MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing

Siddhant Agarwal[1*], **Shivam Sharma**[2,3*], **Preslav Nakov**[4], **Tanmoy Chakraborty**[2]
[1]Indraprastha Institute of Information Technology Delhi, India
[2]Indian Institute of Technology Delhi, India [3]Wipro R&D (Lab45), India
[4]Mohamed bin Zayed University of Artificial Intelligence, UAE
siddhant20247@iiitd.ac.in, {shivam.sharma, tanchak}@ee.iitd.ac.in, preslav.nakov@mbzuai.ac.ae

Abstract

Memes have evolved as a prevalent medium for diverse communication, ranging from humour to propaganda. With the rising popularity of image-focused content, there is a growing need to explore its potential harm from different aspects. Previous studies have analyzed memes in closed settings – detecting harm, applying semantic labels, and offering natural language explanations. To extend this research, we introduce MemeMQA, a multimodal question-answering framework aiming to solicit accurate responses to structured questions while providing coherent explanations. We curate MemeMQACorpus, a new dataset featuring $1,880$ questions related to $1,122$ memes with corresponding answer-explanation pairs. We further propose ARSENAL, a novel two-stage multimodal framework that leverages the reasoning capabilities of LLMs to address MemeMQA. We benchmark MemeMQA using competitive baselines and demonstrate its superiority – $\sim$18% enhanced answer prediction accuracy and distinct text generation lead across various metrics measuring lexical and semantic alignment over the best baseline. We analyze ARSENAL's robustness through diversification of question-set, confounder-based evaluation regarding MemeMQA's generalizability, and modality-specific assessment, enhancing our understanding of meme interpretation in the multimodal communication landscape.[1]

1 Introduction

Memes offer an accessible format for impactful information dissemination for everyone without conventional dependencies of proper formatting or formal language. It provides an easy opportunity for novice content creators and seasoned professionals to propagate information that may sometimes be harmful to the general audience, especially in the age of Internet virality. Previous work has explored aspects such as harmfulness in various forms, such as hate speech (Kiela et al., 2020a), cyber-bullying (Sharma et al., 2022b), and offensive languages (Shang et al., 2021), of memes, typically in a black-box setting.

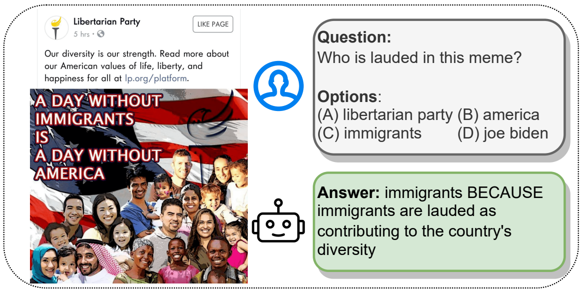


Figure 1: The MemeMQA task: Given an input meme and multiple choices, identify the correct answer and justify.

Memes, with their appealing format and influential nature on social media, necessitate the modeling of complex aspects like harmfulness, targeted social groups, and offensive cues to assess their narrative framing and ensure online content safety. Their growing prevalence as a key medium for information dissemination poses significant societal challenges. A question-answering setup, particularly open-ended or instruction/response formats, offers a user-friendly method for probing models about the potential harmfulness of memes and understanding their responses. This approach enhances model interpretability and serves as an effective tool for content moderation.

In this work, we explore contextualized semantic analysis of memes by introducing a novel multimodal task, MemeMQA (c.f. Fig. 1), which is formulated as follows: Given a meme and a structured question about the semantic role assigned to various entities, (a) deduce the correct answer entity from a set of multiple options, while also, (b) generating succinct explanations towards the answer.

[*] denotes equal contribution
[1]CAUTION: Potentially sensitive content included; viewer discretion is requested.

Findings of the Association for Computational Linguistics: ACL 2024, pages 5042–5078
August 11-16, 2024 ©2024 Association for Computational Linguistics

Building on the work of (Sharma et al., 2022c), we explore the narrative framing of entities like well-known individuals and political figures in online memes. This research is especially important during critical events like elections or pandemics, where the risk of spreading harmful content such as hate speech and misinformation increases, highlighting the need for effective moderation. We adopt terms like 'hero', 'villain', and 'victim' from (Sharma et al., 2022c) to analyze memes' intentions of victimization, glorification, and vilification. Our goal is to deepen the understanding of these memes and contribute to making social media safer. The MemeMQA framework is designed to assist social media users and fact-checkers in evaluating the harmfulness of memes, enabling them to ask questions and receive accurate, informed responses.

Analyzing memes in `MemeMQA` is complex due to their nuanced meanings that demand advanced reasoning, including common sense, and cultural understanding. For instance, the meme in Fig. 1 could simultaneously highlight the role of immigrants in America and promote the Libertarian Party. To correctly answer the question "Who is lauded in this meme?", it's essential to grasp the meme's key themes and the implied message about immigrants enriching diversity, which directly glorifies them. Therefore, "immigrants" is the most suitable answer in this context, rather than "Libertarian Party" or "America", which, despite being referenced positively, would lead to an incorrect conclusion.

In summary, we introduce a new task for answering and explaining multiple-choice questions about political memes, creating a dataset (`MemeMQACorpus`) with $1,880$ questions for $1,122$ memes using `ExHVV` dataset (Sharma et al., 2023). We benchmark `MemeMQACorpus` with various unimodal and multimodal baselines, including recent multimodal LLMs, and propose `ARSENAL`, a novel modular approach that leverages multimodal LLM reasoning capabilities. `ARSENAL` includes rationale, answer prediction, and explanation generation modules. We analyze and compare the performance of `ARSENAL` against these baselines, highlighting its strengths and limitations. Our contributions are summarised as follows[2]:

1. **`MemeMQA`**: A novel task formulation that introduces a multimodal question-answering setup in the context of memes.
2. **`MemeMQACorpus`**: An extension of a previously available dataset to introduce a set of diverse questions and multiple choice settings for `MemeMQA`.
3. **`ARSENAL`**: A multimodal modular framework system architecture that leverages multimodal LLM generated rationales for `MemeMQA`.
4. An exhaustive study in the form of benchmarking, prompt evaluations, detailed analyses of diversified questions, confounding-based cross-examination, implications of multimodality and limitations of the proposed solution.

[2]Supplementary accompanies the source codes and sample dataset.

2 Related Work

This section provides a concise coverage of prominent studies on meme analysis, while also reviewing contemporary works within the domain of Visual Question Answering. Finally we consolidate our assessment of the current state-of-the-art in Multimodal LLMs.

Studies on Memes. Recent collaborative efforts encompass diverse meme analysis aspects, including entity identification (Sharma et al., 2022c; Prakash et al., 2023), emotion prediction (Sharma et al., 2020) and notably, hateful meme detection (Kiela et al., 2020a; Zhou et al., 2021) through methods like fine-tuning Visual BERT, UNITER (Li et al., 2019; Chen et al., 2020), and dual-stream encoders (Muennighoff, 2020; Sandulescu, 2020; Lu et al., 2019; Zhou et al., 2020; Tan and Bansal, 2019). Further studies address anti-semitism, propaganda, harmfulness (Chandra et al., 2021; Dimitrov et al., 2021; Pramanick et al., 2021b; Suryawanshi and Chakravarthi, 2021; Prakash et al., 2023; Sharma et al., 2022a), while recent research explores multimodal evidence prediction, role-label explanations (Sharma et al., 2023), and semantic analysis of hateful memes (Hee et al., 2023; Cao et al., 2022; Chen et al., 2023). Most of these studies are constrained by the schema and quality of the annotations while limiting the open-ended probing of memetic phenomena.

Visual Question Answering (VQA). This subsection explores the evolution of VQA research. Initial pioneering work by Antol et al. (2015) emphasized open-ended questions and candidate answers. Subsequent studies introduced variations, including joint image and question representation, to classify answers (Antol et al., 2015). Researchers further explored cross-modal interactions using various attention mechanisms, such as co-attention, soft-attention, and hard-attention (Lu et al., 2016; Anderson et al., 2018; Malinowski et al., 2018). Notably, efforts were made to incorporate common-sense reasoning (Zellers et al., 2019; Wu et al., 2016, 2017; Marino et al., 2019). Models like UpDn (Anderson et al., 2018) and LXMERT (Tan and Bansal, 2019) harnessed non-linear transformations and Transformers for VQA, while addressing language priors (Clark et al., 2019; Zhu et al., 2020). In a standard Visual-Question-Answering framework, an image is presented alongside a related question and, depending on the setup, multiple-choice options. Memes, however, introduce a more complex layer, combining images with frequently mismatched textual content, making the task more challenging and far from straightforward.

Multimodal Large Language Models. The rise of large language models (LLMs) like ChatGPT (OpenAI, 2022), GPT4 (OpenAI, 2023), Bard (GoogleAI, 2023), LLaMA (Touvron et al., 2023), Vicuna (Chiang et al., 2023), etc., has brought significant advancements in natural language understanding and reasoning. Their affinity towards multimodal augmentation is also reflected for visual-linguistic grounded tasks. Such models augment LLMs via fusion-based *adapter* layers, to excel at various tasks, from VQA to multimodal conversations (Alayrac et al., 2022; Awadalla et al., 2023; Liu et al., 2023a; OpenAI, 2023; Zhu et al., 2023; Gong et al., 2023; Zhao et al., 2023). However, existing multimodal LLMs like LLaVA (Liu et al., 2023a), miniGPT4 (Zhu et al., 2023), and multimodalGPT (Gong et al., 2023) exhibit limitations in grasping nuances like *sarcasm* and *irony* in visual-linguistic incongruity seen in memes. Although few similar works address meme-related tasks, it's mainly limited to visual-linguistically grounded settings of caption generation and VQA (Hwang and Shwartz, 2023). For a more comprehensive range of tasks, they exhibit limitations inherent to LLMs, like *pre-training biases* and *hallucinations* (Zhao et al., 2023).

The dual objectives of `MemeMQA`, encompassing answer prediction and explanation generation, present unique challenges. Existing methods fall short, including the Multimodal CoT (MM-CoT) model (Zhang et al., 2023), a two-stage framework combining DETR-based visual encoding (Carion et al., 2020) and textual encoding/decoding from `unifiedqa-t5-base`[3]. MM-CoT excels in answer prediction but falters in explanations. Instruction-tuned multimodal LLMs like LLaVA, InstructBLIP (Dai et al., 2023), and miniGPT4 show promise in understanding meme semantics but struggle with question-specific accuracy, prioritizing broader meme context over precise answers. In this work, our focus is on addressing challenges pertaining to complex visual-semantic reasoning, posed by `MemeMQA` task while considering limitations in current multimodal LLMs and neural reasoning setups for question-answering.

3 The `MemeMQACorpus` Dataset

Current meme datasets typically encompass either categorical labels (Kiela et al., 2020b; Pramanick et al., 2021a; Shang et al., 2021) or their associated explanations (Sharma et al., 2023). Although conventional Visual Question Answering (VQA) (Antol et al., 2015; Lu et al., 2016) frameworks exist, they lack the nuanced complexity of memes. These include tasks like detection, segmentation, conditional multimodal modeling (such as caption generation, visual question answering, and multiple-choice VQA), and strong visual-linguistic integration (e.g., setups similar to MS COCO for question-answering that focus on common-sense and objective reasoning) (Antol et al., 2015; Lu et al., 2016). While these areas present their distinct challenges and mark a significant line of inquiry within the intersecting realms of computer vision and natural language processing (multimodality), they fall short of addressing the complexities of multimodal *reasoning, abstract idea representation, and the nuanced use of language mechanisms like puns, humor, and figures of speech, etc.* These elements are often integral to memes. This oversight has generally curtailed the effectiveness of existing multimodal approaches (Pramanick et al., 2021b) in capturing the *nuanced complexities* inherent to memes.

To address this gap, we introduce `MemeMQACorpus`, a novel dataset designed to *emulate* free-form questioning and multiple-choice answering. Given the overwhelming diversity of possible question-answer pairs for the multifarious phenomena presented in memes, we supplement `ExHVV` (Sharma et al., 2023), an existing multimodal dataset consisting of natural

[3] https://huggingface.co/allenai/unifiedqa-t5-base

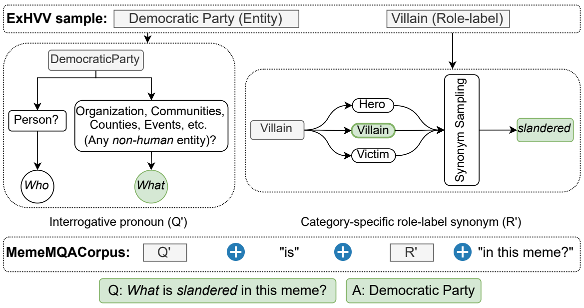


Figure 2: A schematic diagram showing question-answer construction process in `MemeMQACorpus`, using entity and role-label information from `ExHVV`.

Role-label	Counts (%)	Synonyms
hero	222 (17%)	glorified, praised, lauded, idealized
villain	1297 (59%)	vilified, berated, slandered, defamed, denounced, disparaged, maligned
victim	361 (24%)	victimised, exploited, taken advantage of, scapegoated

Table 1: The synonyms used, corresponding to the role-labels *hero*, *villain*, and *victim* (and their proportions) as part of the `MemeMQACorpus` dataset.

Prompt :

```
I will give you a question which will be about one
specific entity in one specific meme and its
categorisation. The question should not ask about any
purpose or concept, it should always be about the
entity. Paraphrase and rewrite this question in five
ways keeping the meaning and answer intact and such
that the new questions are longer than the original:
"{question}"
Just give a numbered list of new questions and
nothing else.
New questions:
```

Original Question :

```
Who is denounced in this meme?
```

Generated Questions:

```
1. Which entity is being denounced in this meme?
2. In this meme, which character is denounced?
3. What specific entity is being criticized in this
meme?
4. Who is the subject of denunciation in this meme?
5. In this particular meme, what entity is being
condemned?
```

Figure 3: Description of the prompting setup for free-form synthetic question generation using the LLM, `Llama-2-7b-chat`. The randomly chosen question option is highlighted in yellow.

language explanations for connotative roles for *three* entity types - *heroes, villains*, and *victims*, across $4,680$ instances for $3K$ memes, with automatically constructed, structured questions. This expansion aims to emulate the intricacies of meme interpretation and communication via a *QnA* setup.

Our approach entails crafting structured question sets for distinct role categories – hero, villain, and victim, each linked to unique entities featured in `ExHVV`'s memes, as depicted in Fig. 2. The central goal is to create role-based queries that precisely elicit only one entity as the correct answer within a multiple-choice setting. For instance, when an entity such as *Democratic Party* is labeled as a *villain* in a meme, along with the availability of a corresponding explanation from the `ExHVV` dataset, we formulate the question: "*What is slandered in this meme?*" (c.f. Fig. 2). With *Democratic Party* as the correct choice, distractive choices for answer options are selected based on entities referred within the meme, sampled randomly from the ones not sharing the role label with the ground truth. Sampling occurs from the entire training corpus in cases with insufficient valid entity choices. Additionally, to emulate free-form questioning by increasing the question variability, we incorporate various synonyms of role labels (hero, villain, and victim). The role-synonym mapping and their proportional breakdown, integral to constructing queries in `MemeMQACorpus`, are shown in Table 1. Our curation effort encompasses $1,880$ meme-question pairs, corresponding to $1,122$ distinct memes about *US Politics*. This domain choice is based on diversity in the entity distribution across different roles compared to the other subset (on *Covid-19*) of `ExHVV` dataset. To further examine the robustness of different modeling approaches, we curate additional variants of `MemeMQACorpus`, with (a) Question Diversification, and (b) Confounding Analysis, the details of which are presented in Sec. 7. As our question enhancement approach is automated, we achieve valid questions seamlessly linked to `ExHVV` instances, relying on the pre-existing annotations.

3.1 Prompting for Question Diversification

To achieve diversity in the framing of the original questions, a pre-trained LLM, `Llama-2-7b-chat`, is utilised for inferencing via zero-shot prompting. In this setting, the LLM is provided a context about the setting of the question which is followed by asking the model to rewrite the question in multiple ways without changing the meaning of the quetion. This ensures that the original meaning and, hence, the validity of the original option set remains intact. One out of the five rephrased questions provided by the LLM is then chosen at random. This chosen question replaces the original question in

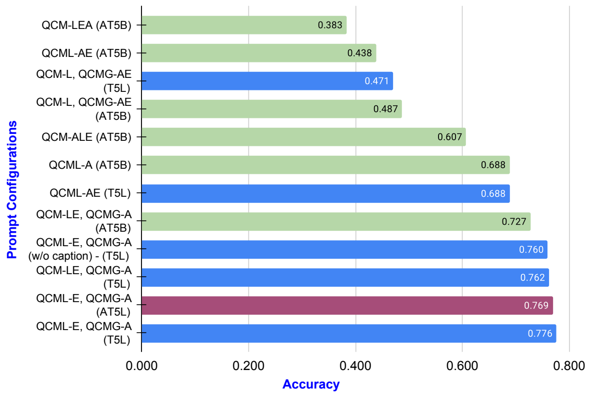


Figure 4: Comparison of various prompt configurations examined. Bar color scheme – Green: `unifiedqa-t5-base`, Magenta: `unifiedqa-t5-large`, and Blue: `t5-large`.

`MemeMQACorpus`, extending the questioning style of `MemeMQACorpus` to emulate free-form question answering more closely.

4 The `ARSENAL` Model

Prior to exploring an effective design towards addressing `MemeMQA`, we analyze different prompting configurations using meme-based inputs to determine the optimal strategy. This section begins by outlining the optimal prompting strategy, then details the structural aspects of `ARSENAL`.

Prompting Configurations: In multimodal question-answering with CoT reasoning (Zhang et al., 2023), the setup includes a question, context (text associated with an image), options, lecture (detailed generic context), explanation (a concise contextual statement), answer, and intermediate generated text.[4] Prompt configurations are represented as `input→output`, combining elements from `QCMLEAG`. Prior one-stage approaches (`QCM→LA` or `QCM→AL`) have limitations, prompting a two-stage setup with improved performance (Zhang et al., 2023). Since `MemeMQA` involves more complex reasoning than ScienceQA (Lu et al., 2022), we first examine 11 prompt configurations for `MemeMQA`, with lectures (`L`) as detailed role definitions, using one/two-stage methods and base models `unifiedqa-t5-base/large` (AT5B/L) and `t5-large` (T5L). Our findings (c.f. Fig. 4) corroborate the applicability of the two-stage framework for `MemeMQA`.[5]

4.1 System Architecture

Our input consists of three parts: (i) the meme image, $Meme_I$, (ii) the OCR Text, $Meme_T$, and (iii) the question `Q` with its corresponding multiple options `M`. The expected output consists of two parts – (i) the answer, Y_{answer}, and (ii) the explanation, Y_{exp}. We propose a multi-stage setup for `ARSENAL` to leverage individual strengths of MM-CoT and multimodal LLMs towards the overall objective of `MemeMQA`. The framework is a two-stage process consisting of answer prediction and explanation generation. It has a modular design, incorporating LLM-inferred rationale in both stages. The initial stage includes two steps: generating an intermediate rationale and predicting the answer, while the second stage focuses on generating explanations. A schematic of the proposed framework is depicted in Fig. 5.

Rationale Generation: We curate "generic rationale," $R_{generic}$, offline in the *first* stage to provide semantic information about the meme in a textual form, which is generally not captured well by the OCR information alone. $R_{generic}$ is developed using the multimodal LLM, LLaVA-7B using zero-shot inference with the prompt, $P_{generic}$ *"Explain this meme in detail."* The multimodal LLaVA-7B LLM is built on the base LLM, `Vicuna-7B`. The $R_{generic}$ thus generated captures relevant semantic information deemed useful for providing semantic clues in further stages of the proposed framework. This process can be expressed as follows:

$$R_{generic} = Model_{LLaVA}(Meme_I, P_{generic}) \quad (1)$$

In the *second* stage, LLaVA-7B model is again used to generate an "answer-specific rationale," $R_{specific}$, by prompting the model with a combination of the answer generated in the first stage and an answer-specific prompt, $P_{specific}$. $P_{specific}$ is of the form – "`How is [answer] [rephrased question]`", where the rephrased question is framed by removing the first two words of the question. For example, for a question, `Q`, given as *"Who is victimised in this meme?"* with the answer *'Joe Biden'*, the rephrased question would be given as *"How is Joe Biden victimised in this meme?"*. This is represented as,

$$R_{specific} = Model_{LLaVA}(Meme_I, P_{specific}) \quad (2)$$

[4]Unless stated otherwise, these are typically abbreviated as `QCMLEAG` – question `Q`, context `C`, multiple options `M`, lecture `L`, explanation `E`, answer `A`, and generated intermediate text `G`.

[5]Refer App. D for more details on *Prompting Configuration Assessment*.

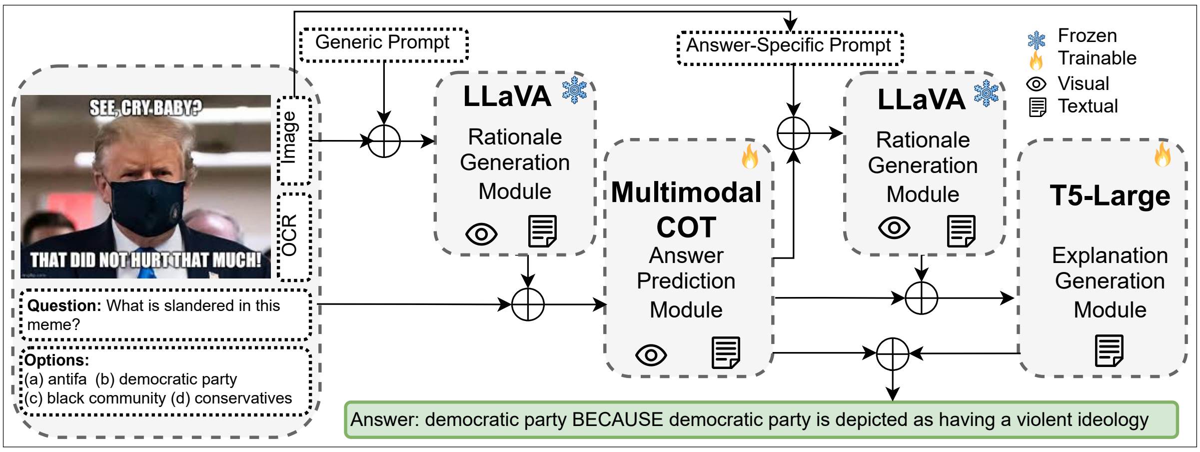


Figure 5: A schematic diagram of ARSENAL for the MemeMQA task ($\oplus$: fusing the information via concatenation).

Stage 1 - Answer Prediction: This stage implements Mutimodal CoT model with two-stage training. It uses the T5-large model with the prompting strategy of QCM→LE followed by QCMG→A. The model is provided with visual data in the form of embeddings obtained from the DETR model. These embeddings are used by adding a gated-cross attention layer in the encoder stack of the T5 model as follows: $H_{fuse} = (1-\lambda)\cdot H_{language} + \lambda \cdot H^{attn}_{vision}$, where λ is the sigmoid-activated output of fused image+text embeddings, $H_{language}$: text-input embeddings and H^{attn}_{vision}: output of text+vision cross-attention. We provide this model with additional contextual cues from $R_{generic}$. The model is then fine-tuned on MemeMQACorpus, with the first step of training for five epochs being a text generation task with the objective of generating text, G, of the form $R_{generic}$ $[SEP]$ Y_{exp}.

$$G = Model_{mm-cot}(Meme_I, Meme_T, Q, M) \quad (3)$$

This is followed by another training step for five epochs to fine-tune the model for generating Y_{answer}.

$$Y_{answer} = Model_{mm-cot}(Meme_I, Meme_T, Q, M, G) \quad (4)$$

Stage 2 - Explanation Generation: The second stage focuses on generating an explanation for the answer obtained from the previous stage. To this end, the LLaVA-7B model is used again for its superior reasoning capacity to generate an *answer-specific* rationale, $R_{specific}$. This provides us with a highly informative rationale that focuses specifically on the chosen answer and provides a highly relevant explanation. However, this generation lacks the structure and the conciseness of the expected explanation. To this end, $R_{specific}$ is provided along with the question and correct answer to a unimodal T5-large model for text-to-text generation. This T5-large model is fine-tuned for two epochs in a text-to-text generation setting for generating the expected explanation. The prompt P_{T5}, given to the T5 model, is "`Summarize the explanation for question based on the answer`". The task of T5 is as follows:

$$Y_{exp} = Model_{T5}(P_{T5}, Q, Y_{answer}, R_{specific}) \quad (5)$$

While we fine-tune it for the conditional generation objective and obtain the T5-decoder's language modeling loss $\mathcal{L}^{\text{EXP}} = -\log(p_{y_t}) = -\log(p(y_t|y_{<t}))$. The resultant explanation is combined with the previously obtained answer to obtain our final result of the form – "`Answer: [answer] BECAUSE [explanation]`".

5 Experiments

In our study, ARSENAL is rigorously tested against various models, with results averaged over five runs. The MemeMQA task involves two components: answer prediction and explanation generation, each evaluated using different metrics. Answer prediction is measured for accuracy due to entity imbalance and open-ended nature in the ExHVV dataset. Explanation quality is assessed against ExHVV ground truth using metrics like BLEU-1, BLEU-4, ROUGE-L, METEOR, CHRF, and BERTScore. Baseline comparisons span uni-modal (text, image) and multi-modal settings. Additionally, ARSENAL's robustness is evaluated through di-

Type	Models	Accuracy	BLEU-1	BLEU-4	ROUGE-L	METEOR	CHRF	BERTScore
UM	UM.TEXT.T5	0.53	0.59	0.15	0.44	0.41	0.35	0.901
	UM.TEXT.GPT3.5	0.28	-	-	-	-	-	-
	UM.IMAGE.ViT.BERT.BERT	0.46	0.51	0.10	0.45	0.44	0.38	0.911
	UM.IMAGE.BEiT.BERT.BERT	0.40	0.50	0.11	0.44	0.44	0.38	0.909
MM	MM.ViT.BERT.BERT	0.45	0.51	0.11	0.46	0.44	0.38	0.911
	MM.BEiT.BERT.BERT	0.44	0.48	0.09	0.45	0.45	0.39	0.910
	MM-CoT (w/o OCR)	0.59	0.58	0.13	0.53	0.50	0.47	0.891
	MM-CoT	0.67	0.59	0.12	0.54	0.51	**0.49**	0.894
	ViLT	0.43	-	-	-	-	-	-
	•MM-CoT (w/ Lecture)	0.69	0.59	0.13	0.54	0.51	**0.49**	0.895
	miniGPT4 (ZS)	0.32	0.09	0.00	0.14	0.21	0.23	0.753
	miniGPT4 (FT)	0.28	0.12	0.00	0.16	0.23	0.26	0.771
	LLaVA (ZS)	-	0.05	0.00	0.09	0.17	0.18	0.837
	MM-CoT (QCML→A, w/ LLaVA rationales)	0.66	0.59	0.12	0.54	0.51	**0.49**	0.896
	ARSENAL (w Entity-Specific Rationale)	**0.87**	0.58	0.17	0.53	**0.56**	0.48	0.932
	★ARSENAL (w Generic Rationale)	**0.87**	**0.63**	**0.19**	**0.55**	**0.56**	0.46	**0.934**
	$\Delta_{\star-\bullet}$(%)	18↑	4↑	4↑	1↑	5↑	1↓	2↑

Table 2: Benchmarking results for MemeVQA, comparing the proposed approach vs unimodal and multimodal baselines. Table Footnotes: **highest**, second-highest, •: MM-CoT (w Lecture) – Best Baseline, and ★: ARSENAL (proposed approach). ARSENAL variants – *(a). w Entity-Specific:* Utilizes rationale conditioned upon the answer predicted by the first module; and *(b). w Generic:* Utilizes generic rationale.

verse question types, confounding-based tests, and multimodal and error analyses.

6 Benchmarking MemeMQA

As noted in Table 2, the T5-based text-only model performs well in answer prediction with an accuracy of 0.53, outperforming image and multimodal models. However, its explanations are incomplete, repetitive, and lack coherence, resulting in low ROUGE-L (0.44), CHRF (0.35), and METEOR (0.41) scores, second only to the LLM-based miniGPT model.

The ViT model, a strong unimodal image baseline, has low answer prediction accuracy and fluent yet repetitive explanations like the T5 baseline, quantified by low ROUGE-L (0.45), METEOR (0.44) and CHRF (0.38) scores. Both ViT and BEiT unimodal baselines perform poorly, with BEiT scoring 0.40 accuracy. Multimodal baselines (ViT+BERT, BEiT+BERT) yield answer prediction accuracy (0.45 and 0.44, respectively) similar to that of ViT but slightly outperform the ViT-based unimodal model in terms of the generated explanation qualitatively. This underscores ViT's robustness over BEiT for unimodal and multimodal settings and, consequently, for the Vicuna-based miniGPT4 and LLaVA-based ARSENAL.

In the closed-form Visual Question Answering domain, we benchmark against models like the multimodal ViLT, which achieves a fine-tuned accuracy of 0.43. LLM-based models like miniGPT4 and GPT3.5 show *low* answer prediction accuracy in both zero-shot (0.32) and visual description-based fine-tuning (0.28) for the former, and 0.28 for GPT3.5. These models' explanations lack specificity, as indicated by miniGPT4's BLEU-1 score of 0.12 and ROUGE-L score of 0.16 post-fine-tuning. Despite their detailed and reasonable reasoning, they fall short in standard evaluations due to their excessive length and nonspecific content. However, BERTScore values of 0.771 for miniGPT4 (FT) and 0.837 for LLaVA-based models suggest a reasonable coherence with the memes in question.

Our primary comparison is to models leveraging the MM-CoT model in various prompt and input settings. The utility of the OCR text is proven by the 8% drop in accuracy on eliminating the OCR text from the input. The addition of *generic lectures* (L) also improves the model's performance, with a 2% increment in answer prediction accuracy. Introducing a contextual rationale using zero-shot inferencing using an LLM such as LLaVA presents qualitative improvements in the explanation generation quality of the MM-CoT model.

It is also worth noting, that the MM-COT model underperforms in understanding memes compared to the new model, which excels in accuracy and explanation due to its *Rationale Generation Module*, offering a deeper contextual grasp of meme content. Table 3 presents the average scores of ARSENAL across the seven primary metrics explored, over five independent runs.

Question: Who is disparaged in this meme?.
Options: (a) barack obama (b) donald trump (c) daily wire (d) green party

Rationale (LLaVA): In the meme, a series of images are presented with a common theme: they all seem to make fun of or mock Donald Trump. One of the images shows a **man with a pointing finger, which could represent a news story or an editorial commentary about Trump's policies or actions**. Another image displays a **man with his hands out, possibly expressing exasperation or frustration with the politician**. The meme also includes a picture of a **man with a red face, which could symbolize emotions such as anger or disapproval towards Trump**. Overall, the meme appears to take a critical stance towards Trump and his actions, **suggesting that he is being unfairly targeted or scrutinized**.

ARSENAL	- "answer: barack obama because barack obama is depicted as having committed crimes"
MM-COT	- "the answer is (b) because barack obama is portrayed as crimesining his against"
UM-Text-only	- "answer: barack obama because barack obama"
UM-Image-only	- "answer: donald trump because donald trump is portrayed as unintelligent"
MM	- "answer: donald trump because donald trump is portrayed as hateful"

Figure 6: Comparison of ARSENAL's output for a sample meme, with four baselines. The LLaVA-based rationale depicted is used for generating the explanation by ARSENAL. The font color scheme is as follows: correct, incorrect, and partially-correct.

Measures	Average	Std. Dev
Accuracy	0.87	0
BLEU-1	0.54	0.13
BLEU-4	0.15	0.06
ROUGE-L	0.50	0.08
METEOR	0.54	0.03
CHRF	36.63	20.29
BERTScore	0.92	0.02

Table 3: Averages and Std. Dev. of ARSENAL's performance measured across primary evaluation metrics, over five independent runs.

Discussion: Our analysis of 60 random test samples compared ARSENAL with other methods in terms of answer quality, explanation coherence, and modality-specific nuances. ARSENAL particularly through the LLaVA approach, excels in reasoning and explaining by effectively integrating details from various meme modalities, as shown in Figs. 6 and 12. In contrast, the MM-CoT model struggles with syntactic and grammatical correctness (c.f. Figs 6, 9, and 10). A T5-based text-only model often produces incoherent and incomplete outputs (c.f. Fig. 9). The UM.IMG.ViT.BERT.BERT model faces challenges in contextualization and alignment, with explanations that are semantically related yet irrelevant. Image-only approaches and multimodal baselines show a lexical bias, and the MM.ViT.BERT.BERT multimodal setup, despite striving for fluency, fails in complex reasoning, leading to generic explanations (refer to Figs 6 and 13).[6] The performance difference might not be as evident from a 2% quantitative increment observed for a metric like a BERTScore, relative to the 18% enhancement for answer prediction accuracy but is distinctly visible for the demonstrative example depicted in Fig. 6, and Appendix K.

[6]For more details, see App. E.

Experiment	Q_{div}	Yes/No	None (All)	None (Train)
UM.TXT.T5	0.351	0.805	0.461	0.457
UM.ViT.BERT.BERT	0.273	0.373	0.328	0.253
MM.ViT.BERT.BERT	0.341	0.295	0.474	0.438
ARSENAL	0.818	0.769	0.692	0.721

Table 4: Robustness Analysis: (a) Question Diversification (Q_{div}); (b) Confounder Setting (three scenarios).

7 Robustness Analysis

A key factor that is expected to characterize the efficacy of a model for a task like MemeMQA, is it's robustness to variations within the question/answer formulation. This is also critical due to the resultant variability within the LLM's generated responses (Salinas and Morstatter, 2024). To this end, we examine ARSENAL's performance in comparison to other contemporary baselines, by (a) Question Diversification, and (b) Confounding Analysis (c.f. Table 4).

Question Diversification: In our analysis, we evaluate the performance of ARSENAL and current baselines using more naturally framed questions than those in MemeMQACorpus. We achieve question diversity by employing the Llama-2-7b-chat model to generate five unique variations of each original question. Each question is then randomly replaced with one of these generated alternatives, ensuring a wide range of questioning styles.[7]

As an indicator of the robustness of ARSENAL to diversity in questions, when trained and tested on the new diverse questions, we obtained an answer prediction accuracy of 0.82 (c.f. Table 4). This is a marginal decline from its performance of 0.87 on the original setting, having a structured question set. In comparison, the UM.TEXT.T5 baseline descends from an accuracy of 0.53 to 0.35, UM.ViT.BERT.BERT from 0.46 to 0.27 and MM.ViT.BERT.BERT from 0.45 to 0.34. These results are a clear indication that ARSENAL is able to accommodate significant variations and diversity in the question framing setup while other models are not as robust to these changes.

Confounding Analysis: Our study evaluates the robustness of ARSENAL against contemporary baselines through three confounding settings, crafted to challenge the model with scenarios differing from typical tasks. These settings involve alterations in questions and options. We compare ARSENAL with three contemporary baseline models: UM.TEXT.T5, UM.IMAGE.ViT.BERT.BERT, and MM.ViT.BERT.BERT. Analyzing ARSENAL across the *following* settings and against these baselines is crucial for understanding its real-world applicability and performance. For detailed information on these confounding tasks, see App. J.

Confounder A – Yes/No Confounding: Transforming dataset to binary 'yes or no' questions (50% chance), reshaping 'yes' as "Is [answer] [rephrased question]?" and altering 'no' by modifying role labels.

Confounder B – None Sampling Across All Sets: Replacing 20% of answers with 'None' by swapping role labels, maintaining consistency; $M_{new} = \{M, None\}$ across sets.

Confounder C – None Sampling Across Train Only: Introducing 20% random 'None' answers in training; model adapts to 'None' while testing remains unchanged; $M_{new} = \{M, None\}$ across sets.

The 'yes or no' confounding setting in our study allows for assessing the model's reasoning robustness. Models depending on statistical probabilities fail here, as answers can be paired with either correct or incorrect role labels regardless of their dataset frequency. ARSENAL and UM.TEXT.T5 demonstrate strong reasoning skills, with scores of 0.77 and 0.80 respectively, indicating they rely on reasoning over statistics. In contrast, the UM.IMAGE.ViT.BERT.BERT-based model and MM.ViT.BERT.BERT-based model score poorly at 0.37 and 0.29, highlighting their reliance on statistical likelihoods of answers based on dataset frequency.

We also evaluated the robustness and generalizability of ARSENAL, using two settings involving "None" answers. Notably, only ARSENAL delivers good performance (0.69 accuracy) in the more challenging original testing set compared to the revised set, showcasing its better generalizability despite being trained on valid "None" answers data.

8 Conclusion and Future Work

This study introduced MemeMQA, a task that involves multimodal question answering for image-text memes, delving into their intricate visual and linguistic layers. Utilizing recently open-sourced LLMs, especially their multimodal adaptations, we tackled the challenge of complex, non-trivial multimodal content. Through a new dataset, MemeMQACorpus, we assessed systems' reasoning in assigning semantic roles to meme entities via *question-answering* and *contextualization* based objectives. Our experiments showcased the efficacy of the proposed two-stage training framework, ARSENAL, while leveraging existing language models and multimodal LLMs, to outperform the state-of-the-art by a remarkable 18% accuracy gain. This study reveals the potency and limitations of multimodal LLMs, enabling the scope for sophisticated setups embracing diverse questions, domains, and emotional nuances conveyed through memes. Ultimately, our findings steers future exploration and the development of comprehensive systems dedicated to deciphering memetic phenomena.

Our future aim is to create sophisticated, multi-perspective sets for MemeMQA, moving beyond standard QnA towards an optimal multimodal solution.

Acknowledgements

The work was supported by Wipro research grant.

Limitations

This section highlights ARSENAL's limitations, including semantically inconsistent rationales, factual errors, and multimodal bias, inherent to

[7]Refer App. 3.1 for more details on *Question Diversification*.

LLaVA's generation capacity. For some cases, LLaVA's rationales seem to be mining the inductive biases due to the co-occurrences of disparate keywords while being influenced by LLM's pre-training corpus and web data, exhibited mostly for *missing-modality* and high *inter-modal incongruity*. An example for the latter shown in Fig. 8 (c.f. Appendix I) illustrates how biased inference by LLaVA dilutes ARSENAL's output due to inaccurate contextualization, whereas MM-CoT deduces the answer accurately, possibly due to standardized definitions being used instead of LLM-based rationales.

Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Appendix A and Table 5.

User Privacy. The information depicted/used does not include any personal information.

Biases. Any biases found in the source dataset ExHVV are attributed to the original authors (Sharma et al., 2023), while the ones in the newly constructed dataset is unintentional, and we do not intend to cause harm to any group or individual.

Misuse Potential. The ability to identify implied references in a meme could enable wrongdoers to subtly express harmful sentiments towards a social group. By doing this, they aim to deceive regulatory moderators, possibly using a system similar to the one described in this study. As a result, these cleverly crafted memes, designed to carry harmful references, might escape detection, thereby obstructing the moderation process. To counteract this, it is advised to incorporate human moderation and expert oversight in such applications.

Intended Use. We make use of the existing dataset in our work in line with the intended usage prescribed by its creators and solely for research purposes. This applies in its entirety to its further usage as well. We do not claim any rights to the dataset used or any part thereof. We believe that it represents a useful resource when used appropriately.

Environmental Impact. Finally, large-scale models require a lot of computations, which contribute to global warming (Strubell et al., 2019). However, in our case, we do not train such models from scratch; instead, we fine-tune them on a relatively small dataset.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. WebSci '21, page 148–157.

Keyu Chen, Ashley Feng, Rohan Aanegola, Koustuv Saha, Allie Wong, Zach Schwitzky, Roy Ka-Wei Lee, Robin O'Hanlon, Munmun De Choudhury, Frederick L. Altice, Kaveh Khoshnood, and Navin Kumar. 2023. Categorizing memes about the ukraine conflict. In *Computational Data and Social Networks*, pages 27–38, Cham. Springer Nature Switzerland.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV'20*, pages 104–120.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans.

GoogleAI. 2023. Bard: A large language model from google.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020a. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557.

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

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. 2018. Learning visual question answering by bootstrapping hard attention. In *ECCV'18*, page 3–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *ArXiv preprint*, abs/2012.07788.

OpenAI. 2022. Chatgpt: A generative pre-trained transformer for conversational applications.

OpenAI. 2023. Gpt-4 technical report.

Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. TotalDefMeme: A multi-attribute meme dataset on total defence in singapore. In *Proceedings of the 14th Conference on ACM Multimedia Systems*. ACM.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *ArXiv preprint*, abs/2012.13235.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media.

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, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022b. Detecting and understanding harmful memes: A survey. In *IJCAI-ECAI '22*, IJCAI-ECAI '22.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022c. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381.

Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4622–4630. IEEE Computer Society.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2019, Long Beach, CA, USA, June 16-20, 2019, pages 6720–6731. Computer Vision Foundation / IEEE.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023. Chatbridge: Bridging modalities with large language model as a language catalyst. *ArXiv preprint*, abs/2305.16103.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *ICMEW*, pages 1–6.

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

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1083–1089. ijcai.org.

Modality	Model	LR	BS	# Params (M)
UM	TEXT T5	1.00E-4	4	222.9
	IMG ViT-BERT	5.00E-5	4	333.7
	IMG BEiT-BERT	5.00E-5	4	333.0
MM	ViT-BERT	5.00E-5	4	333.7
	BEiT-BERT	5.00E-5	4	333.0
	ViLT	5.00E-5	8	113.4
	MM-CoT (allenai-t5-base)	5.00E-5	4	226.6
	MM-CoT (t5-large)	5.00E-5	4	744.2
	MM-CoT (allenai-t5-large)	5.00E-5	4	744.2
	ARSENAL - Answer Prediction	5.00E-5	4	744.2
	ARSENAL - Explanation Generation	1.00E-4	4	737.6

Table 5: Hyper-parameters

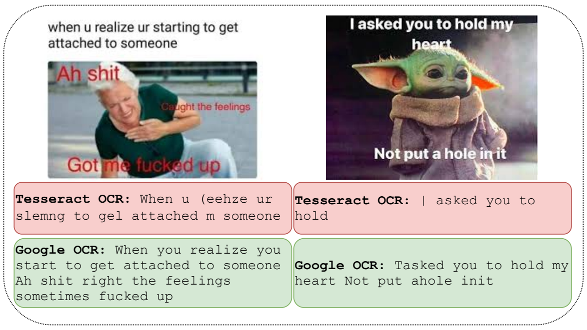


Figure 7: Comparison b/w the quality of the OCR-extracted text via (a) Tesseract OCR, and (b) Google OCR.

A Hyper-parameter and Implementation

We train all the models using PyTorch on an actively dedicated NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-12.2 and cuDNN-7.6.5 installed. For all the models with the exclusion of LLaVA and MiniGPT4, we import all the pre-trained weights from the `huggingface`[8] API. Additionally, we used a series of architectural additions and delta weights to obtain `LLaVA-7B-v0`[9] from the base `LLaMA-7B` model available under an academic license from Meta. We randomly initialize the remaining weights.

Most of our models are implemented using the Adam optimiser (Kingma and Ba, 2015) with a learning rates as specified in Table 5, a weight decay of $1e^{-5}$. We use a Cross-Entropy (CE) and a language modeling loss (LML) as per the applicability. We conducted a thorough empirical analysis before freezing the optimal set of hyperparameters for the current task for all the models examined. We also early stop to preserve our best state convergence for each experiment. Further details of hyperparameters employed can be referred to from Table 5. On average, it took approx. 2:30 hours to train a typical multimodal neural model on a dedicated GPU system.

For `ARSENAL`, we use a learning rate of $1e^{-4}$, with `eps`=($1e^{-30}$, $1e^{-3}$), `clip_threshold`=1.0, `decay_rate`=-0.8, `weight_decay`=0.0. Moreover, we set the `max_source_length` = 512 and `max_target_length` = 256 in first-step, QCM-LE task of MM-CoT, `max_target_length` = 16 in the second-step QCMG-A task of MM-CoT, and `max_target_length` = 32 in the explanation generation module using T5-Large.

A note on *fine-tuning*: Fig. 5 illustrates that the *Answer Prediction Modules* and *Explanation Generation Module* are fine-tuned (please refer to the symbol legend at the top-right corner of Fig. 5) components within the proposed framework. The *Answer Prediction Module* implementation follows guidelines from the multimodal question-answering with CoT study (Zhang et al., 2023). The LLaVa model generates a rationale that is utilized as a proxy for the intermediate rationale for MM-CoT. Furthermore, the *Explanation Generation Module* is fine-tuned using the T5-Large model, employing a training approach akin to that used for the UM_T5 model, also detailed in the implementation of the source code.

B Text Extraction via OCR

The OCR data is part of the original ExHVV dataset, as released with the original work (Sharma et al., 2023), which was extracted using Google GCV OCR[10] (GOCR) as primary inputs. The OCR for each meme is available as part of ExHVV, and we have not made any modifications to this data field. Text retrieval through *optical character recognition* (OCR) is crucial for extracting text from memes. The efficacy of the OCR method impacts the system's overall performance. Towards examining the quality of the GOCR approach used originally, we examine the text retrieval capabilities of *two* widely-used OCR-based APIs: Google Tesseract API[11] (TOCR) and Google GCV API (GOCR), for this purpose.

Our qualitative assessment of 30 varied memes reveals occasional errors in TOCR and fewer in GOCR. TOCR errors are frequent in challenging scenarios, such as overlapping text and images, low-quality graphics, or small text. In contrast,

[8]https://huggingface.co/models
[9]https://github.com/haotian-liu/LLaVA
[10]Google Cloud Vision OCR API
[11]Google's Tesseract-OCR API

GOCR often outperforms TOCR, even in simpler situations. Figure 7 illustrates the disparity in text extraction accuracy between TOCR and GOCR. The first example in Fig. 7 (*left*) shows a combination of straightforward and complex elements like clear black text on a white background and intricate visual-text overlaps, where TOCR fails but GOCR succeeds. Conversely, the second example in Fig. 7 (*right*), a simpler meme, presents more difficulties for TOCR, while GOCR maintains clarity.

C Motivation for ARSENAL's Design

Within the context of reasoning-based question-answering setup for memes, relevant solutions are scarce within the realm of neural frameworks and multimodal LLMs as we transition between them. For existing multimodal-LLM-based systems, eliciting relevant answers and generating concise explanations for memes is challenging. Strategies would typically solicit systematic instruction-tuning for fine-grained meme-related use cases instead of typical vision+language tasks (Liu et al., 2023a; Zhu et al., 2023; Zhao et al., 2023), which itself has been an active research domain. An alternate solution would be to improve existing multimodal-neural frameworks (Zhang et al., 2023) that perform reasoning-based question answering, albeit with constrained reasoning capacity and generative coherence for memes. In this work, we primarily focus on the latter while assessing the potential and limitations of other contemporary solutions.

The capability of the proposed approach (ARSENAL) towards addressing the nuanced complexity posed by memetic content stems from the choice of leveraging detailed rationales generated via multimodal LLM's, while adapting conventional approaches involving chain-of-thought reasoning which we found in our study are more suited for more accurate answer prediction and focused explanation generation.

D More on Prompting Configuration Analysis

Using the `allenai/unifiedqa-t5-base`-based MM-CoT setup, we first evaluate the optimal ordering of lecture (`L`), explanation (`E`), and answer (`A`) components for `MemeMQA`. Comparing `LEA` and `ALE` configurations, we find a significant 22% accuracy difference, emphasizing ordering importance. The two-stage setup generally outperforms one-stage, except for `QCML`→`A` (AT5B), suggesting optimal answer inference with lecture/explanation-based reasoning. The first-stage training with rationale/explanation benefits `QCM`→`LE` and `QCML`→`E` configurations. Among three language models (`unifiedqa-t5-base/large`, `t5-large`), the two-stage `t5-large` achieves the highest accuracy of 0.776, slightly better than `unifiedqa-t5-large`.

It is also worth noting that the accuracy of the two-stage framework, with configuration [`QCM`→`L`, `QCMG`→`AE`] using `allenai/unifiedqa-t5-base` comes out to be 1.5% higher than that from `t5-large`. This could be attributed to the format-agnostic design of the former, the efficacy of which could be best seen for the challenging `*`→`AE`-based scenarios in the two-stage setup (see Figure 4). In addition to this, performing inference with `AE` as outputs mainly yields poor results, as can be observed for the experiments with configurations as `QCML`→`AE` (AT5B), [`QCM`→`L`, `QCMG`→`AE`] (T5L), and [`QCM`→`L`, `QCMG`→`AE`] (AT5B), on average yielding an accuracy of 0.47. This could be due to the distributional differences between the answer choices and explanations, which the MM-CoT-based setup is unable to adjudicate as part of modeling.

A high-level overview of *prompting scenarios*: Our experiments utilize prompting across *three* distinct scenarios and configurations. Sec. 4 addresses the prompting setups for the Multimodal CoT model within the answer prediction module. The prompting structure, as explained in the paragraph on *prompting configurations* in Sec 4, follows an `input`→`output` format, with both the input and output comprising combinations of elements denoted by `QCMLEAG`. Here, `Q` stands for Question, `C` for Context, `M` for multiple options, `L` for lecture, `E` for explanation, `A` for answer, and `G` for generated intermediate text. In the ARSENAL framework, a two-stage setup is implemented, with prompts formatted as `QCM`→`LE` initially, then `QCMG`→`A`. An illustrative example is provided below: `QCM` - "Question: What is slandered in this meme?
nContext: ocr text
nOptions: (a) antifa (b) democratic party (c) black community (d) conservatives" `LE` - "Solution: lecture = generic rationale, R_generic explanation"
`QCMG` - "Question: What is slandered in this meme?
nContext: ocr text
nOptions: (a) antifa (b) democratic party (c) black

community (d) conservatives
ngenerated text" A - "The answer is (a)". The input for the explanation generation module is detailed in the description leading upto the equation # 5, as 'Summarize the explanation for question based on the answer. Explanation: R_specific or entity-specific rationale' Additionally, Sec. 3.1 elaborately discusses the prompt setups used for *question diversification*.

E Multimodal Analysis of ARSENAL

Cross-modal reasoning is a pivotal aspect of LLaVA's capability, particularly evident in situations where textual information falls short. Impressively, LLaVA harnesses its adeptness in detailed visual assessment and intricate reasoning, leading to the generation of semantically accurate rationales, as depicted in Fig. 14, 15, and 16. However, the landscape of cross-modal noise, demonstrated by the example in Fig. 17, introduces an intriguing challenge. This pertains to cases like *visual exaggeration*, where multimodal models tend to anchor their explanations across multiple modalities without a clear emphasis on a primary one, which could otherwise be self-explanatory. On a related note, the phenomenon of *multimodal hallucinations*, represented by Fig. 18, 10, 19, 20, 21, and 22, brings about an intriguing facet of LLaVA's capabilities. In these instances, the model's explanations may indeed prove accurate, despite the rationales not always aligning with factual accuracy. Such discrepancies might arise due to extrapolated ideas or statements, as well as visual misinterpretation, yet these rationales consistently maintain a high degree of semantic relevance, an observation supported by Fig. 10 and 23. In light of these intriguing insights, multimodal analysis error analysis emerges as a critical component for understanding LLaVA's performance and refining its cross-modal reasoning and explanation generation abilities.

F Difference with MM-CoT framework

The original MM-COT model, while being a strong comparative baseline, lags behind the proposed model, both in terms of answer prediction accuracy and explanation generation quality (18%-Accuracy and 2%-BERTScore performance difference w.r.t. ARSENAL), because of its inability to interact and reason well w.r.t. Visual-linguistic semantics of memes. Memes require a deeper understanding of the humour, sarcasm, and hidden meaning of the content, which the MM-COT model is observed to fall short of. The introduction of the Rationale Generation Module is a major contributing factor in the performance of the proposed framework as it provides deeper contextual information about the meme.

Approaches	WER	MEL	WIL	WIP	CER
ARSENAL	0.60	0.57	0.77	0.23	0.41
MM-CoT (w Lecture)	0.37	0.37	0.58	0.42	0.31
UM.TEXT.T5	0.67	0.65	0.82	0.18	0.53
UM.IMAGE.BEiT.BERT.BERT	0.90	0.81	0.95	0.05	0.60
MM.ViT.BERT.BERT	0.89	0.81	0.95	0.05	0.60

Table 6: Error rate comparison between ARSENAL, MM-CoT, unimodal (image and text), and multimodal baselines.

G Comparison with GPT 3.5 and GPT4

As a proxy for comparison with the closed and commercial models like GPT-3.5 and GPT-4, we have provided a comparison with open-source multimodal LLM alternatives in Table 2 in the form of a comparison with LLaVA and miniGPT4 (in zero-shot and fine-tuned settings). The primary reason for this comparison was the accessibility of the technical and background details of these systems in the public domain and to encourage healthy competition within open-source community, especially considering their impressive performance on various multimodal tasks like miniGPT4 exhibiting various emerging capabilities (Liu et al., 2023b) and LlaVA achieving SOTA on 11 benchmarks (Zhu et al., 2023), with rarely any in-depth study w.r.t content like memes, which have very strong visual-linguistic incongruity, in contrast to typically visual-linguistic grounding tasks and datasets.

H A note on Ablation Study

Our ablation analysis begins with a detailed discussion on the investigating *Prompting Configuration* (c.f. Sec. 4, second paragraph, and Fig. 4), and is then reflected as part of *Benchmarking* ARSENAL (c.f. Sec. 6, and Table 2). The specific experiments reflecting the ablation results are outlined below:

Prompting Configuration (c.f. Fig. 4): We have explored various permutations of the elements denoted by the acronym QCMLEAG (Question Q, Context C, Multiple Options M, Lecture L, Explanation E, Answer A, and Generated Intermediate Text G). These elements are crucial to the task and solution framework proposed (ARSENAL), with the goal of

Rationale: In the meme, the Democratic Party is portrayed as a joke by using a humorous symbol associated with the party. The image features a green triangle with the words "The Ban," which likely refers to the Democratic Ban on certain fireworks, and a party symbol. This visual humor is intended to make the viewer laugh and possibly make a mockery of the political party and its policies. The meme's intention is to create a negative and amusing perception of the Democratic Party, which can be considered as an example of political humor or satire. However, it is important to note that humor and satire can be subjective, and different people might find different aspects funny or offensive.

Question : What is vilified in this meme?

Options : `['democratic party', 'green party', 'scott morrison', 'bernie sanders']`

MM-COT : The answer is (B) BECAUSE the green party is portrayed as una unacceptable much restrictive thes lives.
ARSENAL: Democratic party BECAUSE democratic party is depicted as having a ban on fireworks.

Figure 8: An example of the error-type committed by ARSENAL (proposed approach) vs. the correct inferencing by the MM-COT based approach.

identifying the most effective input-output configurations for the foundational multimodal framework. These experimental explorations were carried out using initial lectures (excluding the more complex LlaVa-based justifications).

Benchmarking MemeMQA (c.f. Table 2): The experiments labeled under "Model" entries such as MM-CoT (without OCR), MM-CoT, MM-CoT (with Lecture), MM-CoT (QCML→A, with LLaVA rationales), ARSENAL (with Entity-Specific Rationale), and ARSENAL (with Generic Rationale), collectively contribute to the ablation analysis for ARSENAL. These experiments cover both the basic MM-CoT frameworks and the evolving ARSENAL variants, leading up to the solution ultimately proposed.

I Error Analysis

Among various errors in ARSENAL's outputs, we found errors due to (a) semantically inconsistent rationales caused by LLaVA, (b) factually incorrect rationales, and (c) multimodal bias. *Semantically inconsistent rationales* are prominent when high inter-modal incongruity occurs. Illustrated in Fig. 8 (c.f. Appendix I), a biased inference towards the 'democratic party' by LLaVA leads to incorrect predictions in ARSENAL. Despite a *green triangle* and the term *party* in the meme, the model lacks cues to understand context. It seems to capture inductive biases from the co-occurrence of 'party' and 'ban', likely influenced by media coverage and the LLM's training. Whereas, MM-CoT approach accurately predicts the meme's answer and produces somewhat aligned explanations. This is achieved through standardized definitions replacing rationales, aiding the T5 model's inference to connect visual elements and text to the second option.[12]

The LLMs are instruction fine-tuned for controllable behavior, so if a meme has something controversial, there is a higher chance, the LLM would attempt to normalize the harm intended within the meme, by attributing the content to the humorous and light-hearted mannerism, a typical meme is known for, which the model always seem to factor-in while generating any explanation/rationale. For instance, a couple of lines from a sample meme's explanation via a multimodal LLM states: "...It is important to note that this is a form of political humor and should not be taken seriously. The meme is simply meant to be amusing and provocative, rather than intentionally malicious or offensive." (c.f. Fig. 12). Such statements are critical w.r.t the safe deployment of such systems, yet they inhibit their capacity for pragmatic content generation.

For *quantitative assessment* of the errors commited, we compare generated text (hyp) and ground truth references (ref) in Table 6. Metrics include word error rate (WER), match error rate (MER), word information lost (WIL), word information preserved (WIP), and character error rate (CER), computed via minimum edit distance (I, S, D). distance$(D) = (I + S + D)/N$, with N as total words/characters in the reference. The error rates depicted in Table 6 elucidate the relative challenges different approaches face toward capturing the required linguistic nuances and, indirectly, the overall semantics. As expected, unimodal image-only and multimodal conventional approaches fail to emulate the reasoning necessary for producing coherent and meaningful explanations, and yield the worst scores, with an average error rate of 0.89 and 0.81, respectively. While their word information preser-

[12]For more error-type details, see Appendix. I.

vation is equally abysmal, both attain a meager score of 0.05. In contrast, a unimodal text-only system, being fundamentally built for tasks pertaining to NLU (given text-formatted input/output configurations), produce a moderate average error rate of 0.67, and a WIP score of 0.18.

The best rates are exhibited by the top two systems in our experimental suite, with MM-CoT achieving the best overall average error rate of 0.41, and a WIP score of 0.42, suggesting the potential for enhanced multimodal reasoning, with a modeling approach, not as large-sized as recent LLM-based solutions. But with the downside of the *sub-par* coherence, fluency, and complex reasoning capacity, these models do not produce explanations/answers inferencing of acceptable quality with a few exceptions as demonstrated via the example in Fig. 8, while the proposed approach (`ARSENAL`) demonstrates exceptional inferencing and rationalizing capacity, with a few critical constraints like factuality and too much detailing, while yielding second best average error rate of 0.59, with a decent WIP score of 0.23 (c.f. Fig. 6).

The one-stage approaches like the T5-based unimodal text-only model and MM models have direct accessibility to the meme's content; hence it always attempts to ground its generated explanation w.r.t the meme's content. Whereas `ARSENAL` is observed to suffer when the rationales contributing towards the explanation generation are noisy and irrelevant. This also solicits the requirement for utilizing meme text during the second stage fine-tuning as in T5 text-to-text or the conventional MM-CoT setup (c.f. Fig. 26 and 27).

J Confounding Analysis

Yes/No Confounding: In this setup, we alter `MemeMQACorpus`to shift from multiple options to a 'yes or no' format. Each question has a 50% chance of becoming a 'yes' or 'no' answer. If a question is changed to 'yes,' it's rephrased as `"Is [answer] [rephrased question]"`. For instance, *'Who is maligned in this meme?'* with *'Joe Biden'* becomes *'Is Joe Biden maligned in this meme?'* with *'yes'*. To change a question to 'no,' we adjust the role label to be incorrect for the discussed entity.

None Sampling Across All Sets: In this setup, 20% of answers are randomly changed to *None*. To implement this, semantic role labels for a meme ($\in R_{pos}$) like *hero, villain*, or *victim* (effectively their synonyms) are replaced with a synonym sampled randomly from negative role-categories ($R'_{pos} \in R_{swap}$), where, $R_{pos} \bigcap R_{swap} = \phi$. To maintain consistency, existing role labels for entities in a meme are removed, ensuring the validity of the question. The new option set, $M_{new} = \{M, None\}$, is applied to 20% of the dataset, including validation and test sets, while the entire dataset gets the new option setting.

None Sampling Across Train Only: In this new setting, compared to the previous *None* sampling, the training set incorporates a 20% random sampling of *None* answers, while validation and testing sets remain unchanged. The model now learns from data where answers are *None* with 20% probability, while during testing, *None* is never the answer. This added complexity challenges the model. The option set for any meme remains $M_{new} = \{M, None\}$ across all sets.

K Examples

The rest of the Appendix consists of multiple examples, consisting memes, questions, multiple options, generic and answer-specific rationale, model outputs from (a) `ARSENAL`, (b) MM-CoT (w Lecture), (c) Unimodal Text-only, (d) Unimodal Image-only, and (c) Multimodal baseline setups.

3 / 30 / 308 (3 / 308)

Source 0

Source 1_quest_ans	Who is disparaged in this meme?. Options: (a) barack obama (b) donald trump (c) daily wire (d) green party
Source 2_gen_rationale	The meme consists of a collage of images featuring President Trump and his policies. In one of the images, there is a caption that reads, "They keep accusing Trump of criminal acts, but they keep ignoring the fact that he's the president of criminal acts!" This statement seems to be a commentary on the ongoing discussions and scrutiny of President Trump's actions and policies. The meme also includes a few other images, making it a multi-panel display that combines both visuals and text to express the creator's opinion or perspective on the subject.
Source 3_ans_rationale	In the meme, Barack Obama is disparaged by showing a picture of him with a caption that reads, "They keep accusing Trump of criminal acts." The meme implies that people have constantly accused Trump of criminal activities during his presidency, while ignoring or downplaying Obama's actions. This statement is meant to be humorous by juxtaposing the two president's names and using a visual pun to create a critical commentary on the political climate and public perception of their respective presidencies.
Reference 0	**barack obama is** portrayed as having spied political oponents
Reference 1	**barack obama is** framed for disowning crimes
1_ARSENAL	answer: **barack obama** because barack obama **is depicted as having committed crimes**
2_mmcotlec	**the** answer **is (b)** because **barack obama is portrayed as crimesining his against**
3_umtxt	answer: barack obama because barack obama
4_umimg	answer : donald trump because donald trump **is portrayed as unintelligent**
5_mm	answer : donald trump because donald trump **is portrayed as hateful**

Figure 9: Example 3

 # 4 / 30 / 308 (4 / 308)

Source 0	
Source 1_quest_ans	Who is victimised in this meme?. Options: (a) donald trump (b) joseph goebbels (c) barack obama (d) daily wire
Source 2_gen_rationale	The meme consists of a series of images featuring President Trump and a black man. The black man appears in one image with his arms outstretched, possibly giving a speech or presentation. In another image, he seems to be responding to a question or interacting with someone off-camera. The remaining images mainly focus on President Trump, showing him in different situations, such as giving a speech or being in the spotlight. The meme seems to be highlighting the differences between the two individuals, their roles in politics, and possibly their contrasting styles or actions. The juxtaposition of the two men in different situations creates an interesting and thought-provoking visual analysis of their respective positions and the events they are involved in.
Source 3_ans_rationale	In the meme, a series of images are presented with a common theme: they all seem to make fun of or mock Donald Trump. One of the images shows a man with a pointing finger, which could represent a news story or an editorial commentary about Trump's policies or actions. Another image displays a man with his hands out, possibly expressing exasperation or frustration with the politician. The meme also includes a picture of a man with a red face, which could symbolize emotions such as anger or disapproval towards Trump. Overall, the meme appears to take a critical stance towards Trump and his actions, suggesting that he is being unfairly targeted or scrutinized.
Reference 0	**donald trump is** portrayed as being spied upon
Reference 1	**donald trump is** depicted as a victim of false allegations
1_ARSENAL	answer: **donald trump** because donald trump **is portrayed as being attacked**
2_mmcotlec	**the** answer **is (a)** because **donald trump is ed as beingcommitting racist of racism accus**
3_umtxt	answer: joseph **goebbels** because joseph **go**
4_umimg	answer : barack **obama** because barack **obama's presidency is portrayed as being catastrophic**
5_mm	answer : **donald trump** because donald trump **is portrayed as being killed**

Figure 10: Example 4

12 / 30 / 308 (12 / 308)

Source 0	
Source 1_quest_ans	Who is scapegoated in this meme?. Options: (a) slaves (b) irish (c) america (d) arvin vohra
Source 2_gen_rationale	The image features a group of seven women with red hair standing next to each other, possibly posing for a picture. They are all wearing green shirts, which suggests they might be part of an event or organization with a specific theme or color coordination. The women are arranged in a row, with some standing in the front and others in the back. Their red hair and matching attire make them appear as a cohesive group, and their presence in the image might be intended to convey a sense of unity or belonging to a particular group, cause, or event.
Source 3_ans_rationale	In the meme, the phrase "Irish were the first slaves brought over, do we get paid reparations?" plays on the idea that the Irish people were historically enslaved, which is not entirely accurate. The meme takes a jibe at the idea of reparations by suggesting that the Irish people, who have long hair and red hair, were the ones enslaved, and now they are seeking payment for the past wrongs. This is a form of scapegoating, as it unfairly targets the Irish people and perpetuates a misconception about history for the purpose of humor. It is important to recognize and acknowledge the true history and the complexities of the past to foster better understanding and promote equality and justice in the present.
Reference 0	the **irish** are shown as being enslaved
Reference 1	**irish** are portrayed to be abused as slaves.
1_ARSENAL	answer: slaves because **irish** people **are** portrayed **as being** treated unfairly
2_mmcotlec	**the** answer **is (b)** because **irish are** portrayed **as have neglecteddeprived** slaves
3_umtxt	answer: slaves because slaves **are** depicted **as being** treated unfairly
4_umimg	answer : people because **the** people **are** portrayed **as being exploited**
5_mm	answer : slaves because slaves **are** depicted **as being opressed by democrats**

Figure 11: Example 12

 # 24 / 30 / 308 (24 / 308)

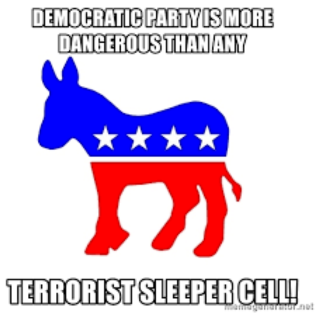
Source 0	
Source 1_quest_ans	What is berated in this meme?. Options: (a) terrorist sleeper cell (b) democratic party (c) republican (d) donald trump
Source 2_gen_rationale	The meme in the image is a playful and creative depiction of a donkey, which is associated with the Democratic Party. The donkey is shown with a bomb symbol above its head, and the words "terrorist sleeper cell" are written beneath it. This meme implies that the Democratic Party is perceived as a potential threat or as having terrorist connections. It is important to note that this is a form of political humor and should not be taken seriously. The meme is simply meant to be amusing and provocative, rather than intentionally malicious or offensive.
Source 3_ans_rationale	In the meme, the Democratic Party is portrayed as a dangerous entity, evoking a sense of caution or warning. The image features a donkey, which is commonly associated with the Democratic Party, and it is depicted as a ticking time bomb or a bomb with a fuse. The message behind this visual representation is that the meme-maker perceives the Democratic Party as a potential threat or danger, suggesting that the party's policies or actions might lead to negative consequences or even destruction. This visual representation is meant to be humorous or satirical, but it also reflects the current political climate and the polarization that often occurs around election time or in discussions about political parties.
Reference 0	**democratic party is** portrayed as a dangerous element to society
Reference 1	the **democratic party is** allegedly compared to **terrorist sleeper cell**
1_ARSENAL	answer: **democratic party** because democratic party **is portrayed as dangerous**
2_mmcotlec	the answer **is (b)** because the **democratic party is portrayed framed as besers cells**
3_umtxt	answer: **democratic party** because the democratic party **is** shown **as an unfavourable** option
4_umimg	answer : **democratic party** because the democratic party **is portrayed as spreading misinformation**
5_mm	answer : **democratic party** because the democratic party **is** shown **as a repulsive** option

Figure 12: Example 24

9 / 30 / 308 (9 / 308)

Source 0	
Source 1_quest_ans	Who is praised in this meme?. Options: (a) barack obama (b) oprah winfrey (c) monsato workers (d) democrats
Source 2_gen_rationale	The image displays a meme featuring President Obama, who is seen giving a thumbs up and smiling. The meme also includes a rating of 53% approval, which is likely a reference to the president's overall approval rating during his time in office. The meme is intended to be humorous, as it combines a positive Obama approval rating with the president's cheerful expression.
Source 3_ans_rationale	In the meme, Barack Obama is praised by being presented with a 55% approval rating. The image shows a happy Obama laughing, while the text reads, "Barack Obama's Approval Rating." The goal of the meme is to highlight the president's popularity or support from the public, using a visual representation that plays on Obama's own image. This is likely meant to convey a positive message about the former president and his tenure in office.
Reference 0	**barack obama is** depicted to have a good approval rating
Reference 1	**barack obama is praised** for his very high approval rating as a president
1_ARSENAL	answer: **barack obama** because barack obama **is** portrayed as having **a 55% approval rating**
2_mmcotlec	**the** answer **is (a)** because **barack obama is** portrayed **for his achievements favorable approval rating wella president**
3_umtxt	answer: barack obama because barack obama
4_umimg	answer : **barack obama** because barack obama **is** portrayed as **classy**
5_mm	answer : **barack obama** because barack obama **is** portrayed as having **a double standard**

Figure 13: Example 9

17 / 30 / 308 (47 / 308)

Source 0	
Source 1_quest_ans	Who is idealized in this meme?. Options: (a) kamala harris (b) donald trump (c) joseph (d) jewish
Source 2_gen_rationale	The meme consists of two images, both featuring President Trump and a young boy sitting in chairs. In one image, the President is pointing at the boy, while in the other, the boy is pointing back at the President. The meme suggests a playful interaction between the two, possibly discussing or disagreeing on political topics. The image also contains a reference to the President's catchphrase, "This is the calm before the storm," which is written on a placard in the scene. This phrase is often associated with President Trump's unpredictable and bold decision-making style.
Source 3_ans_rationale	In the meme, Donald Trump is idealized as a loving and caring leader, standing next to a young boy. The meme juxtaposes two different aspects of the president: on one hand, it shows a stern and controversial figure associated with division and criticism; on the other hand, it presents a warm and compassionate leader who cares for the child. This contrast creates a humorous and satirical portrayal of Trump, suggesting that he could be a more empathetic and nurturing figure if only he were given the chance to govern with love and kindness rather than controversy and division.
Reference 0	**donald trump is** depicted as being attacked by his political opponents
Reference 1	**donald trump is** praised for being inclusive of all religion.
1_ARSENAL	answer: **donald trump** because donald trump **is depicted as caring** for **children**
2_mmcotlec	the answer **is (b)** because **donald trump is** portrayed for **being bold of minor racess**
3_umtxt	answer: **donald trump** because donald trump **is shown**
4_umimg	answer : kamala harris because kamala harris **is** portrayed **as** the **ideal choice**
5_mm	answer : barack obama because barack obama **is depicted as having hateful views**

Figure 14: Example 47

22 / 30 / 308 (52 / 308)

Source 0	
Source 1_quest_ans	What is slandered in this meme?. Options: (a) democratic party (b) protest (c) starter pack (d) leftists
Source 2_gen_rationale	The image features a meme with various objects related to a peaceful protest. Among these objects, there is a black mask, a Molotov cocktail, a lighter, a water bottle, a baseball bat, and a badge. The meme suggests that these items are part of a "peaceful protest starter pack," implying that they could be used as tools or symbols during a non-violent demonstration. The objects are arranged in a way that emphasizes the various items in the pack, with each object placed next to each other, creating a visually striking composition.
Source 3_ans_rationale	In the meme, the Democratic Party is slandered by being associated with various negative and violent symbols. The image includes a picture of a Democratic Party starter pack, which contains a gun, a brick, a bottle, and a lighter. These items are arranged to give the impression that the Democratic Party is linked to violence, fire, and other destructive behaviors. This is a form of slander as it falsely portrays the Party and its supporters in a negative light, without any basis in fact. Such memes can contribute to the spread of misinformation and reinforce biases, so it is essential to be cautious when sharing or believing such content.
Reference 0	**democratic party is** portrayed as supporting anti-social elements
Reference 1	the **democratic party is** framed as protesting dangerously
1_ARSENAL	answer: **democratic party** because democratic party **is depicted as** having **a violent ideology**
2_mmcotlec	the answer **is (c)** because the **democratic party is** framed **as havinging withly**
3_umtxt	answer: **democratic party** because the democratic party **is shown as supporting indecent individuals**
4_umimg	answer : **democratic party** because the democratic party **is** framed **as** having **double standards**
5_mm	answer : **democratic party** because the democratic party **is portrayed as supportive of illegal immigrants**

Figure 15: Example 52

26 / 30 / 308 (56 / 308)

Source 0	 Oldie from last year but still relevant.
Source 1_quest_ans	Who is exploited in this meme?. Options: (a) black people (b) democratic party (c) arab muslims (d) nancy pelosi
Source 2_gen_rationale	The image is a political cartoon that features a group of people, some of whom are chained together, symbolizing political prisoners. The main idea is that the Democratic Party will never speak about it because it would be an uncomfortable or unpopular topic. The cartoon seems to convey a message about the current state of political discussions and the potential silence or avoidance of certain issues by politicians or the media.
Source 3_ans_rationale	The meme exploits the idea of political correctness by showing a cartoon depiction of a group of black people being held captive by another person, who appears to represent a Democratic Party official. The message implies that the Democratic Party will never speak about the issue because it would be considered politically incorrect. The meme takes a controversial and sensitive topic, such as slavery, and uses it as a form of humor to make a political statement. This type of humor can be seen as offensive and disrespectful to some viewers, especially those who have a personal connection to the issues being discussed.
Reference 0	**black people** are depicted as being enslaved
Reference 1	the **black people** are portrayed as a victim of slavery
1_ARSENAL	answer: **black people** because black people **are** portrayed **as being oppressed**
2_mmcotlec	**the** answer **is (c)** because **black** black **people are** portrayed **as beingill target of**
3_umtxt	answer: **black people** because black people **are** portrayed **as being** targetted
4_umimg	answer : **black people** because black people **are** portrayed **as being** targetted
5_mm	answer : muslims because muslims **are** portrayed **as being opressed for years**

Figure 16: Example 56

 # 21 / 30 / 308 (21 / 308)

Source 0

RIP Democratic Party

Source 1_quest_ans	What is scapegoated in this meme?. Options: (a) mainstream media (b) radical islam (c) barack obama (d) democratic party
Source 2_gen_rationale	The image features a headstone with a meme etched on it. The meme reads "Democratic Party" offended to death." This is a play on words, combining a political statement with a pun about the offended emoji. The headstone is part of a larger stone structure, and there are trees in the background.
Source 3_ans_rationale	In the meme, the Democratic Party is scapegoated by having a mock funeral organized for it. The party's name is displayed on the headstone, and the inscription reads "Democratic Party, officially offended to death." This is a play on words, as the party's name is associated with death and the end of an era. The meme is implying that the Democratic Party has been "buried" or is no longer relevant, which is a form of political humor that aims to highlight a particular issue or statement. However, it is important to note that the humor may not be suitable for everyone, and it is essential to consider the sensitivities surrounding political topics and the memebers of the party.
Reference 0	**democratic party is** depicted to have recieved a lot of hate
Reference 1	the **democratic party is** depicted as a target for hateful trolls.
1_ARSENAL	answer: **democratic party** because democratic party **is** portrayed as being **killed**
2_mmcotlec	the answer **is (d)** because the **democratic party is ed** as **beingcommitting target for off peopleshugss**
3_umtxt	answer: **democratic party** because the democratic party **is depicted** as being **insulted**
4_umimg	answer : **democratic party** because the democratic party **is** portrayed as being **full of hateful people**
5_mm	answer : **democratic party** because the democratic party **is** portrayed as **having met its demise**

Figure 17: Example 21

2 / 30 / 308 (2 / 308)

Source 0

Source 1_quest_ans	Who is denounced in this meme?. Options: (a) donald trump (b) barack obama (c) kamala harris (d) democratic party
Source 2_gen_rationale	The meme is a humorous image featuring an edited picture of former President Barack Obama with a disappointed or funny expression, along with a First Name tag. The meme has been photoshopped to make it appear as though Obama is reacting to his own name being displayed on the "Disappointed" banner. The image is captioned with "First Name" and is set against a backdrop of the U.S. Capitol building, which adds to the playful and lighthearted nature of the meme.
Source 3_ans_rationale	In the meme, Barack Obama is denounced by showing a picture of him with a red, white, and black caption above his image. The caption reads, "First Name Ever," which implies that the meme is mocking Obama's famous statement, "First Name Ever." This suggests that the meme is created to make fun of the president's name and popular catchphrase, rather than genuinely conveying any political message or dissatisfaction with his policies. The meme is intended to be humorous, rather than thought-provoking or insightful.
Reference 0	**barack obama is** depicted as a disappointment
Reference 1	**barack obama is** portrayed as ill-suited for president
1_ARSENAL	answer: **barack obama** because barack obama **is depicted as an unsuitable choice for** the **presidency**
2_mmcotlec	the answer **is (b)** because **barack obama is** portrayed **as havinga-suited**
3_umtxt	answer: barack obama because barack obama
4_umimg	answer : donald trump because donald trump **is** portrayed **as being incompetent**
5_mm	answer : **barack obama** because barack obama **is depicted as engaging in inappropriate actions**

Figure 18: Example 2

5 / 30 / 308 (5 / 308)

Source 0	
Source 1_quest_ans	What is glorified in this meme?. Options: (a) democratic party (b) communist party (c) republican party (d) kegger party
Source 2_gen_rationale	The image consists of four portraits of famous historical figures, each of them associated with a specific political party. The first person is Karl Marx, who is often linked to the Democratic Party. The second person is John Adams, who is typically connected to the Republican Party. The third person is Alexander Hamilton, who is commonly associated with the Democratic Party as well. The fourth person is Thomas Jefferson, also a part of the Democratic Party. The idea behind the meme is to create a humorous visual by combining the faces of these historical figures with the names of the political parties they are associated with. The meme is intended to be funny because it is not typical to see these people's faces connected to specific political parties, as they are more commonly known for their individual achievements and contributions rather than their political affiliations.
Source 3_ans_rationale	In the meme, the Democratic Party is glorified by showing three different images of the party's founder, Karl Marx, with one of the images having an American flag superimposed over him. This is an attempt to associate the values and principles of the Democratic Party with the iconic figure of Karl Marx, thereby implying that the party still upholds his ideologies and legacy. The inclusion of the American flag in one of the images suggests a connection between the party and the nation's flag, emphasizing the party's identification with American values and patriotism.
Reference 0	**communist party is** portrayed **in** a positive light
Reference 1	the **communist party is** regarded as a distinct preference.
1_ARSENAL	answer: democratic **party** because democratic party **is portrayed** as the **ideal** choice
2_mmcotlec	the answer **is (c)** because **commun communist party is shownportrayed** as **thea** suitable party
3_umtxt	answer: republican **party** because the republican party **is depicted** as **a**
4_umimg	answer : shrek super **party** because shrek super party **is** shown as **a** suitable **option**
5_mm	answer : democratic **party** because the democratic party **is** shown as **an unfavourable** choice

Figure 19: Example 5

7 / 30 / 308 (7 / 308)

Source 0	
Source 1_quest_ans	What is maligned in this meme?. Options: (a) donald trump (b) democrats (c) xi jinping (d) covid19
Source 2_gen_rationale	The meme in the image is a playful alteration of a photograph featuring a Chinese politician, Xi Jinping, who is shown with a humorous text superimposed over his face. The text reads, "We deceitfully hid the 9/11 whistler." The reference to 9/11 and the whistler is a play on Xi's name, which sounds similar to "911" and is a nod to the famous phone call between him and former US President George W. Bush after the 9/11 attacks. The meme is a creative way of mixing politics and humor, but it should be noted that it may not be suitable for all audiences due to its political nature and could be considered offensive to some.
Source 3_ans_rationale	In the meme, a Chinese politician, possibly President Xi, is depicted with a caption that reads, "We deceitfully hid the 98th whistle-blower." The meme implies that the politician has been dishonest and tried to cover up a scandal involving the 98th whistle-blower. It suggests that the politician has been involved in a conspiracy to suppress the truth, possibly related to corruption or other wrongdoing. The meme is a form of political humor that uses irony and exaggeration to make a statement about the politician's alleged actions and character.
Reference 0	**democrats** are insinuated to wrongly accuse their opponents
Reference 1	the **democrats** are accused of improperly deflecting any blame.
1_ARSENAL	answer: **democrats** because democrats **are depicted** as **being incompetent**
2_mmcotlec	the answer **is (b)** because the **democrats are of beingly targetingfrauing** votes votes **onto**
3_umtxt	answer: **democrats** because the democrats **are** portrayed as **supporting stupid policies**
4_umimg	answer : **democrats** because the democrats **are** portrayed **to have colluded with foreing interests**
5_mm	answer : **democrats** because democrats **are framed** as **conspiring to spread** the **coronavirus**

Figure 20: Example 7

11 / 30 / 308 (11 / 308)

Source 0

Source 1_quest_ans	Who is berated in this meme?. Options: (a) democrats (b) democratic party (c) donald trump (d) kids
Source 2_gen_rationale	The meme in the image is a humorous speech bubble featuring a politician giving a speech to a large crowd of people. The politician appears to be a caricature of former President Bill Clinton, and the speech bubble contains a playful message. The meme reads, "You've always wondered what happened to the slow kids?" This implies that the politician is joking about the audience's slow kids, using a light-hearted tone to engage the crowd. The meme is a playful and entertaining way to express humor in a political context.
Source 3_ans_rationale	In the meme, a large crowd of people is gathered, and they are all looking at a man who appears to be a public speaker. The meme cleverly takes a shot at Donald Trump by using a picture of him with a text that reads, "You've always wondered what happened to the slow kids?" The implication is that the crowd is reacting to a speech that Trump gave, and the meme suggests that the reaction might be due to his perceived slow or unintelligent delivery. The humor in this meme is derived from the juxtaposition of the serious political event with the playful visual content, which is a common practice in internet culture to critique or satirize public figures.
Reference 0	**donald trump is** depicted to be supported by unintelligent people
Reference 1	**donald trump is** portrayed as having inept followers.
1_ARSENAL	answer: **donald trump** because donald trump **is depicted** as **slow**
2_mmcotlec	**the** answer **is (c)** because **donald trump is portrayed** as **being misconsider policies**
3_umtxt	answer: **donald trump** because donald trump **is**
4_umimg	answer : **donald trump** because donald trump **is insinuated** as **hateful**
5_mm	answer : **donald trump** because donald trump **is depicted** as **stupid**

Figure 21: Example 11

 # 26 / 30 / 308 (56 / 308)

Source 0	 Oldie from last year but still relevant.
Source 1_quest_ans	Who is exploited in this meme?. Options: (a) black people (b) democratic party (c) arab muslims (d) nancy pelosi
Source 2_gen_rationale	The image is a political cartoon that features a group of people, some of whom are chained together, symbolizing political prisoners. The main idea is that the Democratic Party will never speak about it because it would be an uncomfortable or unpopular topic. The cartoon seems to convey a message about the current state of political discussions and the potential silence or avoidance of certain issues by politicians or the media.
Source 3_ans_rationale	The meme exploits the idea of political correctness by showing a cartoon depiction of a group of black people being held captive by another person, who appears to represent a Democratic Party official. The message implies that the Democratic Party will never speak about the issue because it would be considered politically incorrect. The meme takes a controversial and sensitive topic, such as slavery, and uses it as a form of humor to make a political statement. This type of humor can be seen as offensive and disrespectful to some viewers, especially those who have a personal connection to the issues being discussed.
Reference 0	**black people** are depicted as being enslaved
Reference 1	the **black people** are portrayed as a victim of slavery
1_ARSENAL	answer: **black people** because black people **are** portrayed **as being oppressed**
2_mmcotlec	**the** answer **is (c)** because **black** black **people are** portrayed **as beingill target of**
3_umtxt	answer: **black people** because black people **are** portrayed **as being** targetted
4_umimg	answer : **black people** because black people **are** portrayed **as being** targetted
5_mm	answer : muslims because muslims **are** portrayed **as being opressed for years**

Figure 22: Example 45

10 / 30 / 308 (10 / 308)

Source 0	
Source 1_quest_ans	Who is slandered in this meme?. Options: (a) donald trump (b) democratic party (c) russians (d) rural america
Source 2_gen_rationale	The image features a red and white sign on a fence, which reads "Rural America Party." This sign is a political statement, possibly associated with the Democratic Party. Beyond the fence, there is a field with a mix of vehicles, including cars and trucks, parked or driving by. The scene suggests that the vehicles might be attending an event or visiting a rural area in America. The presence of the sign and the vehicles in the background indicates that this location could be a gathering place or an area of interest for political activities or events.
Source 3_ans_rationale	In the meme, a sign is displayed with a picture of Donald Trump and a message accusing him of being a "racist, sexist, homophobic, xenophobic, and climate change denier." The meme implies that the creators of the sign are politically opposed to Trump and are using this sign to express their disagreement with his views and policies. The meme is intended to convey the message that the sign-carrying creators do not agree with Trump's views and consider him to be a negative figure.
Reference 0	**donald trump is** portrayed as unsuitable for being president
Reference 1	**donald trump is** insinuated for his election to the presidency.
1_ARSENAL	answer: **donald trump** because donald trump **is depicted as an unsuitable choice for president**
2_mmcotlec	the answer **is (a)** because **donald trump is sinuated as wrong crimes** the **presidency**
3_umtxt	answer: **donald trump** because donald trump **is depict**
4_umimg	answer : **donald trump** because donald trump **is insinuated as** the **worst president in hisptry**
5_mm	answer : **donald trump** because donald trump **is framed as a traitor**

Figure 23: Example 10

1 / 30 / 308 (1 / 308)

Source 0

Source 1_quest_ans	Who is maligned in this meme?. Options: (a) vacation (b) barack obama (c) bernie sanders (d) democrats
Source 2_gen_rationale	The meme features a picture of President Obama with a caption that reads, "I didn't catch that, but thanks for the thought about the next 30 million dollars worth of vacation that you're actually paying for." The meme humorously suggests that President Obama is not paying attention to the question asked and is instead thinking about the cost of the next 30 million dollars worth of vacation.
Source 3_ans_rationale	In the meme, Barack Obama is maligned by being portrayed as a villain or a cartoon character with devil horns, resembling a typical evil character. The meme reads, "I didn't catch that, thank you, Mr. President," which implies that the person responsible for the meme is mocking the former president's policies or actions. This type of meme is a form of political humor that aims to convey a negative opinion about Obama's leadership or administration by using exaggerated and satirical elements.
Reference 0	**barack obama is** framed as misuing public money
Reference 1	**barack obama is** being accused of taking unjustified vacations on taxpayer money.
1_ARSENAL	answer: **barack obama** because barack obama **is depicted as being wrongly blamed**
2_mmcotlec	**the** answer **is (b)** because **barack obama is of over toosuit vacation vacations hiss**
3_umtxt	answer: barack obama because barack obama
4_umimg	answer : **hillary** clinton because **hilary** clinton **is framed as** a **whiner**
5_mm	answer : **barack obama** because barack obama **is framed as having** a **liar**

Figure 24: Example 1

30 / 30 / 308 (60 / 308)

Source 0

(GC)

Source 1_quest_ans	What is exploited in this meme?. Options: (a) democratic party (b) republicans (c) black community (d) joe biden
Source 2_gen_rationale	The image is a meme featuring a black and white picture of a man pointing a gun at another man who is sitting in a chair, with the phrase "how could the republicans do this" written on it. The meme is a play on the idea of gun control and political affiliations. The man pointing the gun represents the Republicans, while the sitting man symbolizes the Democratic Party. The meme humorously suggests that the Republicans might try to enact gun control policies supported by the Democratic Party.
Source 3_ans_rationale	In the meme, a Democrat Party member is pointing a gun at a Republican Party member and pulling the trigger. The target, a stuffed animal or dummy representing the Republican Party, is depicted as being shot. This visual humor is playing on the idea of a political "shooting" or conflict between the two parties, with the Democrat Party member being the aggressor. The meme is intended to be funny by taking a serious situation and portraying it in a lighthearted and satirical manner.
Reference 0	republics are portrayed as being wrongly accused
Reference 1	the **republicans** are depicted to be blamed for false allegations
1_ARSENAL	answer: republicans because republicans **are portrayed as being shot**
2_mmcotlec	the answer is **(b)** because the republicans **are ed as have targetingd for** the **ideas**
3_umtxt	answer: republicans because republicans **are portrayed as being insulted**
4_umimg	answer : democratic party because the democratic party is **portrayed as being influenced** by **corporate money**
5_mm	answer : republican party because the republican party is **depicted as being attacked** by **political opponents**

Figure 25: Example 60

6 / 30 / 308 (6 / 308)

Source 0

Source 1_quest_ans	Who is disparaged in this meme?. Options: (a) covid19 (b) americans (c) democrats (d) xi jinping
Source 2_gen_rationale	The meme in the image displays a picture of a man, likely a Chinese politician, with a caption that reads "We deceitfully hid the 98th whistle-blower." The meme seems to be making a joke or a statement about a whistle-blower scandal involving the politician. The politician is wearing a suit and tie, giving the impression that the meme is taken from a professional setting. The image is a play on words, using "whistle-blower" as a metaphor for a person who exposes wrongdoing, and "98th" referring to the 98th whistle-blower, implying that there are multiple whistle-blowers involved in the scandal.
Source 3_ans_rationale	In the meme, xi jinping, the Chinese president, is disparaged by being shown in a photo with a humorous message. The message reads, "We deceitfully hid the 90's whistle and made it look like a Democrat would blame Trump for it." The meme implies that xi jinping is being blamed for a problem that occurred during the 1990s, which is attributed to a Democratic politician, and it suggests that the politician would blame Trump for the issue. This is a playful and sarcastic way of implying that the politician would accuse Trump of causing the problem, even though it happened in the 1990s, and Trump was not yet in office.
Reference 0	**xi jinping is** framed as hiding critical information from the world
Reference 1	**xi jinping is** accused of tampering with the covid-19 news.
1_ARSENAL	answer: **xi jinping** because xi jinping **is** depicted **as being wrongly blamed**
2_mmcotlec	the answer **is (d)** because **xi jinping is of beingsampering** with the **electionvid19**
3_umtxt	answer: xi jinping because xi jinping
4_umimg	answer : **xi jinping** because xi jinping **is** depicted **to be interfering** with **american politics**
5_mm	answer : barack obama because barack obama **is framed as unintelligent**

Figure 26: Example 6

13 / 30 / 308 (13 / 308)

Source 0

Source 1_quest_ans	What is disparaged in this meme?. Options: (a) barack obama (b) democratic voters (c) democratic party (d) hillary clinton
Source 2_gen_rationale	The image is a meme featuring headshots of nine politicians, both Democrat and Republican, with text stating "The Radicalization of the Democratic Party by these UnAmerican Sheep." The meme implies that the politicians are extreme or have shifted towards more radical positions. The image is designed to be humorous and to provoke a reaction from the viewers by using strong political rhetoric.
Source 3_ans_rationale	In the meme, a series of portraits of Democratic voters are presented as caricatures, which can be considered as a form of political satire or humor. However, some might perceive it as a means to belittle or disparage these voters. The meme shows the various portrayals of people who are likely to be supporting Democratic candidates or policies. By presenting these voters in a comical or exaggerated manner, the meme might be attempting to elicit a negative or sarcastic response from the viewers towards the Democratic Party or its supporters. It is essential to be aware that such depictions can be interpreted in different ways, and some people might find them offensive or biased.
Reference 0	**democratic party is** portrayed as being radicalized
Reference 1	**democratic party is** villified for its alleged radicalization.
1_ARSENAL	answer: democratic voters because democratic voters are depicted **as unfavourable**
2_mmcotlec	the answer **is (c)** because the **party is ctorified as radical radicalradi radicalization**
3_umtxt	answer: democratic voters because democratic voters are depicted **as being radicalised**
4_umimg	answer : **democratic party** because the democratic party **is framed as manipulative**
5_mm	answer : **democratic party** because the democratic party **is portrayed as having a weak leadership**

Figure 27: Example 13