple prompting and fine-tuning techniques. Results show that even the best-performing model (GPT-4o-mini) achieves only 55.93% accuracy with CoT prompting, confirming the difficulty of this task.

2 Related Work

Bangla Meme Dataset. The MemoSen dataset (Hossain et al., 2022a) focuses on sentiment analysis of Bengali memes. This multimodal dataset, containing 4,417 memes annotated with positive, negative, and neutral sentiments. The study demonstrates that multimodal approaches outperform unimodal counterparts, emphasizing the value of combining visual and textual features for sentiment analysis. The MUTE dataset (Hossain et al., 2022b), addresses this gap by providing a multimodal hate speech dataset comprising 4,158 memes with Bengali and code-mixed captions. The study highlights the challenges of analyzing memes in low-resource languages. Subsequently, BanglaAbuseMeme dataset (Das and Mukherjee, 2023) addresses the detection of abusive memes in Bengali. This dataset fills a critical gap in low-resource language settings, providing a benchmark for classifying abusive content. The corresponding study underscores the effectiveness of multimodal models over unimodal ones. The authors also highlight the challenges of detecting abusive content in memes due to the interplay of text and imagery.
LLMs Prompting for Meme Identification. The application of VLMs and LLMs for meme content detection has gained significant traction recently. Research in this domain has focused on various aspects. LLMs, for instance, have been utilized to detect hate content and enhance meme explanations (Cao et al., 2024; Lin et al., 2024, 2023; Jha et al., 2024). Specifically, Huang et al. (Huang et al., 2024) employed LLM agents for identifying harmful content in memes. Other investigations, such as those by (Hwang and Shwartz, 2023; Hessel et al., 2022), have explored cartoon content to assess VLMs' capacity for understanding visual metaphors and humor, particularly in English-centric image contexts. Furthermore, various studies (Jha et al., 2024; Cao et al., 2023b; Liu et al., 2024c; Cao et al., 2023a; Agarwal et al., 2024) have concentrated on improving LLM effectiveness for meme identification, often by refining prompting

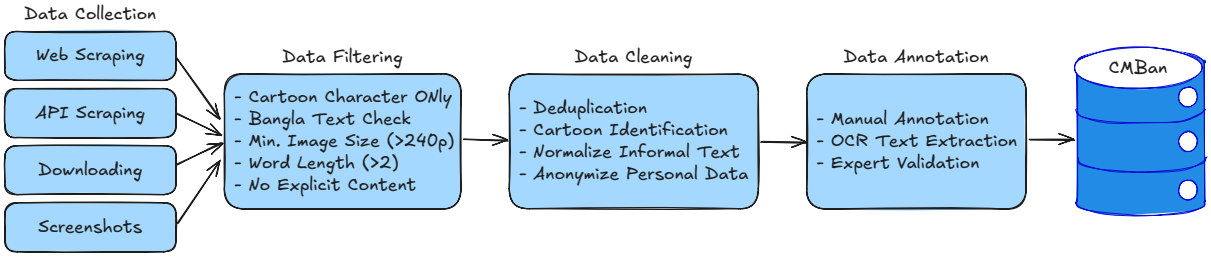


Figure 2: Overview of the CMBAN dataset creation pipeline. This diagram illustrates the sequential stages involved: Data Collection, initial Data Filtering, comprehensive Data Cleaning, and the final Data Annotation process.

techniques or integrating visual questions pertinent to meme interpretation.

While prior studies have significantly advanced Bangla meme analysis, a notable gap persists concerning the popular cartoon-based memes. This work specifically addresses that gap by introducing a novel dataset dedicated to this modality and systematically examining how VLMs interpret and detect such content. We enhance performance by using a base prompt along with visual questions tailored to cartoon-based content.

3 CMBAN: Dataset Creation

We introduce CMBAN, a curated dataset of 2,641 Bangla cartoon-driven memes, designed to support nuanced contextual understanding. Each meme is annotated across five essential dimensions: Humor, Sarcasm, Offensiveness, Motivation, and Overall Sentiment, following the Memotion framework (Sharma et al., 2020; Mishra et al., 2023). The pipeline for CMBAN dataset creation is depicted in Figure 2.

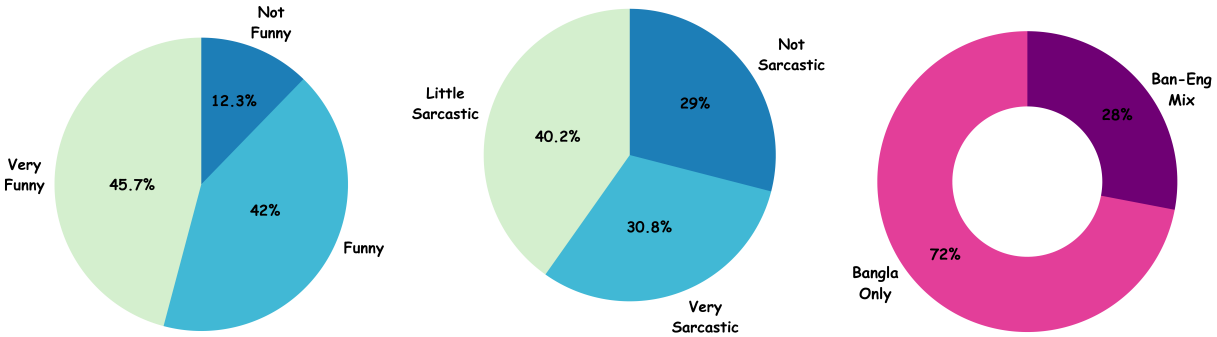
Data Collection. The data collection process aimed to acquire a diverse and culturally representative sample of Bangla-language memes, encompassing both Bangladeshi and Indian Bengali communities. This broad scope captures different regional linguistic nuances and distinct cultural expressions prevalent within these communities. Sources included public domains on popular social media platforms such as Facebook, Instagram, and Pinterest, as well as web repositories like Google Images and specialized meme forums (e.g., Reddit Bangladesh communities). Over 8,000 raw memes were initially collected, capturing temporal trends from August 2024 to February 2025.

Data Filtering. We implemented a systematic filtering process to ensure data quality and relevance. This involved ensuring images prominently featured at least one cartoon or comic character, discarding those without. We also verified the presence of Bangla text, prioritizing memes with over 70% Bangla script. Low-resolution images (below 240p) were removed to ensure visual clarity. Memes with fewer than two words were excluded to maintain meaningful textual context. Finally, any memes containing nudity, graphic violence, or other explicitly inappropriate visual content were removed. This initial filtration step removed approximately 3,854 memes.

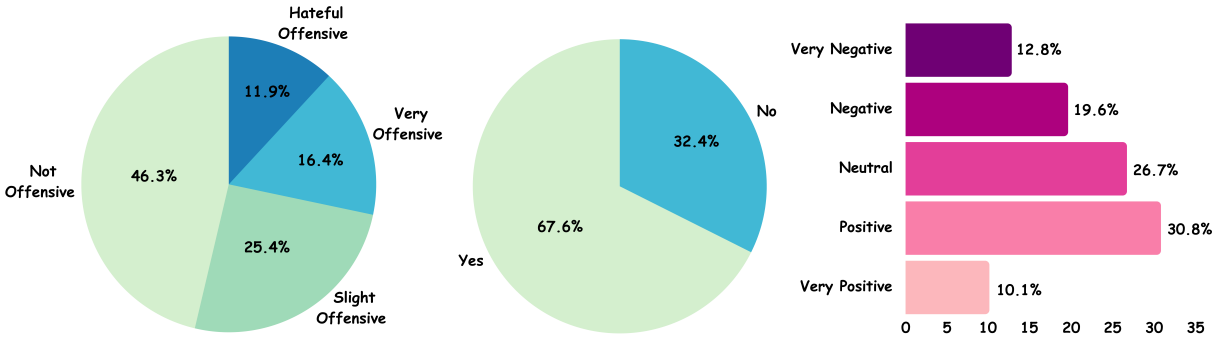
Data Cleaning. Following filtering, a dedicated cleaning phase further refined the dataset. This involved deduplication to remove redundant instances. We then performed precise Cartoon Template Identification, categorizing over 821 unique meme templates in the final corpus. Text was Manually Normalized to standardize informal language, correct misspellings, and regularize transliterated Bangla. Finally, we Anonymized Personal Data to remove any identifiable information. This comprehensive process eliminated approximately 1,500 additional memes, resulting in the final dataset of 2,641 high-quality samples.

Data Annotation. For annotating our dataset, we recruited three annotators with a one-month deadline. All annotators were undergraduates and native Bangla speakers, with extensive experience on social media. They were also active members or administrators of several meme-focused Facebook groups. The annotators were fairly compensated at a rate of BDT 5 per sample. Since the annotators were already familiar with tools like Google Sheets, we chose not to develop a separate annotation platform. All annotators agreed with this approach. We provided interactive guidance and clear annotation guidelines, which are outlined in the Appendix A.

Data Validation. Each data sample was annotated by three annotators to ensure consistency To assess the quality of the annotations, we calculated the inter-annotator agreement score using Fleiss's Kappa score (Fleiss, 1971). The agreement scores



(a) Humor Distribution (b) Sarcasm Distribution (c) Language Distribution



(d) Offensive Content (e) Motivational Content (f) Overall Sentiment

Figure 3: Illustration of the data distributions for (a) Humor, (b) Sarcasm, (d) Offensive Content, (e) Motivational Content, and (f) Overall Sentiment, alongside the Language Distribution (c) within the CMBAN dataset.

for each label in the categories are reported in Appendix A Table 2. From the table, we observed that no agreement score was below than 0.62. According to (Islam et al., 2021; Fleiss, 1971), an inter-annotator agreement score of 0.62 or higher, when considering three annotators, indicates strong agreement across the dataset. An overview of our dataset can be found in Fig 4.

4 Data Analysis.

Statistical Analysis. The CMBAN dataset comprises 2,641 memes, with its detailed statistical distributions presented in Figure 3. Analysis of the humor content reveals a strong prevalence of overtly humorous content, suggesting entertainment as a primary objective of these cartoon memes, as depicted in Figure 3(a). The Sarcasm category also constitutes a significant portion of the memes, implying that humor often intertwines with nuanced, indirect expression in Bangla meme culture, as shown in Figure 3(b). Furthermore, in the Offensiveness category, an overwhelming majority of memes are consistently classified as either non-offensive or slightly offensive, as illustrated in Figure 3(d). This distribution clearly reflects the effectiveness of our rigorous data cleaning processes in mitigating overtly harmful content. Motivational content is present in a notable proportion of memes, as shown in Figure 3(e), suggesting a broader communicative scope beyond purely humorous or satirical expression. For Overall Sentiment, the dataset exhibits a balanced distribution across positive, neutral, and negative categories, with extreme instances being less frequent, indicating that meme expressions often lean towards more moderate emotional tones, as seen in Figure 3(f). Linguistically, the dataset is predominantly composed of pure Bangla memes, with a significant portion also featuring code-mixed Bangla-English. This hybrid language presence reflects the nature of online communication in the region, as presented in Figure 3(c), and underscores the necessity for models capable of handling linguistic diversity.

Cartoon Template Analysis. Our dataset contains 821 unique templates. These templates were modified with minor adjustments to create a variety of memes, which are then used in different contexts. This adaptability allows for a broad range of creative expressions while maintaining a consistent visual foundation. Fig 4 illustrates two pairs of

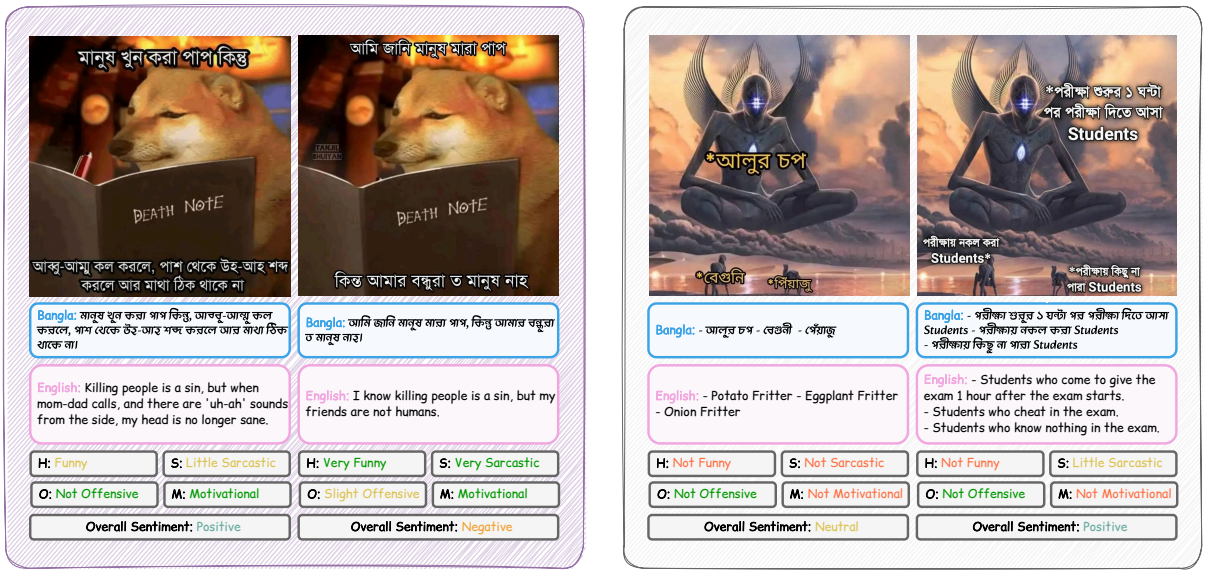


Figure 4: Illustration of cartoon-based meme images sharing similar templates but conveying distinct meanings and contextual interpretations. Two templates are presented, each with two example memes, showing their adaptability.

cartoon-based meme images that share similar templates or backgrounds but differ in their content and contextual meaning.

5 Experiment Setup

In this study, we design three distinct experimental setups: (i) Zero-Shot Prompting, (ii) VQA-Style Prompting, and (iii) Supervised Fine-Tuning. These setups are employed to evaluate the classification accuracy of multiple Vision-Language Models (VLMs) on cartoon memes across the test dataset. To facilitate this, the final dataset was split into training (80%) and test (20%) sets, stratified by the five target categories.

Zero-Shot Prompting. In our zero-shot prompting setup, we present a meme to the models and ask them to classify it across five dimensions: Humor, Sarcasm, Offensiveness, Motivation, and Overall Sentiment. We evaluate a selection of open-source models alongside two proprietary, closed-source vision-language models. The open-source models can be categorized into two main types: (i) Monolingual (English-centric) and (ii) Multilingual models. For the monolingual open-source models, we consider LLaVA-1.5-7B (Liu et al., 2024a), LLaVa-NeXT (Liu et al., 2024b), LLaVa-OneVision–Qwen2-7B (Li et al., 2025), Molmo-7B (Deitke et al., 2024), and SmolVLM-500M (Marafioti et al., 2025). For the multilingual open-source models, we evaluate Pali Gemma-3B (Beyer et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), Phi-3.5-Vision (Abdin et al., 2024), Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024), and InternVL2-8B (Chen et al., 2025). In addition to the open-source models, we also consider the proprietary Gemini-2.0-Flash and GPT-4o-mini models. For the GPT family, we choose GPT-4o-mini due to its cost-effectiveness compared to other versions.

VQA-Style Prompting. Inspired by the work of (Jha et al., 2024; Cao et al., 2023a; Agarwal et al., 2024), which explores Vision Question Answering (VQA) settings to improve the performance of VLMs in meme identification, we also incorporate a set of seven specific questions. These questions are designed to capture various aspects of memes, including the characteristics of different cartoon characters, their tone, the text content, and the visual metaphors present in the images. The full list of questions is provided in the Appendix C Table 4.

This approach allows us to probe the models' ability to understand the nuanced relationships between visual elements and textual content in memes, enhancing their ability to correctly interpret and classify complex meme features. By integrating these targeted questions, we aim to further refine VLM performance, particularly in the context of cartoon-based memes, where the interplay of visual and textual elements plays a vital role in conveying humor and meaning. The detailed prompts are given in Appendix C.

CoT Prompting. Our study also integrates a structured Chain-of-Thought (CoT) prompting approach, inspired by its ability to enhance com-

Models	Label Attributes					
	Humor	**Sarcasm**	**Offen.**	**Motiv.**	**Senti.**	**Avg.**
Base Prompt						
Open VLM Zero-Shot [Monolingual]						
LLaVA-1.5-7B	41.75	31.58	75.44	46.67	23.51	43.79
LLaVA-NeXT	41.75	31.58	75.44	46.67	23.51	43.79
LLaVA-OneVision-Qwen2-7B	32.28	39.30	74.74	52.98	31.19	46.01
Molmo-7B	40.70	38.00	38.60	53.33	28.77	39.88
SmolVLM-500M	41.75	40.00	0.09	46.31	20.70	29.77
Open VLM Zero-Shot [Multilingual]						
Pali Gemma-3B	41.75	39.29	0.02	46.67	18.59	29.26
Qwen2.5-VL-7B	42.11	30.88	75.09	54.04	**40.35**	48.49
Phi-3.5-Vision	34.38	35.43	74.38	54.73	18.24	43.43
Llama-3.2-11B-Vision-Instruct	14.39	31.93	74.74	53.68	24.91	39.93
InternVL2-8B	41.75	35.43	50.87	53.33	22.11	40.70
Close VLM Zero-Shot						
Gemini 2.0 Flash	43.51	40.00	72.29	54.39	35.44	49.17
GPT-4o-mini	**44.50**	**42.78**	**75.50**	**55.32**	37.50	**51.12**
VQA Style Prompt						
Open VLM Zero-Shot [Monolingual]						
LLaVA-1.5-7B	22.46	38.77	73.55	46.38	15.33	39.30
LLaVA-NeXT	42.86	34.07	70.33	53.30	31.32	46.38
LLaVA-OneVIsion-Qwen2-7B	13.96	38.87	74.72	53.96	28.79	42.06
Molmo-7B	38.93	35.00	20.99	54.96	20.61	34.01
Open VLM Zero-Shot [Multilingual]						
Phi-3.5-Vision	36.30	33.45	75.80	53.02	22.42	44.20
Qwen2.5-VL-7B	41.34	40.28	74.91	53.36	32.86	48.55
Llama-3.2-11B-Vision-Instruct	32.98	36.14	75.44	52.63	33.33	46.10
InternVL2-8B	42.77	32.08	74.84	52.20	**43.30**	49.01
Close VLM Zero-Shot						
Gemini 2.0 Flash	45.68	43.4	76.08	56.84	37.98	52.00
GPT-4o-mini	**48.34**	**46.35**	**80.24**	**58.75**	40.00	**54.74**
CoT Prompt						
Open VLM Zero-Shot [Monolingual]						
LLaVA-1.5-7B	48.57	39.29	19.12	50.00	3.57	32.11
LLaVA-OneVIsion-Qwen2-7B	41.75	32.63	75.44	53.33	29.47	46.52
Molmo-7B	38.25	32.28	68.77	49.12	25.61	42.81
Open VLM Zero-Shot [Multilingual]						
Qwen2.5-VL-7B	32.98	20.70	52.63	37.89	7.37	30.31
Phi-3.5-Vision	41.75	31.58	9.12	52.98	8.77	28.84
Llama-3.2-11B-Vision-Instruct	41.75	40.35	75.44	53.33	**43.86**	50.95
InternVL2-8B	42.11	32.98	74.39	53.33	25.96	45.75
Close VLM Zero-Shot						
Gemini 2.0 Flash	46.75	44.20	77.33	57.55	39.22	53.01
GPT-4o-mini	**49.12**	**47.80**	**81.10**	**59.38**	42.18	**55.92**
LoRA Fine Tuning						
Llama-3.2-11B-Vision-Instruct	41.75	**40.35**	75.44	53.33	43.86	**50.95**
LLaVA-NeXT	41.24	39.30	75.00	53.64	44.28	50.69
Qwen2.5-VL-7B	42.11	37.89	75.44	54.04	44.21	50.74
LLaVA-1.5-7B	**42.75**	36.14	**75.68**	**54.22**	**42.34**	50.26

Table 1: The model benchmarking on the test split of the CMBANdataset is reported in terms of accuracy percentage. In the results, "Sarc," "Offen," "Motiv," and "Sent" refer to the categories of Sarcasm, Offensiveness, Motivation, and Sentiment, respectively.

plex problem-solving through intermediate reasoning steps, for meme classification. This method guides VLMs through a step-by-step analysis involving visual understanding (Scene Understanding, Character Identification, Expression/Style Analysis) and textual interpretation (Text Extrac-

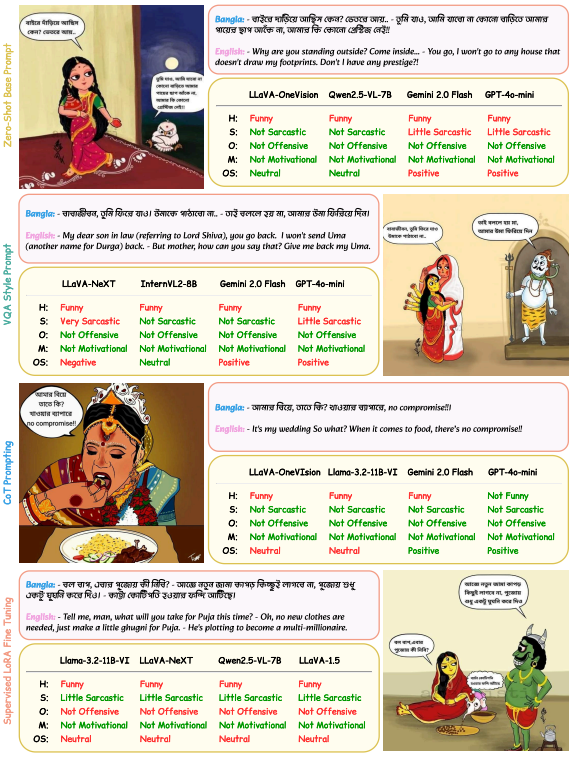


Figure 5: Error Analysis of VLMs performance on Bangla Meme Classification. Red and Green indicate wrongly and correctly classified labels, respectively. Abbreviations: 'H' (Humor), 'S' (Sarcasm), 'M' (Motivational), 'O' (Offensive), 'OS' (Overall Sentiment).

tion/Interpretation, Message/Intent Analysis, Emotional/Social Tone, Visual Technique Check). Following this internal reasoning, models provide a detailed descriptive analysis of the meme, culminating in evaluative labels across the five target categories: Humor, Sarcasm, Offensiveness, Motivation, and Overall Sentiment. The complete CoT prompt structure is detailed in Appendix C.

Supervised Fine-Tuning. To further enhance the model performance, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tuning. Instead of directly modifying all the pre-trained parameters, LoRA utilizes a low-rank matrix. This matrix requires significantly fewer parameters to represent the task-specific adaptations. If W represent the frozen parameters of the pre-trained model. LoRA introduces a low-rank update $\Delta W = A \times B^T$, where A $(d \times r)$ and B $(r \times d)$ are trainable matrices with a much smaller rank r compared to d. These matrices capture task-specific adjustments with fewer parameters. The updated weights W' are the sum of the original weights W and the fine-tuned update ΔW: $W' = W + \Delta W = W + AB^T$. The details about the LoRA configuration is provided in Appendix B.

6 Result Analysis

In Table 1, we present the results of the multimodal models evaluated in all three experimental settings. Based on these results, we derive the following key observations.

Zero-Shot Base-Prompt Setting. Closed models dominate performance, demonstrating better generalization across tasks. GPT-4o-mini achieves the highest overall accuracy at 51.12. Furthermore, monolingual models such as LLaVA-1.5 exhibit higher peak performance in specific attributes. However, they also show greater inconsistency when classifying overall sentiment. This pattern holds true for multilingual models as well, where a decline in performance is observed in the sentiment classification category referring to randomness in prediction. Nevertheless, an exception to this trend is Qwen2.5-VL-7B, a multilingual model that maintains relatively stable performance across attributes where it outperforms the GPT-4o mini in the overall sentiment category in which every other open weight model has struggled.

Effect of VQA-Style Prompting. For monolingual models, VQA-style prompting often yields detrimental effects; for instance, LLaVA-1.5-7B's performance degrades from 43.79% to 39.30%, with similar reductions observed for LLaVA-OneVision and Molmo-7B. This indicates their struggle to leverage additional context in code-mixed data. In contrast, multilingual models show more consistent improvements, with InternVL2-8B improving from 40.70% to 49.01% and Llama-3.2-11B gaining over 6 percentage points in overall accuracy. Closed models exhibit the most robust gains; GPT-4o-mini advances from 51.12% to 54.74% and Gemini 2.0 Flash from 49.17% to 52.00%, likely due to their extensive pretraining on diverse multilingual datasets.

Effect of CoT Prompting. Analysis of the CoT Prompting results reveals that, while some open models show comparable or even lower overall accuracy compared to other prompting methods, certain models demonstrate notable improvements. For instance, Llama-3.2-11B-Vision-Instruct significantly advances its overall accuracy and specifically boosts its Sentiment classification performance with CoT prompting. Similarly, closed models like Gemini 2.0 Flash and GPT-4o-mini also exhibit gains in their average performance.

These findings suggest that CoT prompting offers nuanced benefits, particularly in refining the interpretative capabilities of certain models for challenging attributes like sentiment, even if its general impact on all models is not uniformly positive.

Effect of LoRA Fine-Tuning. For LoRA fine-tuning, we selected both monolingual (LLaVA-1.5-7B, LLaVA-NeXT) and multilingual models (Llama-3.2-11B, Qwen2.5-VL-7B). Our fine-tuning specifically targets zero-shot prompting rather than VQA Style.

From Table 1, a notable observation is that models that initially exhibited weaknesses in specific categories showed noticeable improvements after LoRA fine-tuning. For instance, Llama-3.2-11B-Vision-Instruct initially struggled with humor detection in the zero-shot setting, but its performance improved after fine-tuning catching up to its zero shot counterparts. Another critical perspective emerges in the sentiment analysis task. Across both monolingual and multilingual models, sentiment classification proved to be challenging in the zero-shot setting, with models demonstrating inconsistent performance. However, after LoRA fine-tuning, all models experienced a boost in sentiment classification accuracy, suggesting that fine-tuning helped refine their ability to differentiate emotional tones within the dataset. Given that the dataset contains code-mixed text (Bengali-English), it is likely that LoRA fine-tuning allowed the models to adapt better to mixed-language inputs, making them more robust in capturing sentiment patterns.

Performance Across Label Attributes. Analysis of Table 1 indicates that models generally excel at classifying Offensiveness and Motivational content, with GPT-4o-mini consistently leading. However, Sentiment classification proves the most challenging, particularly for open-source models. While performance on Humor and Sarcasm is varied, LoRA fine-tuning significantly boosts accuracy in the challenging Sentiment category.

Error Analysis. Figure 5 presents a qualitative error analysis of VLM performance on Bangla cartoon memes. Across all prompting setups, Humor was consistently misclassified, especially in Zero-Shot Base and VQA-Style Prompting, where even advanced models like InternVL2-8B struggled. Overall Sentiment was also difficult to predict, while Motivational and Offensive labels were mostly correct, showing some basic understanding of meme context. Fine-tuning with Supervised LoRA led to better results for Sarcasm and Motivational, but misclassifications in Humor and Offensive remained common. Introducing Chain-of-Thought (CoT) Prompting helped improve accuracy for Sarcasm and reinforced already strong performance on Motivational content. However, it did not significantly reduce misclassifications in Humor or Overall Sentiment. These results highlight that even with guided reasoning or fine-tuning, VLMs struggle with nuanced and subjective categories, especially humor, in Bangla meme contexts.

7 Conclusion

Our study introduces CMBAN, a novel dataset of 2,641 cartoon-based Bangla memes, annotated for humor, sarcasm, offensiveness, motivation, and sentiment. Evaluating 12 vision-language models, we find code-mixed meme analysis remains difficult, with GPT-4o-mini achieving the highest accuracy of 54.74% under VQA-style prompting. LoRA fine-tuning improves performance for some models, and CoT prompting provides nuanced gains in complex attributes like sentiment. However, error analysis shows consistent misclassification of Humor and Overall Sentiment across all setups. Motivational content is reliably predicted, while Offensiveness and Sarcasm yield mixed results. These challenges highlight key gaps in multimodal understanding. We believe our dataset and findings will support the development of more capable digital platforms and better meme classification systems.

Limitations

A primary limitation of our study is the relatively constrained dataset size, comprising 2,641 memes. This scale is primarily dictated by the demanding nature of high-quality, culturally informed human annotations, particularly for nuanced tasks like sentiment, sarcasm, and humor categorization within bilingual and bimodal content. Given the inherent complexity of code-mixed Bangla-English memes and the intricate visual-textual reasoning required, the annotation process proved both time-consuming and resource-intensive. While our evaluated models demonstrate promising performance, the current scale of the dataset may inherently limit the generalizability of their findings. Future work should therefore prioritize the exploration of scalable annotation strategies, such as active learning or semi-automated pipelines, to facilitate dataset expansion without compromising the crucial quality of annotations.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. Mememqa: Multimodal question answering for memes via rationale-based inferencing. *arXiv preprint arXiv:2405.11215*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Jingtao Cao, Zheng Zhang, Hongru Wang, Bin Liang, Hao Wang, and Kam-Fai Wong. 2024. Ospc: Detecting harmful memes with large language model as a catalyst. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1892–1895.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A dataset for Bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. *arXiv preprint arXiv:2411.05383*.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Memeguard: An llm and vlm-based framework for advancing content moderation via meme intervention. *arXiv preprint arXiv:2406.05344*.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *TMLR*.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint arXiv:2312.05434*.

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

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

Junxi Liu, Yanyan Feng, Jiehai Chen, Yun Xue, and Fenghuan Li. 2024c. Prompt-enhanced network for hateful meme classification. *arXiv preprint arXiv:2411.07527*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, et al. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv:2303.09892*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd ACL*, Bangkok, Thailand. Association for Computational Linguistics.

A Annotation Guideline

A comprehensive annotation framework was developed to standardize the labeling process, with clear definitions and criteria for each dimension. Nine parameters were annotated: Humour , Sarcasm , Offensiveness , Motivational Quality , Overall Sentiment , Meme Type , Confidence Score , Language Dominance,and Extracted Text.

- Humor was evaluated on a three-point ordinal scale: Not funny (1) indicated no discernible comedic intent, Funny (2) denoted mild amusement, and Very funny (3) reflected strong humor likely to provoke laughter. For example, memes employing exaggerated irony in Bangla text (e.g., "This is our dream" paired with an image of a flooded street) were classified as "very funny."

- Sarcasm was assessed based on ironic or mocking undertones: Not sarcastic (1) implied literal intent, Little sarcastic (2) indicated mild irony requiring cultural context (e.g. ["University exams are easy"] paired with a stressed student meme), and Very sarcastic (3) denoted blatant mockery, often through visual-textual contrast.

- Offensiveness was categorized into four levels: Not offensive (1) for neutral content, Slight offensive (2) for mildly provocative material, Very offensive (3) for derogatory targeting of groups/individuals, and Hateful offensive (4) for content violating hate speech policies (e.g., memes ridiculing religious practices). Cultural sensitivity was prioritized, with annotators flagging Bangla-specific slurs or stereotypes.

- Motivational Quality used a binary classification: Not motivational (1) for neutral or negative content, and Motivational (2) for uplifting messages (e.g., ["Success comes with effort"] paired with a graduation image).

- Overall Sentiment aggregated interpretations into five categories: Very negative (1) , Negative (2) , Neutral (3) , Positive (4) , and Very positive (5) . For example, a

meme criticizing political corruption would score "Very negative," while one celebrating cultural festivals would score "Positive."

Meme Type required annotators to identify the underlying cartoon template (e.g., "Distracted Boyfriend," "Surprised Pikachu") or label it as "Original" if no template was detected. Confidence Score (1–5) allowed annotators to self-report uncertainty, with "1" indicating low confidence and "5" denoting high certainty. Language Dominance recorded the primary language(s) in the meme (e.g., "Bangla-only," "Bangla-English mix"), while Extracted Text preserved OCR-extracted Bangla/English captions for downstream analysis.

Category	Labels	Kappa
Humor	Not Funny	0.70
	Funny	0.65
	Very Funny	0.85
Sarcasm	Not Sarcastic	0.78
	Little Sarcastic	0.72
	Very Sarcastic	0.65
Offensive	Not Offensive	0.75
	Slight Offensive	0.65
	Very Offensive	0.62
	Hateful Offensive	0.64
Motivational	Yes	0.86
	No	0.88
Overall Sentiment	Very Negative	0.63
	Negative	0.68
	Neutral	0.64
	Positive	0.72
	Very Positive	0.75

Table 2: Inter-annotator agreement (Fleiss's Kappa) scores per label for the CMBAN dataset.

B LoRA Finetuning Configuration & Ablation

This table presents the results of a LORA (Low-Rank Adaptation) ablation study conducted on the LLaVA 1.5-7B model, evaluating different configurations based on Loran Rank (64 and 128) and Lora Alpha (64 and 128), over multiple epochs (1 and 3). All of these setting a learning rate of $1e^{-4}$ and a batch size 32 has been used. We train all the models for 1 epoch. The configurations were tested on categories such as Humor, Sarcasm, Offensive, Motivational, and Overall performance. The Loran Rank 64, Lora Alpha 64 (Epoch 1) setting produced the best overall performance, with Humor at 42.75%, Sarcasm at 36.14%, and a strong Offensive score of 75.44%, while the Motivational score was 53.33%, and the Overall score stood at 43.86%. The Loran Rank 128, Lora Alpha 128 (Epoch 1) setup led to a slight drop in Humor (26.67%) and Overall (33.33%), but Offensive remained strong (76.67%) and Motivational was at 50.00%. In the third epoch with Loran Rank 64, Lora Alpha 64, Humor dropped to 26.67%, but there was a notable increase in Sarcasm (66.67%), and Offensive stayed at 76.67%. Motivational remained stable at 50.00%, while Overall performance slightly improved to 36.67%. These results highlight that the Loran Rank 64, Lora Alpha 64 (Epoch 1) configuration generally produces better performance, particularly in the Offensive category, and demonstrates the significant impact of Loran Rank, Lora Alpha, and epoch variations on performance, especially in the Sarcasm and Motivational categories.

The LoRA configuration is set with α and r = 64, a dropout rate of 0.01, a learning rate of $1e^{-4}$ and a batch size ranging from 4 to 32. We train all the models for 1 epoch. Our training data set initially comprises 2251 Question-Answer (QA) pairs. To address the limited availability of multiple choice data, we increase each pair of QA n times, where n corresponds to the number of categories 2 in our dataset. For VLM inference, we utilize vLLM (Kwon et al., 2023), and LLaMA-Factory (Zheng et al., 2024) is used for LoRA finetuning. To ensure reproducibility, greedy decoding, the temperature set as 0 without any sampling mechanism, is used during evaluation.

*ing

Loran Rank	Lora Alpha	Epoch	Humor	Sarcasm	Offensive	Motivational	Overall
64	64	1	**42.75**	36.14	75.44	**53.33**	**43.86**
128	128	1	26.67	36.67	**76.67**	50.00	33.33
64	64	3	26.67	**66.67**	**76.67**	50.00	36.67

Table 3: LORA Ablation Study Results for LLaVA 1.5-7B

C Prompting

C.1 Questions

Focus	Questions
Content	What is shown in the image?
Characters	What cartoon character(s) are depicted in the meme?
Exaggeration	Does the meme utilize exaggerated facial expressions or actions typical of cartoons?
Theme	What is the main theme or message of the meme?
Tone	What is the tone or mood conveyed by the cartoon character's expression?
Text Content	What is the text content in the meme?
Visual Metaphors	Does the meme use visual metaphors or exaggeration (like oversized heads or limbs) typical of cartoons?

Table 4: List of questions for VQA-style Prompt

C.2 Base Prompt

Based Prompt Used for Prediction Part-1

You are an expert multimodal AI assistant. Your task is to answer multiple-choice questions based on an image.

Write given question's answer in JSON FORMAT, PLEASE DO NOT FORGET JSON, ALSO START WITH THE JSON AND NOT ANY THING ELSE, FIRST CHAR IN YOUR RESPONSE IS ITS OPENING BRACE

Response Format: Your response must strictly follow this format:
- Each answer should be on a new line. - Use the format: 'X. Y'
- 'X' is the question number.
- 'Y' is the letter of the correct answer. - Do **not** provide explanations, reasoning, or any additional text.

Example Response: { "Question-1": "B",
"Question-2": "C",
"Question-3": "A",
"Question-4": "B",
"Question-5": "D"
}
Rules: - Do not forget to put the response inside the curly braces else THIS WILL CAUSE FIRE AND MILLIONS WILL BE HARMED**
- **REMEMBER TO CLOSE THE DICTIONARY WITH '}'BRACE, IT GOES AFTER THE END OF DESCRIPTION-YOU ALWAYS FORGET IT, THIS WILL CAUSE A LOT OF ISSUES**
- Do not generate any text outside the expected format.
- Do not rephrase, explain, or add comments.
- Only output the answer choices as specified.
- If uncertain, make your best guess.

Based Prompt Used for Prediction Part-2

Now, answer the following question based on the provided image:

Question-1: How would you rate this image according to Humor? (Humor: the quality of being amusing or comic, especially as expressed in literature or speech)
Options:
A. Not funny
B. Funny
C. Very funny

Question-2: Do you find this image Sarcastic? If then Rate from the following options. (Sarcastic: marked by or given to using irony in order to mock or convey contempt.)
Options:
A. Not sarcastic
B. Little sarcastic
C. Very sarcastic

Question-3: Do you find this image Offensive? If then Rate from the following options. (Offensive: causing someone to feel resentful, upset, or annoyed.)
Options:
A. Not offensive
B. Slight offensive
C. Very offensive
D. Hateful offensive

Question-4: Do you find this image Motivational? Rate from the following options. (Motivational: designed to promote the desire or willingness to do or achieve something.)
Options:
A. Not motivational
B. Motivational

Question-5: Lastly, How would rate the image overall? select from the following options.
A. Very negative
B. Negative
C. Neutral
D. Positive
E. Very positive

C.3 VQA Style Prompt

VQA Style Prompt Used for Prediction Part-1

You are an expert multimodal AI assistant. You are given a cartoon based Meme image.
Now provide detailed answer to Visual Question Answering (VQA) questions:
VQA-1: What is shown in the image?
VQA-2: What cartoon character(s) are depicted in the meme?
VQA-3: Does the meme utilize exaggerated facial expressions or actions typical of cartoons?
VQA-4: What is the main theme or message of the meme?
VQA-5: What is the tone or mood conveyed by the cartoon character's expression?
VQA-6: What is the text content in the meme?
VQA-7: Does the meme use visual metaphors or exaggeration (like oversized heads or limbs) typical of cartoons?

Based on your responses above, select the best answer for the following multiple-choice questions:

MCQ-1: How would you rate this image according to Humor? (Humor: the quality of being amusing or comic, especially as expressed in literature or speech)
Options:
A. Not funny
B. Funny
C. Very funny

MCQ-2: Do you find this image Sarcastic? If then Rate from the following options. (Sarcastic: marked by or given to using irony in order to mock or convey contempt.)
Options:
A. Not sarcastic
B. Little sarcastic
C. Very sarcastic

MCQ-3: Do you find this image Offensive? If then Rate from the following options. (Offensive: causing someone to feel resentful, upset, or annoyed.)
Options:
A. Not offensive
B. Slight offensive
C. Very offensive
D. Hateful offensive

VQA Style Prompt Used for Prediction Part-2

MCQ-4: Do you find this image Motivational? Rate from the following options. (Motivational: designed to promote the desire or willingness to do or achieve something.)
Options:
A. Not motivational
B. Motivational

MCQ-5: Lastly, How would rate the image overall? select from the following options.
A. Very negative
B. Negative
C. Neutral
D. Positive
E. Very positive

VQA Answer Response Format (Strictly Follow This Format):
- Provide detailed answers in complete sentences.

Example Response Format-VQA:

VQA-1-Answer: your_detailed_answer,
VQA-2-Answer: your_detailed_answer,
VQA-3-Answer: your_detailed_answer,
VQA-4-Answer: your_detailed_answer",
VQA-5-Answer: your_detailed_answer,
VQA-6-Answer: your_detailed_answer,
VQA-7-Answer: your_detailed_answer,

MCQ Answers Response Format(Strictly follow given format):
- Answer using only the letter option (A, B, C, etc.) without any explanation.
- Do not add question with the answer-YOU ALWAYS FORGET THAT, THIS WILL CAUSE LOT OF ISSUE

Response Format-MCQ:

MCQ-1-Answer: correct_letter,
MCQ-2-Answer: correct_letter,
MCQ-3-Answer: correct_letter,
MCQ-4-Answer: correct_letter,
MCQ-5-Answer: correct_letter

C.4 CoT Prompt

Chain-of-Thought Prompt Structure

You are an expert multimodal AI assistant. You are shown a cartoon-style meme image. ## Think step by step through the following process before providing your final analysis:

1. Scene Understanding:
- Carefully observe the setting and elements in the image.
- Identify key objects, actions, and the general context.

2. Character Identification:
- Determine if any recognizable cartoon characters are present.
- If unknown, describe the appearance and expression of the characters.

3. Expression and Style Analysis:
- Examine the characters' facial expressions and body language.
- Note any exaggerated or stylized features typical of cartoons (e.g., oversized heads, dramatic poses).

4. Text Extraction and Interpretation:
- Read and understand any text or captions included in the meme.
- Reflect on how the text interacts with the visuals (e.g., sarcasm, irony, humor).

5. Message and Intent Analysis:
- Infer the main theme or message being conveyed.
- Consider whether the meme is humorous, sarcastic, motivational, offensive, or neutral.

6. Emotional and Social Tone:
- Assess the emotional tone (funny, mocking, inspirational, etc.).
- Determine if it carries positive, negative, or mixed social implications.

7. Visual Technique Check:
- Identify use of visual metaphors or exaggeration.
- Comment on how these enhance or distort the message.

After completing these reasoning steps internally, summarize your final analysis in structured form.

Your final response should include:
- A detailed paragraph describing the image and its message based on the above reasoning.
- A set of evaluative labels (choose only one for each):
Humor: [A. Not funny, B. Funny, C. Very funny]
Sarcasm: [A. Not sarcastic, B. Little sarcastic, C. Very sarcastic]
Offensiveness: [A. Not offensive, B. Slight offensive, C. Very offensive, D. Hateful offensive]
Motivation: [A. Not motivational, B. Motivational]
Overall Sentiment: [A. Very negative, B. Negative, C. Neutral, D. Positive, E. Very positive]