M3Hop-CoT: *M*isogynous Meme Identification with *M*ultimodal *M*ulti-hop *C*hain-of-*T*hought

Gitanjali Kumari[1] **Kirtan Jain** [1] **Asif Ekbal**[2]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
[2]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
[1]{gitanjali_2021cs03,kirtan_2101cs38}@iitp.ac.in,[2]asif@iitj.ac.in

Abstract

In recent years, there has been a significant rise in the phenomenon of hate against women on social media platforms, particularly through the use of misogynous memes. These memes often target women with subtle and obscure cues, making their detection a challenging task for automated systems. Recently, Large Language Models (LLMs) have shown promising results in reasoning using Chain-of-Thought (CoT) prompting to generate the intermediate reasoning chains as the rationale to facilitate multimodal tasks, but often neglect cultural diversity and key aspects like emotion and contextual knowledge hidden in the visual modalities. To address this gap, we introduce a **M**ultimodal **M**ulti-hop CoT (M3Hop-CoT) framework for **M**isogynous meme identification, combining a CLIP-based classifier and a multimodal CoT module with entity-object-relationship integration. M3Hop-CoT employs a three-step multimodal prompting principle to induce emotions, target awareness, and contextual knowledge for meme analysis. Our empirical evaluation, including both qualitative and quantitative analysis, validates the efficacy of the M3Hop-CoT framework on the SemEval-2022 Task 5 (**MAMI task**) dataset, highlighting its strong performance in the macro-F1 score. Furthermore, we evaluate the model's generalizability by evaluating it on various benchmark meme datasets, offering a thorough insight into the effectiveness of our approach across different datasets [1].

1 Introduction

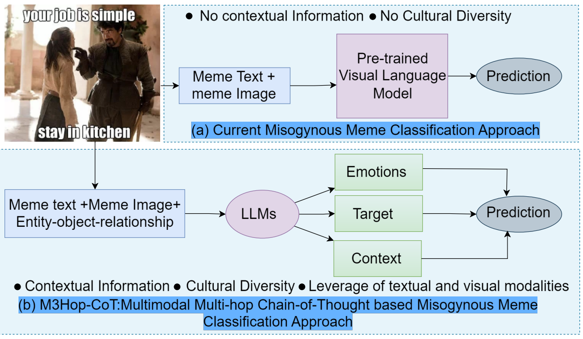

your job is simple
stay in kitchen
● No contextual Information ● No Cultural Diversity
Meme Text + meme Image
Pre-trained Visual Language Model
Prediction
(a) Current Misogynous Meme Classification Approach
Meme text +Meme Image+ Entity-object-relationship
LLMs
Emotions
Target
Context
Prediction
● Contextual Information ● Cultural Diversity ● Leverage of textual and visual modalities
(b) M3Hop-CoT:Multimodal Multi-hop Chain-of-Thought based Misogynous Meme Classification Approach

Figure 1: Comparison between (a) fine-tuning visual language model approach and (b) Chain-of-Thought based approach.

In recent years, the proliferation of memes on social media platforms like Facebook, Twitter, and Instagram has gained significant attention due to their widespread influence and potential to shape public discourse. While many memes are created for entertainment, some serve political or activist purposes, often employing dark humor. Misogynous memes[2], however, stand apart by propagating hatred against women through sexist and aggressive messages on social media (Attanasio et al., 2022; Zhou et al., 2022; Arango et al., 2022). These memes exacerbate sexual stereotyping and gender inequality, mirroring offline societal issues (Franks, 2009) and have become a concerning issue (Chen and Chou, 2022; Zhang and Wang, 2022; Fersini et al., 2022).

Identifying misogynous memes is much more challenging than other memes, as the task demands an understanding of world knowledge and common sense (Singh et al., 2024). Despite the challenges, developing deep learning models to classify such memes can provide sociological benefits, such as understanding hidden meanings, supporting humanities research, and raising awareness on a large scale (Kumari et al., 2024). Previous research has primarily focused on developing robust deep-learning models that learn cross-modal interactions (c.f. Figure 1 (a)) from scratch to identify these memes (Rijhwani et al., 2017; Sharma et al., 2020; Kiela et al., 2020a; Suryawanshi et al.,

---

[1] Codes are available at this link: https://github.com/Gitanjali1801/LLM_CoT

[2] WARNING: This paper contains meme samples with slur words and sensitive images.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22105–22138
November 12-16, 2024 ©2024 Association for Computational Linguistics

2020; Pramanick et al., 2021a; Hossain et al., 2022; Sharma et al., 2022a). However, learning complex multimodal interactions can be difficult with limited data (Cao et al., 2023). The advent of Large Language Models (LLMs) offers a way to bridge this gap. Although LLMs are highly adept at question-answering and reasoning tasks, they often overlook the cultural diversity of human reasoners, crucial for tasks demanding commonsense, contextual knowledge, and multimodal reasoning (Li and Zhang, 2023; Cao et al., 2024). However, the recent concept of **Chain-of-Thought (CoT)** has demonstrated the potential of LLMs for multi-hop reasoning (Wei et al., 2023; Fei et al., 2023; Wu et al., 2023), showing that LLMs can perform chain-style reasoning effectively with the right prompts. Nonetheless, most CoT reasoning studies focus primarily on the language modality (Lu et al., 2023; Wang et al., 2022), often overlooking multimodal contexts. Analyzing memes is particularly challenging because their implicit meanings are not fully conveyed through text and images. In such a scenario, neglecting one modality in meme detection can negatively impact model performance.

As depicted in Figure 1 (b), if an LLM can not only interpret emotions, such as anger or disgust from the text, “Your job is simple. Stay in Kitchen," but also analyze the visual elements of the meme featuring a woman and a man, which would further enhance its ability to recognize emotions by considering their facial expressions and body language. This is crucial for identifying sexist stereotypes. Moreover, the LLM can determine if the meme targets women by evaluating both textual and visual modalities. Furthermore, understanding the broader context, which encompasses societal and cultural discussions on gender roles and equality, is crucial, and this can also be achieved using LLMs. To achieve this, the ability to perform **multi-hop reasoning Chain-of-Though(CoT)** (i.e., inferring emotion, target, and then understanding the context) is indispensable. By hierarchically considering key aspects of misogynous memes, such as *emotions*, *targets*, and *contextual backgrounds*, we can create general-purpose models that are in sync with human intent for real-world tasks like meme identification.

Our proposed work is motivated by the aforementioned discussion, where we introduce a deep learning-based framework named M3Hop-CoT (**M**isogynous **M**eme Identification with **M**ultimodal **M**ulti-**hop** **C**hain-of-**T**hought) a modular approach that leverages an LLM as the “reasoning module” and operates over a given meme. In the M3Hop-CoT approach, we first extract the Entity-object-relationship (EORs) of the meme-image using a scene graph (Tang et al., 2020). Subsequently, the meme text, image, and EORs are fed into the multi-hop CoT LLM, enabling it to identify three crucial hidden cues for inferring the meme’s rationales: (i) emotion, (ii) target, and (iii) context. M3Hop-CoT eliminates the need for external resources, also bridging the gap between the modalities by utilizing both textual and visual aspects of the meme in rational generation at zero cost. To ensure the weighted contribution of each reasoning step, we employ a hierarchical cross-attention mechanism that assesses the contribution of each rationale in decision-making.

The main contributions of this work are summarized below: (i) This is the first study where we introduce multimodal LLM in a CoT manner to identify the misogynous memes. (ii) We introduce the M3Hop (Misogynous Meme Identification with Multimodal Multi-hop Chain-of-Thought) framework, where we utilize the meme text and EORs of the meme-image as a prompt to the LLM in a multi-hop CoT manner, enabling it to identify three crucial rationales helpful to detect misogynous memes: (a) emotion, (b) target, and (c) context. (iii) Our empirical evaluation, including both qualitative and quantitative analysis, validates the efficacy of the M3Hop-CoT framework on several datasets, highlighting its strong performance.

2 Related Work

Detection of Misogynous memes. Previous studies on memes have predominantly focused on identifying hate or offensive content (Rijhwani et al., 2017; Sharma et al., 2020; Kiela et al., 2020a; Suryawanshi et al., 2020; Sharma et al., 2022a; Hossain et al., 2022; Yadav et al., 2023). While most of the existing meme research has focused on refining multimodal representations by exploring interactions between textual and visual elements (Kumari et al., 2021; Akhtar et al., 2022; Sharma et al., 2022b; Bandyopadhyay et al., 2023; Sharma et al., 2023), still error analyses in these studies have revealed a significant gap in the contextual comprehension of memes (Cao et al., 2022). While existing research on detecting misogynous content has largely focused on unimodal data (primarily text) (Hewitt et al., 2016; Fersini et al., 2018; Nozza et al., 2019), the integration of multimodality (text and image), on the other hand, is still a work in progress (Zhou et al., 2022; Zhi et al., 2022; Arango et al., 2022; Singh et al., 2023).

Large Language Models. Pre-training of language models has garnered significant interest for its ability to enhance downstream applications (Raffel et al., 2019). Recently, large-scale language models (LLMs), such as GPT-3 (Kojima et al., 2023), ChatGPT (Ouyang et al., 2022), LLaMA (Touvron et al., 2023) etc., have demonstrated remarkable potential for achieving human-like intelligence. LLMs have shown exceptional capabilities in common-sense understanding (Paranjape et al., 2021; Liu et al., 2022) with the incorporation of the chain of thought (CoT) method which has revolutionized the way machines approach reasoning-intensive tasks (Wei et al., 2023; Zhou et al., 2023). While previous research has often struggled with incorporating external knowledge and common-sense understanding and has been limited by fitting spurious correlations between multimodal features, our proposed model **M3Hop-CoT** bridges this gap with a novel prompt-based approach by employing a *Multimodal Multihop CoT* based approach to simultaneously analyze the meme text and the entity-object relationship of the meme image, thereby deciphering the emotional, targeted, and contextual dimensions of a misogynous meme. By doing so, we aim to integrate culturally diverse reasoning into our proposed misogynous meme classifier.

3 Dataset

For our experiments, we employ two misogynous meme datasets: MAMI (SemEval2022 Task 5, Subtask A) (in English) (Fersini et al., 2022) and MIMIC (in Hindi-English Code-Mixed) (Singh et al., 2024) (Refer to Table 1). To demonstrate the generalizability of our CoT-based approach, we conduct experiments on three benchmark meme datasets: Hateful Memes (Kiela et al., 2020b), Memotion2 (Ramamoorthy et al., 2022), and Harmful Memes (Sharma et al., 2022a) (See Appendix Table 7 for data statistics)

Dataset	Train set	Test set	Task
MAMI	10,000	1,000	Misogynous Meme Detection
MIMIC	4,044	1,010	Misogynous Meme Detection
Hateful Meme	8,500	1,000	Hateful Meme Detection
Memotion2	7,500	1,500	Offensive Meme Detection
Harmful Meme	3,013	354	Harmful Meme Detection

Table 1: Dataset Statistics

4 Methodology

This section illustrates our proposed M3Hop-CoT model to identify the misogynous meme. The overall workflow of our proposed *M3Hop-CoT* model is shown in Figure 2, and its components are discussed below.

4.1 Problem Formulation

Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ represent the dataset of misogynous memes, where N is the number of samples, $x_i \in \mathcal{X}$ is the i-th meme (comprising text and images), and $y_i \in \{0, 1\}$ is its corresponding misogyny label (1 for misogynous, 0 for non-misogynous). Our objective is to train a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ to predict correct misogynous label $\hat{\mathcal{Y}}$, parameterized by θ, to minimize a loss function $\mathcal{L}(\hat{\mathcal{Y}}|\mathcal{X}, \theta)$, defined over the output space $\mathcal{Y}$ and the predicted label $\hat{\mathcal{Y}}$.

4.2 Encoding of Meme

A meme sample $\mathcal{X}_i$ comprises of meme text $T_i = (t_{i_1}, t_{i_2}, \ldots, t_{i_k})$, which is tokenized into sub-word units and projected into high-dimensional feature vectors, where k is the number of tokens in the meme text, and image I_i with regions $r_i = \{r_{i_1}, r_{i_2}, \ldots, r_{i_N}\}$; for $r_{i_j} \in R^N$, where N is the number of regions. These components are input into a pre-trained CLIP model (Radford et al., 2021) designed to extract features by understanding text and images at a semantic level.

$$ft_i, fv_i = CLIP(t_i, r_i)\ ; \qquad (1)$$

where $ft_i \in \mathbb{R}^{d_t}$ and $fv_i \in \mathbb{R}^{d_v}$ are the extracted text and visual features, respectively, with d_t and d_v denoting the dimensions of the text and visual feature spaces. To integrate these features, we use the Multimodal Factorized Bilinear (MFB) pooling technique (Kumari et al., 2021; Bandyopadhyay et al., 2023). The interaction between textual and visual features was limited in earlier fusion techniques (Zhang et al., 2022) (e.g., concatenation, element-wise multiplication, etc.). These methods did not allow for comprehensive interaction between textual and visual features, essential for generating robust and nuanced multimodal features. Bilinear pooling, while effective in capturing detailed associations between textual and visual features through outer products, introduces high computational costs and risks of overfitting due to the large number of parameters required (Yu et al.,

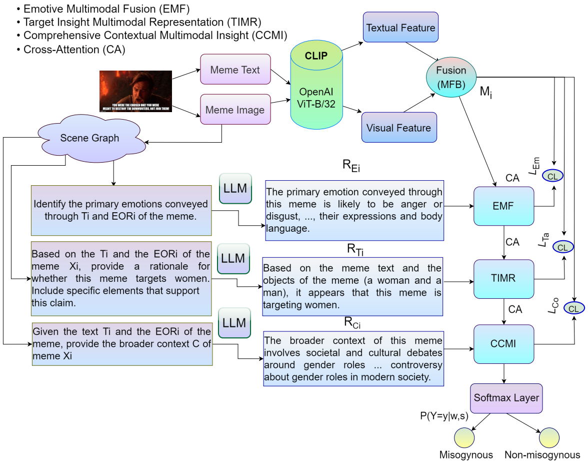

• Emotive Multimodal Fusion (EMF)
• Target Insight Multimodal Representation (TIMR)
• Comprehensive Contextual Multimodal Insight (CCMI)
• Cross-Attention (CA)
Meme Text
Meme Image
CLIP
OpenAI ViT-B/32
Textual Feature
Visual Feature
Fusion (MFB)
M$_i$
Scene Graph
LLM
LLM
LLM
R$_{Ei}$
R$_{Ti}$
R$_{Ci}$
Identify the primary emotions conveyed through Ti and EORi of the meme.
Based on the Ti and the EORi of the meme Xi, provide a rationale for whether this meme targets women. Include specific elements that support this claim.
Given the text Ti and the EORi of the meme, provide the broader context C of meme Xi
The primary emotion conveyed through this meme is likely to be anger or disgust, ..., their expressions and body language.
Based on the meme text and the objects of the meme (a woman and a man), it appears that this meme is targeting women.
The broader context of this meme involves societal and cultural debates around gender roles ... controversy about gender roles in modern society.
CA
CA
CA
EMF
TIMR
CCMI
L_{Em}
L_{Ta}
L_{Co}
CL
CL
CL
Softmax Layer
P(Y=y|w,s)
Misogynous
Non-misogynous

Figure 2: Illustration of the proposed M3Hop-CoT model.

2018). In contrast with this, MFB provides an efficient solution by factorizing the bilinear pooling operation. This approach effectively maximizes the association between textual and visual features while mitigating the computational and overfitting concerns associated with traditional bilinear pooling (Yu et al., 2017; Kumari et al., 2023). MFB combines ft_i and fv_i to produce a multimodal representation M_i with dimensions $\mathbb{R}^{o \times 1}$. The MFB module employs two weight matrices, U and V, to project and sum-pool the textual and visual features, respectively. The resulting fusion is expressed in the following equation:

$$M_i = \text{SumPool}(U^T ft_i \circ V^T fv_i, k) \; ; \qquad (2)$$

Here, $\circ$ represents element-wise multiplication, and $\text{SumPool}(x, k)$ refers to a sum-pooling operation over x using a one-dimensional non-overlapping window of size k.

4.3 Entity-Object-Relationships (EoRs) Extraction

Improving the representation of textual and visual components in meme analysis is crucial to bridge the semantic gap between these modalities. To retrieve a better visual representation, we utilize an unbiased scene graph model proposed by Krishna et al. (2016), which leverages Faster RCNN (Ren et al., 2015) and joint contextual feature embedding to extract unbiased *Entity Object Relationship (EOR)* data from the visual modality of each meme (c.f. Figure 8). For a meme image I_i in meme $\mathcal{X}_i$, the scene graph is defined as $T_I \subseteq (E_I \times R_I \times E_I)$, where T_I is the set of visual triples, E_I is the entity set, and R_I is the relation set, with $R_I \subseteq R$. Each entity $e_{I,k} = (e_{t,I,k}, A_{I,k}, b_{I,k}) \in E_I$ consists of the entity type $e_{t,I,k} \in E_t$, where E_t is the set of entity types (Refer to Appendix Figure 7). For meme $\mathcal{X}_i$, the extracted entity-object-relation triplets from its scene graph are denoted as EOR_i. The notation EOR_i represents the top k (in this case, k=5) entity-object relations from the scene graph of image I_i, expressed as $\text{EOR}_i = (\text{EOR}_i^1, \text{EOR}_i^2, \ldots, \text{EOR}_i^k)$. Integrating this visual understanding into our LLM is intended to uncover such hidden cues in images that are crucial for making informed, human-like decisions for detecting misogynous memes.

4.4 Chain-of-Thought Prompting

Our M3Hop-CoT model (refer to Figure 2) employs a Chain-of-Thought (CoT) prompting approach (Wei et al., 2023; Zhou et al., 2023) to facilitate multi-step reasoning during meme analysis. Rather than directly querying the LLM for the final label $\hat{y}$, we aim for the LLM to infer detailed information about the meme's emotional content, its potential targeting of women, and the broader context of its creation and interpretation. The three-hop prompts are constructed as follows:

Step 1. The first prompt queries the LLM about the emotions E conveyed by the meme with the following template:

> R_{Ei} [Identify the primary emotions conveyed through T_i and EOR_i of the meme $\mathcal{X}_i$.]

R_{Ei} is the first prompt context, which infers the emotions-related rationale to provide the hidden cues for $\mathcal{X}_i$. $R_{Ei} = \text{argmax}(E|T_i, EOR_i)$ where R_{Ei} is LLM-generated output text which explicitly mentions primary emotions E.

Step 2. After that, we ask LLM whether the meme is targeted towards women or not with the following template:

> R_{Ti} [Based on the T_i and the EOR_i of the meme $\mathcal{X}_i$, provide a rationale for whether this meme targets women. Include specific elements that support this claim.]

R_{Ti} is the second prompt context, designed to extract a target-enriched rationale revealing cues of misogynous memes. It is defined as $R_{Ti} = \text{argmax}(T|T_i, EOR_i)$, where R_{Ti} is the LLM-generated text that explicitly provides the rationale for whether the meme targets women.

Step 3. Finally, to understand the broader context of a meme, we ask LLM to define the contextual information of the meme with the following template:

> R_{Ci} [Given the text T_i and the Entity-Object Relationships EOR_i of the meme, provide the broader context $\mathcal{C}$ of meme $\mathcal{X}_i$.]

Finally, R_{Ci} is the third prompt context, aimed at uncovering contextual knowledge highlighting social cues associated with memes. It is defined as $R_{Ci} = \text{argmax}(C|T_i, EOR_i)$, where R_{Ci} is the LLM-generated text explicitly outlining the meme's context C.

4.5 Encoding of LLMs Generated Rationale

To leverage the sequential and contextual information within the LLM-generated rationale $R_{r_i} = \{w_{1_i}, w_{2_i}, \ldots, w_{k_i}\}$, where $r \in \{e, t, c\}$ corresponds to emotion-rich, target-aware, and contextually-enriched rationale of meme sample $\mathcal{X}_i$, respectively, with varying word lengths $k \in \{l, m, n\}$, we employ the textual encoder of the pre-trained CLIP model:

$$(R_{E_i}, R_{T_i}, R_{C_i}) = \text{CLIP}(R_{e_i}, R_{t_i}, R_{c_i}) \ ; \qquad (3)$$

4.6 Enhancing CoT Reasoning via Cross-Attention

We use three-layer hierarchical cross-attention to enable the interaction between the representations of rationales $(R_{E_i}, R_{T_i}, R_{C_i})$ before determining the final label $\hat{y}$.

Emotive Multimodal Fusion (EMF): To derive an emotion-enriched multimodal representation of the meme $\mathcal{X}_i$, we calculate the cross-attention $H1_i$ between M_i from Equation 2 and R_{E_i}. Initially, we perform linear transformations to obtain *query* ($Q_{M_i} = M_i W_{M_q}$), *key* ($K_{R_{E_i}} = R_{E_i} W_{E_k}$), and *value* ($V_{M_i} = M_i W_{M_v}$) vectors for both the multimodal representation and the emotion-rich rationale using learned weight matrices $(W_{M_q}, W_{E_k}, W_{M_v})$:

$$H1_i = \text{softmax}\left(\frac{Q_{M_i} K^T_{R_{E_i}}}{\sqrt{d_k}}\right) V_{M_i} \ ; \qquad (4)$$

where d_k is the dimension of the key vector. The final representation $HF1_i$ is obtained by adding $H1_i$ to the original multimodal representation M_i through a residual connection and then applying layer normalization (Ba et al., 2016):

$$HF1_i = \text{LayerNorm}(H1_i + M_i) \ ; \qquad (5)$$

Target Insight Multimodal Representation (TIMR): To integrate the target-aware information of meme sample $\mathcal{X}i$ into the emotion-enriched representation obtained in Equation 5, we compute the cross-attention $H2_i$ between the emotive representation $HF1_i$ and the target-aware rationale R_{T_i}. We perform linear transformations to obtain *query* ($Q_{HF1_i} = HF1_i W_{HF1_q}$), *key* ($K_{R_{T_i}} = R_{T_i} W_{T_k}$), and *value* ($V_{HF1_i} = HF1_i W_{HF1_v}$) vectors using learned weight matrices $(W_{HF1_q}, W_{T_k}, W_{HF1_v})$:

$$H2_i = \text{softmax}\left(\frac{Q_{HF1_i} K^T_{R_{T_i}}}{\sqrt{d_k}}\right) V_{HF1_i} \ ; \qquad (6)$$

where d_k is the dimension of the key vector. The final target-aware multimodal representation $HF2_i$ is obtained by adding $H2_i$ to the emotive representation $HF1_i$ through a residual connection and applying layer normalization:

$$HF2_i = \text{LayerNorm}(H2_i + HF1_i) \ ; \qquad (7)$$

Comprehensive Contextual Multimodal Insight (CCMI): To obtain a comprehensive contextual multimodal representation of meme sample $\mathcal{X}i$, we compute the cross-attention $H3_i$ between the target-aware representation $HF2_i$ and the context-aware rationale R_{C_i}. We perform linear transformations to obtain *query* ($Q_{HF2_i} = HF2_i W_{HF2_q}$), *key* ($K_{R_{C_i}} = R_{C_i} W_{C_k}$), and *value* ($V_{HF2_i} = HF2_i W_{HF2_v}$) vectors using learned weight matrices $(W_{H2_q}, W_{C_k}, W_{H2_v})$:

$$H3_i = \text{softmax}\left(\frac{Q_{HF2_i} K^T_{R_{C_i}}}{\sqrt{d_k}}\right) V_{H2_i} \ ; \qquad (8)$$

where d_k is the dimension of the key vector. The final comprehensive representation $HF3_i$ is obtained by adding $H3_i$ to the target-aware representation $HF2_i$ through a residual connection and applying layer normalization:

$$HF3_i = \text{LayerNorm}(H3_i + HF2_i) \ ; \qquad (9)$$

4.7 Network Training

We use a singular feed-forward neural net (FFN) with softmax activation, which takes the Comprehensive Contextual Multimodal representation $(HF3_i)$ in Equation 9 as input and outputs class for misogynous meme identification, shown in the following Equation 10:

$$\hat{y}_t = P(Y_i | HF3_i, W, b) = \text{softmax}(HF3_i W_i + b_i); \quad (10)$$

The proposed classifier is trained using cross-entropy loss:

$$\mathcal{L}_1 = -\sum [y_t \log \hat{y}_t + (1 - y_t) \log(1 - \hat{y}_t)] \ ; \qquad (11)$$

Reasoning Revising with Supervised Contrastive Learning Loss: In addition to cross-entropy loss, we incorporate supervised contrastive loss (SCL) to enhance the CoT-based learning and provide empirical evidence of its effectiveness in learning cultural diversity-enriched representations for a more robust classifier (Li et al., 2023; Shen et al., 2021). This loss component encourages well-separated representations for the misogynous meme identification task, creating equitable representations and correct predictions. All three multimodal representations that enhance the CoT reasoning,i.e., ($HF1_i$,$HF2_i$,$HF3_i$ in Equation 5, 7, and 9) and multimodal representation M_i, are assumed to capture similar contexts for a given meme $\mathcal{X}_i$. During training, these representations are aligned within the same semantic space, enabling effective utilization through contrastive learning.

$$\begin{aligned}
\mathcal{L}_{Em} &= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{HF1_i}, \boldsymbol{M_i}\right)/\tau\right)}{\sum_{l=1[l\neq i]}^{2N} \exp\left(\text{sim}\left(\boldsymbol{HF1_i}, \boldsymbol{M_l}\right)/\tau\right)}\,; \\
\mathcal{L}_{Ta} &= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{HF2_i}, \boldsymbol{M_i}\right)/\tau\right)}{\sum_{l=1[l\neq i]}^{2N} \exp\left(\text{sim}\left(\boldsymbol{HF2_i}, \boldsymbol{M_l}\right)/\tau\right)}\,; \qquad (12) \\
\mathcal{L}_{Co} &= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{HF3_i}, \boldsymbol{M_i}\right)/\tau\right)}{\sum_{l=1[l\neq i]}^{2N} \exp\left(\text{sim}\left(\boldsymbol{HF3_i}, \boldsymbol{M_l}\right)/\tau\right)}\,;
\end{aligned}$$

where, sim is the cosine-similarity, N is the batch size, and τ is the temperature to scale the logits. Therefore, the overall loss $\mathcal{L}_F$ is a weighted sum of the cross-entropy loss $\mathcal{L}_1$ in Equation 11, and these contrastive losses ($\mathcal{L}_{Em}, \mathcal{L}_{Ta}, \mathcal{L}_{Co}$) in Equation 12. The weights (α, β,γ, and θ) control the relative importance of each loss.

$$\mathcal{L}_F = \alpha \cdot \mathcal{L}_1 + \beta \cdot \mathcal{L}_{Em} + \gamma \cdot \mathcal{L}_{Ta} + \theta \cdot \mathcal{L}_{Co}\,; \qquad (13)$$

5 Results Analysis

In this section, we present the results of our comparative analysis, which examines the baseline models [3], LLM-based models, our proposed model, and their respective variations for misogynous meme identification tasks[4]. We use the macro-F1 (F1) score on both the dev and test sets as the preferred metrics to measure this.

5.1 Model Results and Comparisons

Models Notation: CLIP_MM: This is the CLIP-based classifier. M3Hop-CoT: Proposed scene graph with CoT-based model with emotion, target, and context-aware prompts. M3Hop-CoT^{-E}: Proposed model without emotion-aware prompt, M3Hop-CoT^{-T}: Proposed model without target-aware prompt, M3Hop-CoT^{-C}: Proposed model without context-aware prompt, M3Hop-CoT^{-SG}: This model is trained solely with all the CoT based modules, excluding the scene graph, M3Hop-CoT^{E}: Proposed model with only emotion-aware prompt, M3Hop-CoT^{T}: Proposed model with only target-aware prompt, M3Hop-CoT^{C}: Proposed model with only context-aware prompt and M3Hop-CoT$^{-\mathcal{L}_k}$ where k$\in$ $\{Em, Ta, Co\}$: Proposed model without respective k$_{th}$ loss.

5.1.1 Results on MAMI Dataset

Comparison with Baseline Models: Table 2 presents the performance of various baseline models on the task of misogynous meme identification. Notably, our CLIP-based baseline classifier (CLIP_MM) achieves superior performance with an F1 score of 73.84% on both the dev and test sets, serving as the foundation for our proposed method. We also observed that multimodal baselines give better results than unimodal ones. Furthermore, our proposed model, M3Hop-CoT, surpasses all other baseline models in terms of F1 scores for both the dev and test datasets. It shows the robustness of our proposed model for such a challenging task.

Comparison with LLMs: When extending the CLIP_MM with a prompt-based approach, Mistral LLM surpasses other LLMs by achieving a $\backsim 2\%$ increment on the dev set, whereas $\backsim 4\%$ higher F1-score on the test set, establishing a strong foundation for subsequent CoT-based methods. Moreover, when implementing the CoT-based approach across various LLMs, M3Hop-CoT, which incorporates Mistral LLM, consistently outperforms other CoT-based models. It validated the robustness of the proposed model, which understands the hidden complex cues of any meme by means of their hidden emotions, target, and contextual information (A detailed discussion about the comparison of only prompt-based models with CoT-based models with different LLMs is given in Appendix Section C).

---

[3] Details of the baseline models are given in the Appendix Section B.1

[4] Additional details of experimental setups and hyperparameters explored are provided in the Appendix Section B.5

	Models	Text	Image	MAMI						MIMIC			
				Dev			Test			Test			
				P	**R**	**M-F1**	**P**	**R**	**M-F1**	**P**	**R**	**M-F1**	**W-F1**
Baseline	$L_{FT}(1)$	✓		56.23	56.69	56.47	44.7	47.9	46.2	47.39	46.93	47.16	47.29
	BERT(2)		✓	63.29	71.81	67.28	58.0	50.9	54.2	61.45	61.06	61.25	61.26
	LaBSE (3)	✓		63.59	61.99	63.72	49.4	54.2	51.6	63.59	61.39	62.48	62.66
	VGG19 (4)		✓	64.29	60.79	62.49	47.40	49.40	48.38	44.48	42.35	43.39	43.84
	ViT (5)		✓	69.21	67.36	68.27	54.30	52.40	53.37	49.99	48.91	49.45	49.29
	Early Fusion												
	(1)+(4)		✓	72.60	62.52	67.19	52.5	47.0	49.6	52.39	50.38	51.37	51.2
	(2)+(4)	✓	✓	58.19	64.48	61.18	54.4	51.3	52.7	64.29	62.49	63.38	63.24
	(2)+(5)	✓	✓	70.81	64.09	68.27	53.48	59.29	56.21	69.49	67.97	68.72	68.49
	(3)+(5)	✓	✓	69.09	61.93	65.28	55.93	51.19	53.0	63.85	63.94	63.89	63.91
	Pretrained Model												
	LXMERT	✓	✓	78.94	69.45	73.88	69.01	65.18	65.9	66.03	61.39	63.63	63.21
	MMBT	✓	✓	73.60	69.09	71.27	56.4	49.0	52.4	68.39	65.91	67.13	67.39
	VisualBERT	✓	✓	81.03	77.79	79.38	78.2	61.2	68.7	73.98	70.39	72.15	72.38
	BLIP	✓	✓	70.95	68.28	69.58	62.39	53.39	57.54	74.39	72.39	73.38	73.74
	ALBEF	✓	✓	72.30	70.98	71.62	59.2	53.5	56.1	71.21	69.38	70.28	70.13
	***CLIP_MM**	✓	✓	**85.3**	**83.4**	**84.3**	**75.4**	**69.2**	**72.1**	**76.39**	**74.05**	**75.24**	**75.25**
Prompt-based	**CLIP_MM**												
	+ ChatGPT	✓	✓	85.89	83.99	84.98	80.0	69.3	74.2	76.71	74.59	75.63	75.34
	+ GPT 4	✓	✓	87.11	84.81	85.93	75.5	71.3	72.3	76.47	72.43	74.39	74.12
	+ Llama	✓	✓	83.70	81.29	82.46	77.83	69.40	73.38	78.01	73.97	75.94	75.75
	+ Mistral (Ours)	✓	✓	**88.80**	**84.76**	**86.72**	**81.20**	**72.70**	**76.94**	**78.15**	**75.39**	**76.75**	**76.35**
CoT-based	**CLIP_MM**												
	+ChatGPT	✓	✓	86.20	84.40	85.29	81.0	76.0	77.0	78.69	76.34	77.49	77.41
	+ GPT4	✓	✓	89.52	85.20	87.38	71.9	70.8	71.4	75.16	73.39	74.26	74.21
	+ Llama	✓	✓	91.38	86.28	88.85	77.50	76.40	76.98	77.17	75.10	76.12	76.91
	M3Hop-CoT$_{Mistral}$ **(Proposed)**	✓	✓	**96.39**	**87.59**	**91.75**	**82.38**	**78.29**	**80.28**	**80.29**	**78.98**	**79.63**	**79.41**

Table 2: Results from the proposed model and the various baselines on the MAMI and MIMIC datasets. Here, the bolded values indicate maximum scores. Here, T: Text, I: Image, M-F1: Macro F1, and W-F1: weighted F1-score. * represents the best-performing baseline model. We observe that the performance gains are statistically significant with p-values (<0.0431) using a t-test, which signifies a 95% confidence interval.

Models	Text	Image	Macro F1-score			
			MAMI		MIMIC	
			dev	**test**	**M-F1**	**W-F1**
M3Hop-CoT (Ours)	✓	✓	**91.75**	**80.28**	**79.63**	**79.41**
M3Hop-CoT^{-E}	✓	✓	86.83	76.3	73.74	73.01
M3Hop-CoT^{-T}	✓	✓	86.92	75.1	75.37	74.92
M3Hop-CoT^{-C}	✓	✓	85.92	75.3	73.91	73.24
M3Hop-CoT^{-SG}	✓	✓	84.21	73.9	72.35	72.47
M3Hop-CoT^{E}	✓	✓	82.99	70.2	69.28	70.14
M3Hop-CoT^{T}	✓	✓	84.21	73.2	73.58	73.76
M3Hop-CoT^{C}	✓	✓	84.38	71.2	75.86	75.97
M3Hop-CoT$^{-\mathcal{L}\_Em}$	✓	✓	89.29	76.2	77.94	77.05
M3Hop-CoT$^{-\mathcal{L}\_Ta}$	✓	✓	88.73	77.0	75.62	75.33
M3Hop-CoT$^{-\mathcal{L}\_Co}$	✓	✓	88.28	77.9	77.95	77.01

Table 3: Ablation Study: Role of different modules in our proposed model. We observe that the performance gains are statistically significant with p-values (<0.0553) using a t-test, which signifies a 95% confidence interval.

5.1.2 Results on MIMIC Dataset

To show the robustness of our proposed model in another language, in Table 2, we have shown the results on the MIMIC dataset, which is in Hindi-English code-mixed. Our proposed model follows a behavior similar to the MAMI dataset (outperforming by more than $\backsim 4-5\%$ from CLIP_MM), whereas CoT-based LLM is not only leveraging the language-related dependency but also performing superbly by utilizing the different cultural-based hidden cues of the dataset (A detailed analysis of results on this dataset is provided in the Appendix Section A).

5.1.3 Ablation Study

To assess our proposed architecture, we created several multimodal variants of our proposed model M3Hop-CoT by training it on MAMI and MIMIC datasets, as shown in Table 3, which allows us to evaluate the contribution of each component to the model's overall performance. M3Hop-CoT emerged as the most effective model, achieving a significant increase of 6-7% in F1 scores for both development and test sets. Additionally, incorporating SCL further enhanced M3Hop-CoT's performance, as evidenced by the impact of each loss component. The model's superior performance is attributed to its balanced use of textual and visual modalities, integration of entity-object relationships, and leveraging key factors such as emotion, target, and context-enriched LLM-generated rationales. M3Hop-CoT effectively captures the semantic relationships between objects in the meme, which is crucial for identifying misogynous content.

5.2 Detailed Analysis

5.2.1 Result Analysis with Case Study

Using Appendix Figure 5, we qualitatively analyze our proposed framework through the predictions obtained from the baseline CLIP_MM and our proposed model M3Hop-CoT. For meme sample (a) with the text “I WAS BROUGHT UP TO NEVER HIT A WOMAN. SHE’S NO WOMAN," and an image showing a slap and a woman, “CLIP_MM," classified it as non-misogynous. In contrast, our model M3Hop-CoT correctly classified it as misogynous using a CoT-based rationale from an LLM with multi-hop reasoning. While CLIP_MM slightly preferred text (as depicted by T= 13.85) over visuals (V= 11.27), M3Hop-CoT provided a balanced contribution by considering both text and visuals and context. It is evident in GradCAM, where M3Hop-CoT distinctly highlights both the slap and the woman, unlike CLIP_MM, which fails to concentrate on these critical elements. Similarly, the meme sample (b) conveys the disrespect towards women using domestic violence. The LLM’s generated rationale offers insight into the meme’s intended message. Once again, CLIP_MM struggles to accurately classify the meme, whereas “M3Hop-CoT" correctly identifies it as misogynous. M3Hop-CoT effectively recognizes the sarcastic nature of memes by underlying emotions, target, and context, showcasing their ability to understand the meme’s subtleties. In example (c), the meme, which compares a woman to a pig, is identified as misogynous. The CLIP_MM fails to classify it correctly, focusing only on the words "EX WIFE/FOLLOW/ WEDDING PHOTOS" and missing the image’s subtle cues. In contrast, “M3Hop-CoT" accurately detects its misogynous nature by considering both modalities and integrating contextual knowledge through multimodal reasoning. Enhanced by CoT prompting and EoRs, M3Hop-CoT provides a more comprehensive analysis and outperforms baseline models in recognizing misogynous content (Similar qualitative analysis for the MIMIC dataset is shown in the Appendix Section A.1.)

5.2.2 Result Analysis for Cultural Diversity

In the Appendix Figure 12, we present three illustrative examples from the MAMI dataset, showcasing how M3Hop-CoT leverages cultural knowledge from diverse demographics. The model better recognizes misogyny by incorporating emotional cues, target identification, and context in a CoT framework. Each example in the figure delves into different cultural references. These include historical beliefs surrounding the Church and women’s roles in the 1500s (c.f. example (i)), comparisons between women and witches within Japanese mythology (c.f. example (ii)), and Christian interpretations of the Bible’s teachings (c.f. example (iii)). Notably, CLIP_MM fails to grasp the underlying misogynistic connotations within these examples. Conversely, our proposed model effectively utilizes these cultural references, leading to accurate predictions of misogynistic labels.

5.2.3 Quantitative Analysis with Error Rates

We illustrate the impact of various M3Hop-CoT model variants on test error rates in the Appendix Figure 3. CLIP_MM model exhibits the highest error rate, highlighting the necessity of LLMs for such complex tasks. Models like M3Hop-CoT^{-E}, M3Hop-CoT^{-T}, and M3Hop-CoT^{-C}, lack emotion, target, and context-aware prompts, respectively, have higher error rates than the proposed M3Hop-CoT, indicating the importance of these components. Additionally, M3Hop-CoT^{-SG}, excluding the scene graph module, shows an increased error rate, emphasizing the significance of visual semantics. Models M3Hop-CoT^{E}, M3Hop-CoT^{T}, and M3Hop-CoT^{C}, focusing on individual rationale, demonstrate that a balanced approach is essential for optimal performance. The M3Hop-CoT model achieves the lowest error rates, demonstrating its superior ability to identify such memes.

5.3 Generalibity of the Proposed Model

To demonstrate the adaptability of our proposed architecture, M3Hop-CoT, we assess its performance across three English benchmark datasets: Hateful Memes, Memotion2, and the Harmful dataset (c.f. Table 4). This evaluation validates the generalizability of our architecture, demonstrating its effectiveness not only in misogynous tasks but also in various benchmark datasets and tasks (See the detailed discussion in Appendix Section D).

5.4 Comparison with State-of-the-Art Models

Table 5 presents a detailed comparison between M3Hop-CoT and other state-of-the-art (SOTA) models. In the MAMI task, M3Hop-CoT surpasses existing SOTA. Despite PromptHate achieving high accuracy on the MAMI dataset, it struggles with contextual knowledge, leading to modality-specific biases. Another model, Multimodal-CoT, attempts to leverage multimodal features with LLM but lacks essential psycholinguistic factors that our model incorporates, such as emotions, target awareness, and contextual information (c.f. Figure 13). Our M3Hop-CoT model outperforms, mainly due

Models	Modality		Memotion2	Hateful		Harmful
	T	I	F1↑	F1↑	AUC↑	F1↑
FasterRCNN		✓	48.9	38.81	59.97	65.9
BERT	✓		50.01	58.41	67.92	77.92
ViT		✓	51.17	—	—	67.88
Late-Fusion	✓	✓	51.4	64.40	72.51	78.50
MMBT	✓	✓	52.1	58.29	76.77	80.2
VisualBERTCOCO	✓	✓	50.86	59.28	73.85	86.1
ALBEF	✓	✓	50.8	—	—	87.5
ViLBERT	✓	✓	49.92	52.60	76.32	85.83
VisualBERT	✓	✓	51.06	67.46	74.63	84.57
UNITER	✓	✓	52.7	61.66	60.02	61.66
LXMERT	✓	✓	52.3	69.45	76.15	69.45
$^{\phi}$SOTA	✓	✓	$^{\varphi}$55.17	$^{\gamma}$66.71	73.43	$^{\gamma}$89.0
DisMultiHate	✓	✓	50.57	63.31	75.97	84.57
Momenta	✓	✓	50.9	66.71	73.43	88.3
PromptHate	✓	✓	50.89	71.22	77.07	89.0
CLIP_MM (Full-Train)	✓	✓	48.4	53.22	75.98	82.9
CLIP_MM+GPT4 (Full-Train)	✓	✓	56.39	62.18	77.13	85.64
CLIP_MM+ ChatGPT (Full-Train)	✓	✓	55.74	60.39	76.21	86.29
CLIP_MM+Llama (Full-Train)	✓	✓	56.23	59.63	78.29	88.29
CLIP_MM+Mistral (Full-Train)	✓	✓	57.75	65.58	79.02	88.75
M3Hop-CoT (Zero-Shot)	✓	✓	**53.47**	**78.36**	**79.93**	**85.38**
M3Hop-CoT (Full-Train)	✓	✓	**59.95**	**79.24**	**83.29**	**91.01**

Table 4: Results from the proposed model and the various baselines on the Memotion2, Hateful Memes, and Harmful Memes datasets. Φ: SOTA model on respective datasets. φ by (Ramamoorthy et al., 2022) for Memotion2, γ by (Cao et al., 2022) for Hateful meme, and Harmful Memes. We observe that the performance gains are statistically significant with p-values (<0.05) using a t-test, which signifies a 95% confidence interval.

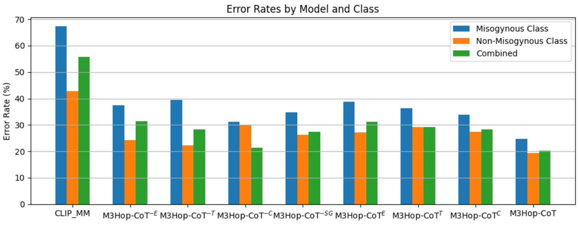


Figure 3: Misclassification rate comparison between proposed model **M3Hop-CoT** and their various variants

to its use of EoRs and the above psycholinguistic factors.

Models	Macro-F1↑	
	dev	test
$^{\Psi}$Zhang and Wang (2022)	83.4	77.6
DisMultiHate (Lee et al., 2021)	67.24	61.89
Momenta (Pramanick et al., 2021b)	72.81	68.29
MMBT (Kiela et al., 2019)	74.8	68.93
PromptHate (Cao et al., 2022)	79.98	73.28
Multimodal-CoT (Zhang et al., 2023)	82.98	72.19
Kumari et al. (2024)	79.59	—
M3Hop-CoT	**91.75**	**80.28**

Table 5: Comparison of our proposed model with the existing SOTA models, Ψ is the SOTA on MAMI Dataset

5.5 Error Analysis

Despite its high performance, our proposed model occasionally misclassifies memes in following scenarios: **(i) Cartoonist image:** In certain scenarios, M3Hop-CoT overlooks the extracted rationale from CoT LLMs and solely concentrates on the image featuring cartoon characters, leading to a misclassification of the meme as "Non-misogynous (c.f. Appendix Figure 11 (a))." **(ii) Reasoning Failure:** M3Hop-CoT sometimes struggle to produce accurate rationales using LLMs due to the implicit nature of memes (c.f. Appendix Figure 11 (b)), such as failing to recognize external references (e.g., the significance of grey sweatpants). **(iii) Overgeneralization of identity terms:** M3Hop-CoT overgeneralize specific "identity terms (e.g., the presence of the word 'SANDWICH')," leading to a misclassify a meme as "Misogynous" based solely on these words while disregarding other information such as images and rationales extracted by LLMs (c.f. Appendix Figure 11 (c)). **(iv) Annotation Error:** In our analysis, we encountered situations where our proposed model accurately predicted the correct label for a given sample. However, due to the problematic annotation issues, misclassification happens (c.f. Appendix Figure 11 (d)). (More detailed error analyses are discussed in Appendix Section E).

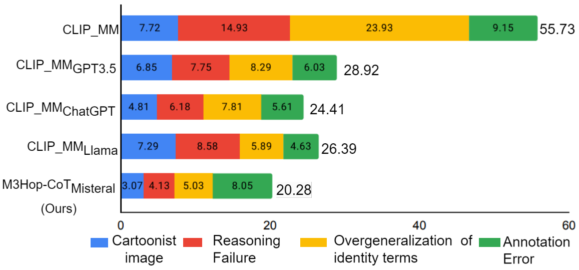


Figure 4: Categorization of error analysis (%) of proposed model **M3Hop-CoT** and other SOTA models

6 Conclusion

In conclusion, our work introduces a novel approach for detecting misogynous content in memes, leveraging the power of LLMs with CoT reasoning. Our proposed model, *M3Hop-CoT*, integrates multimodal information and employs a three-step reasoning process to effectively capture memes' emotional, target-oriented, and contextual nuances. By incorporating scene graphs, we enhance the model's ability to understand the visual aspects of memes. Our results demonstrate that M3Hop-CoT outperforms existing SOTA models, significantly improving F1 scores on both dev and test sets. In the future, we could explore extending our approach to other forms of online content and integrating additional modalities to enhance the model's effectiveness.

Limitations

In Section 5.5, we discussed a few limitations of our proposed model. Despite its strengths, our model encounters difficulties in accurately detecting misogynous memes, especially when the images are cartoonish or when the misogynous references are subtle and require nuanced reasoning. These challenges highlight areas for further refinement and improvement. Understanding these limitations is crucial for advancing our model's capability to identify misogynous content more effectively in future iterations (See a detailed future discussion in the Appendix Section F).

Ethics Statement

Broader Impact: The broader impact of this work is significant in the field of misogynous meme identification. This research promotes a safer and more respectful online environment by developing advanced techniques for detecting misogynous content. Our proposed model, M3Hop-CoT, can help reduce the prevalence of harmful content, fostering a more inclusive and peaceful digital community. Addressing the issue of detecting misogynous memes is essential for promoting equality and fostering peace and justice. We create a more inclusive and fair online environment by developing methods to identify such internet memes. This effort also supports the principle by ensuring that marginalized and vulnerable genders are included in development initiatives. However, it is important to acknowledge the ongoing discussion of automated content moderation and potential biases within such systems. We will explore techniques to ensure fairness, transparency, and accountability in future work in such models (See a detailed future discussion in the Appendix Section F).
Intended Use: This research is intended to advance the detection of misogynous content on social media, aiming to improve the experiences of social media users, content moderators, and the broader online community. By enhancing the ability to identify and moderate such content, we hope to contribute positively to safer online interactions.
Misuse Potential: The dataset utilized in this study includes memes with slur words and offensive images, which are included solely for understanding and analyzing the dataset. It is important to clarify that our use of such content is strictly for research, and we do not intend to harm any individual or group. We emphasize the ethical use of our findings and the importance of handling sensitive content with care.

Acknowledgements

The research reported in this paper is an outcome of the project "**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

References

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2022. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, 13:285–297.

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 581–584, Seattle, United States. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 654–662, Seattle, United States. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Preprint*, arXiv:1607.06450.

Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2023. A knowledge infusion based multitasking system for sarcasm detection in meme. In *Advances in Information Retrieval*, pages 101–117, Cham. Springer Nature Switzerland.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5244–5252, New York, NY, USA. Association for Computing Machinery.

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.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Lei Chen and Hou Wei Chou. 2022. RIT boston at SemEval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 636–641, Seattle, United States. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Mary Franks. 2009. Unwilling avatars: Idealism and discrimination in cyberspace. *Columbia Journal of Gender and Law*, 20:224–238.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA. Association for Computing Machinery.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

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*, volume 33, pages 2611–2624. Curran Associates, Inc.

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. *CoRR*, abs/2005.04790.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*.

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Gitanjali Kumari, Anubhav Sinha, and Asif Ekbal. 2024. Unintended bias detection and mitigation in misogynous memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2719–2733, St. Julian's, Malta. Association for Computational Linguistics.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5138–5147, New York, NY, USA. Association for Computing Machinery.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint*.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *Preprint*, arXiv:2209.14610.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, Suryavardan S, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

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, 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, Ramaneswaran S, Md. Shad Akhtar, and Tanmoy Chakraborty. 2024. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, pages 1–12.

Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5272–5290, Toronto, Canada. Association for Computational Linguistics.

Shivam Sharma, Mohd Khizir Siddiqui, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022b. Domain-aware self-supervised pre-training for label-efficient meme analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 792–805, Online only. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *Preprint*, arXiv:2109.10645.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

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.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. *CoRR*, abs/2002.11949.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1977–1983, Minneapolis, Minnesota. 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. *Preprint*, arXiv:2302.13971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *CoRR*, abs/2207.00747.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6519–6534, Toronto, Canada. Association for Computational Linguistics.

Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. Towards identifying fine-grained depression symptoms from memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8890–8905, Toronto, Canada. Association for Computational Linguistics.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Yong Zhang, Cheng Cheng, Shuai Wang, and Tianqi Xia. 2022. Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling. *Biomedical Signal Processing and Control*, 77:103877.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Jin Zhi, Zhou Mengyuan, Mengfei Yuan, Dou Hu, Xiyang Du, Lianxin Jiang, Yang Mo, and XiaoFeng Shi. 2022. PAIC at SemEval-2022 task 5: Multimodal misogynous detection in MEMES with multitask learning and multi-model fusion. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 555–562, Seattle, United States. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. 2022. DD-TIG at SemEval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, Seattle, United States. Association for Computational Linguistics.

A Detailed Results Analysis on MIMIC Dataset

In Table 2, we have mentioned the results of our proposed model and several baseline models for the MIMIC dataset. Notably, the baseline model CM_CLIP performed better than other baselines, showcasing the efficiency of the pre-trained CLIP model for multimodal data. It is performing better than other baselines, with more than $\backsim 4\%$ increment on the test dataset in terms of macro-F1 and weighted F1-scores. Now, moving towards using LLMs, it replicates results similar to those of the MAMI dataset. Although Llama's performance on the MIMIC dataset is better than the MAMI dataset, Mistral LLM is again providing better context than other LLMs, resulting in an increment of $\backsim 3\%$. For the proposed model

A.1 Qualitative Analysis of the MIMIC Dataset

In Appendix Figure 6, we present a qualitative analysis comparing the performance of our proposed model, M3Hop-CoT, with the baseline model, CM_CLIP, on the MIMIC dataset. Sample (i) depicts a meme intended to degrade women through prejudice. CM_CLIP fails to understand the underlying prejudices. In contrast, M3Hop-CoT, by leveraging its ability to understand emotions, targeted information, and context, correctly identifies the misogynistic nature of this meme. Similarly, samples (ii) and (iii) showcase memes designed to humiliate women by referencing a specific Indian context. M3Hop-CoT demonstrates human-like comprehension of the subtle humiliation conveyed within these memes, leading to accurate predictions of the misogynistic label. These findings highlight the effectiveness of M3Hop-CoT in identifying misogyny compared to the baseline model.

B Experiments

B.1 Baseline Models

To compare the performance of our proposed model with some existing state-of-the-art models, we create several baseline models.

B.2 Unimodal Systems

For the unimodal setting, we implement the following variants of the baseline models:

1. **LSTM with FastText-based Embedding** (L_{FT}): We utilize LSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997) networks combined with FastText embeddings (Joulin et al., 2016) to leverage both sequential processing capabilities and enriched word vector representations.

2. **BERT (Bidirectional Encoder Representations from Transformers):** Next, we leverage the BERT model (Pires et al., 2019) to extract contextually rich feature representations from meme text.
3. **LaBSE (Language-agnostic BERT Sentence Embedding):** We utilized the LaBSE (Feng et al., 2020) model to obtain high-quality language-agnostic text embeddings of the meme text.
4. **VGG-19:** This VGG-19 (Simonyan and Zisserman, 2015) architecture is included to capture visual features from meme images. VGG-19 is highly effective in extracting intricate patterns and textures from visual data, which are crucial for analyzing image-based content.
5. **Visual Transformer (ViT):** The ViT (Dosovitskiy et al., 2020) model applies the principles of transformers for image recognition. This model segments images into patches and processes them sequentially, enabling the capture of global dependencies across the entire image.

After feature extraction from each model, the resulting feature vectors are processed through a softmax function for final prediction.

B.3 Multimodal Systems

Early Fusion: In our early fusion approach, we leverage the strength of combining textual and visual features at an initial processing stage by concatenating them to enhance the model's understanding of misogynous context.

1. **L_{FT}+VGG:** Combines LSTM with FastText-based embedding for text and VGG19 for image features, integrating rich textual embeddings with image features.
2. **BERT+VGG:** This model utilizes BERT for its superior text features and VGG-19 for robust image feature extraction.
3. **BERT+ViT:** This model concatenates BERT's contextual understanding of the text with a visual transformer for image features for the final prediction.
4. **LaBSE+ViT:** In this model, we paired Language-agnostic BERT Sentence Embeddings with a Visual Transformer for processing multilingual text alongside complex image data.

Pre-trained Models: To get a better multimodal representation, we employ different pre-trained models that are specifically designed for handling complex multimodal data:

1. **LXMERT** (Tan and Bansal, 2019): This model is specifically tailored for learning cross-modality representations and has shown exceptional performance on tasks that require joint understanding of text and image content.
2. **VisualBERT** (Zhou et al., 2022): A variant of BERT incorporating visual features into the BERT architecture, enhancing its applicability to scenarios where visual context is crucial.
3. **MMBT** (Supervised Multimodal Bitransformers) (Kiela et al., 2019): MMBT integrates information from heterogeneous sources (text and image) using transformer architectures, making it well-suited for tasks where both modalities are equally important.
4. **BLIP** (Li et al., 2022): We utilize this model to bridge the gap between vision and language tasks by effectively leveraging image-language pre-training.
5. **ALBEF** (Li et al., 2021): The Alignment of Language and Vision using BERT leverages a dual-transformer structure that synchronizes learning between visual and textual representations.

B.4 LLM Based Models.

We used ChatGPT (Ouyang et al., 2022), LLaMA (Touvron et al., 2023), GPT 4 (OpenAI et al., 2024), along with Mistral LLMs for zero-shot simple prompt-based and CoT-based models.

B.5 Experimental Details

All models, including baselines, were implemented using the Huggingface Transformers library[5], with a fixed random seed of 42 for consistency. The details of hyper-parameters are given in the Appendix Table 6. The training was conducted on a single NVIDIA-GTX-1080Ti GPU with 16-bit mixed precision. For the proposed model, hyperparameters α, β, γ, and θ in the overall loss function $\mathcal{L}_F$ (Equation 13) were determined through grid search and set to 0.5, 0.5, 0.3, and 0.4, respectively.

LLM: For our proposed model M3Hop-CoT, we used Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) LLM, which has 7 billion parameters.

Tokenizer: To extract the textual and visual features, we have utilized a pre-trained CLIP

---

[5] https://huggingface.co/docs/transformers/index

Hyper-Parameter	MAMI	MIMIC	Hateful	Memotion2	Harmful
epoch	60	60	60	60	60
batch size	64	64	64	64	64
Learning Rate	3e-5	3e-5	1e-4	3e-5	5e-4
Optimizer	Adam	Adam	Adam	Adam	Adam
Image Size	224	224	224	224	224
Random seed	42	42	42	42	42

Table 6: Details of Hyper-parameters

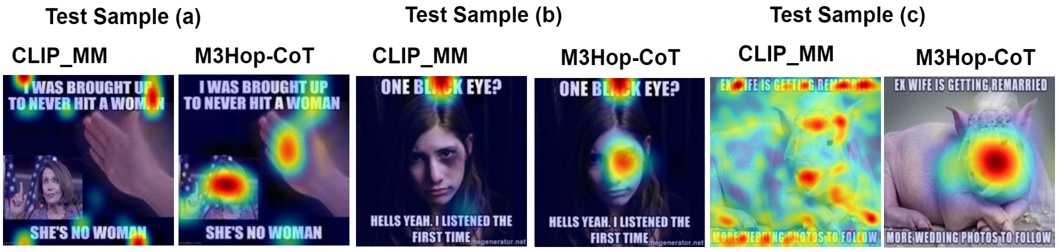
(Contrastive Language-Image Pretraining) model. CLIP is a transformer-based architecture focusing solely on the encoder (no decoder) and utilizes contrastive learning to make textual and visual features semantically similar. Our model leverages the CLIP tokenizer, which employs byte pair encoding (BPE) with a lowercase vocabulary of 49,152 tokens. To facilitate model processing, text sequences are padded with special tokens: "[SOS]" at the beginning and "[EOS]" at the end, signifying the start and end of the sequence, respectively.

For the MAMI dataset, we used the clip (clip-ViT-B-32) model, and for the MIMIC dataset, we utilized multilingual CLIP (mCLIP) (M-CLIP/XLM-Roberta-Large-Vit-L-14).

C Detailed Discussion of Different LLMs-generated Context and the Impact of Prompts

In the results Table 2, we have shown the results of four highly robust LLMs ((a) ChatGPT, (b) GPT 4, (c) Llama and (d) Mistral (Ours)) on the MAMI and MIMIC datasets. We have shown four variations of each LLM used: (i) utilizing simple prompts with only meme text, (ii) utilizing multimodal prompts with meme text with EoRs, (iii) utilizing simple prompts with only meme text by CoT technique, and (iv) utilizing multimodal prompts with meme text with EoRs by CoT technique.

1. **Utilizing only meme text as prompt to LLM:** The first example presents a meme with the text "When I see a woman," accompanied by an image depicting violence against a woman (Refer to Figures 14 (a), 15(a), 16 (a)). In the absence of the image modality by EOR, all three LLMs (ChatGPT, Llama, and Mistral) struggle to identify the misogynistic content within the meme. Llama and ChatGPT respond with uncertainty, indicating a dependence on the broader context. Conversely, Mistral offers a distinct response, stating that the text itself is not inherently misogynistic but could be interpreted as such depending on accompanying visual information. This suggests that Mistral, unlike the other models, attempts to generate context similar to humans even without additional modalities. This behavior presents a promising avenue for further exploration and development. The second example (Figure 21 (a), 22 (a), 23 (a)) presents a meme with the text "Woman in 1500s: Look at the magic trick. The Church." This meme appears to mock and portray a woman's alleged inability to comprehend 15th-century technology, juxtaposed with the Church exhibiting a similar lack of understanding. Here also, unlike other LLMs, Mistral provides a comprehensive explanation, elucidating how this meme perpetuates stereotypes targeting women. Mistral highlights the meme's role in disseminating and reinforcing gender-based and negative stereotypes about the Church.
2. **Utilizing multimodal prompts with meme text and EoRs to LLM:** Now, comparing only text-based LLMs with multimodal LLMs by adding EORs surely adds the extra benefit of the model's understanding of memes while making the prediction. In Figures 14(b), 15(b), and 16(b), the prompt "When I see a woman" is presented alongside an image of man depicting violence against a woman. Including visual elements as EoRs significantly enhances understanding of the meme's context for all LLMs. The Figures 21 (b), 22 (b), 23 (b), demonstrate that EoRs alone are insufficient for grasping the deeper meaning of memes with implicit offensiveness, as seen in the meme text "Woman in 1500s: Look at the magic trick. The Church:". This highlights the limitations of EoRs and underscores the need for a CoT-based approach, as employed by our proposed model, for comprehensive understanding.
3. **Utilizing only meme text as prompt by CoT technique:** Now, we have explored the effectiveness of CoT prompting using only meme text to understand the need for human-like reasoning for identifying misogyny. Figures 17 and 24 showcase the rationale generated by the Mistral model using the CoT technique. This rationale analyzes emotions, targets, and context within the meme. As seen in Figure 17, the model identifies emotions like surprise, anger, and disappointment, along with the targeted nature and context of the meme. This


Test Sample (a)
Test Sample (b)
Test Sample (c)
CLIP_MM
M3Hop-CoT
CLIP_MM
M3Hop-CoT
CLIP_MM
M3Hop-CoT
I WAS BROUGHT UP TO NEVER HIT A WOMAN
SHE'S NO WOMAN
ONE BLACK EYE?
HELLS YEAH. I LISTENED THE FIRST TIME
EX WIFE IS GETTING REMARRIED
MORE WEDDING PHOTOS TO FOLLOW

	Test Sample (a)		Test Sample (b)		Test Sample (c)	
CLIP_MM	I WAS BROUGHT UP TO NEVER HIT A WOMAN. SHE'S NO WOMAN		ONE BLACK EYE? HELLS YEAH. I LISTENED THE FIRST TIME.		EX WIFE IS GETTING REMARRIED. MORE WEDDING PHOTOS TO FOLLOW	
M3Hop-CoT	I WAS BROUGHT UP TO NEVER HIT A WOMAN. SHE'S NO WOMAN		ONE BLACK EYE? HELLS YEAH. I LISTENED THE FIRST TIME.		EX WIFE IS GETTING REMARRIED. MORE WEDDING PHOTOS TO FOLLOW	
Actual Label	Misogynous		Misogynous		Misogynous	
CLIP_MM	Non-Misogynous	T=13.85 V=11.27	Non-Misogynous	T=12.85 V= 12.94	Non-Misogynous	T=14.07 V= 12.04
M3Hop-CoT	Misogynous	T=12.39 V=12.74	Misogynous	T=13.28 V= 13.04	Misogynous	T=11.63 V= 12.34

Legend for Text: 🟥 Misogynous ☐ Neutral 🟩 Non-Misogynous

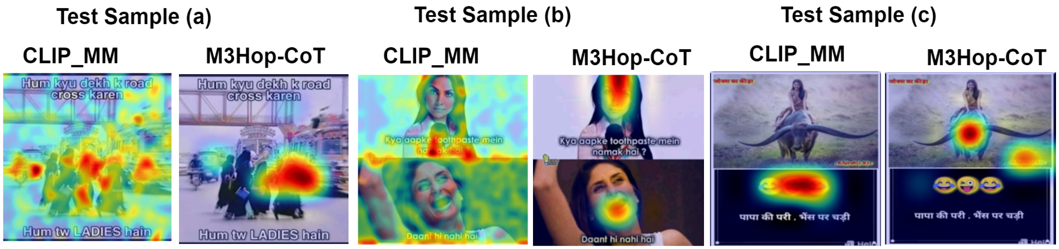
Figure 5: Case studies comparing the attention-maps for the baseline **CLIP_MM** and the proposed model **M3Hop-CoT** using Grad-CAM, LIME (Ribeiro et al., 2016), and Integrated Gradient (Sundararajan et al., 2017) on the **MAMI dataset** test samples. Here, T and V are the normalized textual and visual contribution scores in the final prediction using Integrated Gradient.

analysis helps our proposed model (M3Hop-CoT) to understand the underlying claim of degrading women and make a correct prediction. Similarly, Figure 24 demonstrates how human-like reasoning, achieved through CoT prompting, allows M3Hop-CoT to analyze the three crucial cues (emotions, target, and context) and accurately identify the misogynistic label. However, without a visual element, the LLM fails to generate accurate reasoning about the meme.

4. **Utilizing multimodal prompts with meme text and EoRs by CoT technique:** Finally, we showcase the impact of CoT prompting with multimodal prompts, incorporating both meme text EoRs. Figures 18, 19, 20, and corresponding Figures 25, 26, and 27 present the rationale generated by different LLMs using the CoT technique for the same two examples. The results demonstrate the superiority of Mistral with multimodal prompts. As seen in Figures 18 and 25, Mistral generates highly relatable, human-like rationales for both examples. These rationales consider emotions, targets, context, and cultural nuances within the meme. Compared to Mistral, LLMs like ChatGPT and Llama struggle to produce such comprehensive and culturally rich rationales (Figures 19, 26, 20, 27). This highlights the effectiveness of Mistral in leveraging CoT prompting with multimodal information for superior performance in misogyny detection.

D Result analysis on Hateful meme, Memotion2 and Harmful meme dataset

To evaluate the robustness of our proposed method across various datasets and to understand how common, language-specific taboo elements affect generalization, we conducted a comprehensive generalization study, as highlighted in (Nozza, 2021; Ranasinghe and Zampieri, 2021, 2020). We tested our model on three well-known datasets: the Hateful Memes dataset (Kiela et al., 2020a), the Memotion dataset (Sharma et al., 2020), and the Harmful Memes dataset. Notably, these datasets are predominantly in English and were used to evaluate our model in a zero-shot manner, meaning the model was not directly trained on these specific datasets. These datasets include unique linguistic elements, such as slang and jargon that differ significantly from those found



English Translation	Hum kyun road dekh kar cross kare, hum to ladies hai na?		Is there salt in your toothpaste, I have not even teeth.		Papa's angel, riding the buffalo	
CLIP_MM	Hum kyun road dekh kar cross kare, hum to ladies hai na?		Kya aapke toothpaste me namak hai, daant hi nhi hai.		पापा की परी, भेस पर चढ़ी.	
M3Hop-CoT	Hum kyun road dekh kar cross kare, hum to ladies hai na?		Kya aapke toothpaste me namak hai, daant hi nhi hai.		पापा की परी, भेस पर चढ़ी.	
Actual Label	Misogynous		Misogynous		Misogynous	
CLIP_MM	Non-Misogynous	T=10.92 V=13.28	Non-Misogynous	T=14.34 V= 10.36	Non-Misogynous	T=14.81 V= 12.05
M3Hop-CoT	Misogynous	T=12.34 V=12.47	Misogynous	T=12.76 V= 11.94	Misogynous	T=12.49 V= 12.97

Legend for Text: 🟥 Misogynous ☐ Neutral 🟩 Non-Misogynous

Figure 6: Case studies comparing the attention-maps for the baseline **CLIP_MM** and the proposed model **M3Hop-CoT** using Grad-CAM, LIME (Ribeiro et al., 2016), and Integrated Gradient (Sundararajan et al., 2017) on the **MIMIC dataset** test samples. Here, T and V are the normalized textual and visual contribution scores in the final prediction using Integrated Gradient.

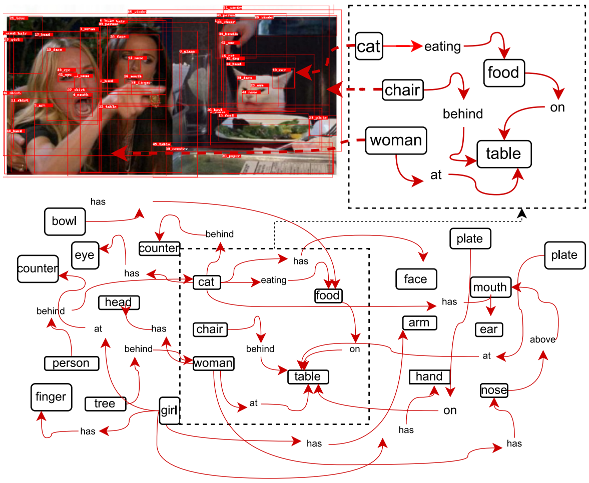


Figure 7: Illustration of scene graph for an image I.

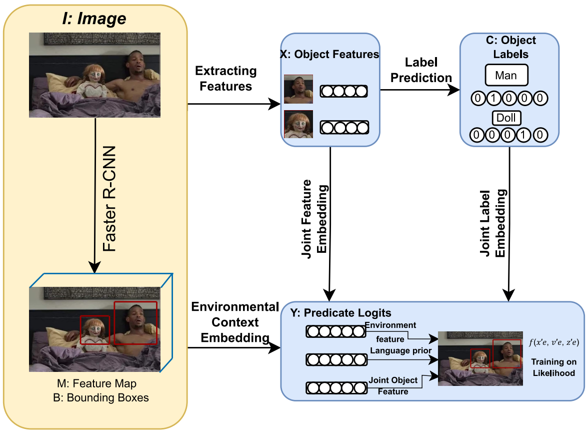


Figure 8: Illustration of the architecture of model used for scene graph.

in misogynous memes, making them challenging out-of-distribution samples for our model.

Despite these challenges, our model demonstrated robust performance across all datasets, underscoring its effectiveness as shown in Table 4 and illustrating its broad applicability. The model's ability to handle linguistic and cultural variances effectively showcases its versatility and potential for widespread use across diverse data sources. The performance metrics from the Hateful Memes dataset, detailed in Table 4, offer valuable insights into how our novel M3Hop-CoT model compares with various baseline and state-of-the-art (SOTA) models.

D.1 Results on Memotion meme dataset

Our proposed model's performance through zero-shot learning on the Memotion dataset gives balanced results. While pre-trained vision and language models deliver significantly good results, the prompt-based model, PromptHate, surpasses all other models, highlighting the efficacy of prompting techniques. However, when evaluating the performance of our proposed model, it shows improvement over other pre-trained models, but the increment is not significant. The performance increment is 5.03% lower compared to the Hateful Meme dataset. This discrepancy can be attributed to the different nature of the Memotion dataset. Unlike the Hateful Memes dataset, which is synthetically generated following a specific template, the Memotion dataset comprises real data collected from social media platforms and features a variety of generalized templates. Considering the real-world nature of the Memotion dataset, an improvement of 5.03% is nevertheless substantial, demonstrating the transferability and robustness of our M3Hop-CoT model. This supports our hypothesis that our model effectively captures critical aspects such as emotion, target, and context of the memes, consistent with the observed trend where social media content frequently targets women. This outcome underscores the prevalence and impact of gender-targeted content in the social media discourse.

Now, when we trained our model on the entire dataset using simple prompts and CoT prompts with LLMs, we obtained better results than the PromptHate, Momenta, and even the SOTA model ($\backsim$ 6% increment). It shows the efficiency of LLMs' understanding of meme's hidden emotions, targeted knowledge, and contextual information, which helped the model outperform baselines.

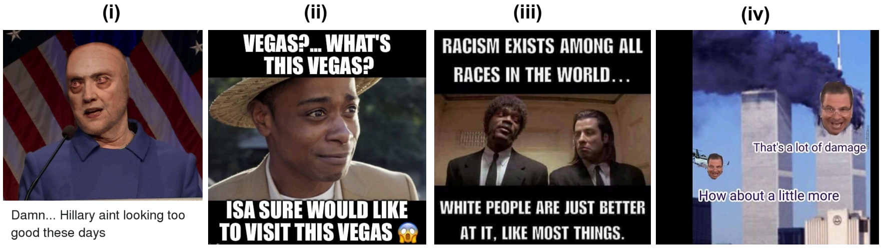
To illustrate the effectiveness of our proposed model, M3Hop-CoT, over baseline and SOTA models, we present a few examples from the memotion dataset in Figure 9. Each example highlights the crucial role emotions, contextual information, and targeted knowledge play in accurately identifying meme labels. For instance, sample (i) shows an offensive image degrading a political leader. While baseline and SOTA models fail to capture these cues, M3Hop-CoT leverages its LLM strength to provide rationales based on three key elements: emotions, context, and targeted knowledge. Similarly, in samples (ii) and (iii), where indirect racism is subtly conveyed within the memes, M3Hop-CoT accurately identifies the offensive nature of the memes. These findings demonstrate M3Hop-CoT's enhanced ability to understand the nuances of complex memes compared to existing models.

D.2 Results on Hateful meme dataset

The Hateful Memes dataset is specifically designed for a hateful meme challenge by synthetically generating memes that alter keywords and images within a template context. Despite being synthetically generated, the Hateful Memes dataset is a task-specific dataset, similar to the MAMI dataset. It encompasses a broad spectrum of what constitutes hate, which also explicitly includes samples that are offensive towards women. This allows for targeted analysis of hate speech and misogyny within memes, reflecting the complexity of the issues being addressed.

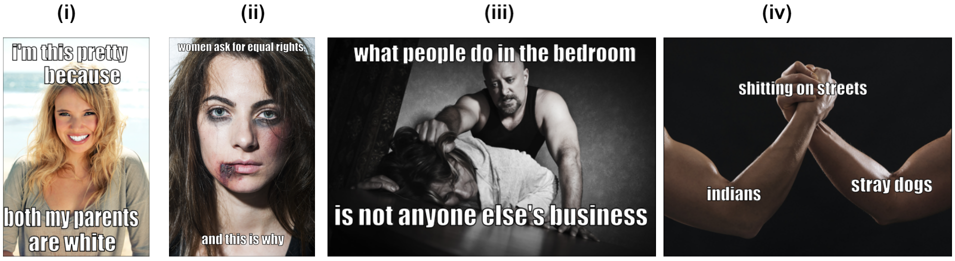
Consequently, when evaluating the Hateful Memes dataset using our M3Hop-CoT model, we achieved remarkably consistent results. Our analysis showed that pre-trained models like VisualBERT, CLIP, and ALBEF also performed well. Among prompt-based models, PromptHate exhibited high accuracy on this dataset. However, a deeper analysis highlighted the significance of our M3Hop-CoT model, which extends beyond mere prompting techniques. M3Hop-CoT enhances its capability by integrating cultural diversity through hierarchical prompts and effectively utilizing Entity-Object Relationships (EoRs), providing a more nuanced understanding of the memes.

Now, training M3Hop-CoT on the complete dataset yielded superior performance compared to all baseline and state-of-the-art (SOTA) models, as shown in Table 4. For a qualitative comparison, in Figure 10, we have shown four meme samples with the actual label of "Offensive." where offensiveness in all the memes are implicit in nature. Compared to the SOTA models like Momenta and PromptHate, M3Hop-CoT excels at identifying implicit offensiveness. While these models leverage image entities and captions respectively, still they fail to grasp the underlying meaning. Notably, LLMs (CLIP_MM+Mistral) are helping in recognizing subtle cues within samples (ii) and (iv). However, LLMs alone struggle with more complex


(i)
(ii)
(iii)
(iv)
Damn... Hillary aint looking too good these days
VEGAS?... WHAT'S THIS VEGAS?
ISA SURE WOULD LIKE TO VISIT THIS VEGAS
RACISM EXISTS AMONG ALL RACES IN THE WORLD...
WHITE PEOPLE ARE JUST BETTER AT IT, LIKE MOST THINGS.
That's a lot of damage
How about a little more

Actual Label	very_offensive	very_offensive	hateful_offensive	hateful_offensive
CM_CLIP	not_offensive	not_offensive	not_offensive	not_offensive
Momenta	Slight Offensive	not_offensive	Slight Offensive	not_offensive
PromptHate	Slight_offensive	Slight Offensive	Slight_offensive	not_offensive
CLIP_MM+Llama	Slight Offensive	not_offensive	Slight Offensive	not_offensive
CLIP_MM+Mistral	Slight Offensive	Slight_offensive	very_offensive	not_offensive
M3Hop-CoT+Misteral	very_offensive	very_offensive	hateful_offensive	hateful_offensive

Figure 9: Predictions from different models on a few test samples from Memotion Dataset.


(i)
(ii)
(iii)
(iv)
i'm this pretty because
both my parents are white
women ask for equal rights
and this is why
what people do in the bedroom
is not anyone else's business
shitting on streets
indians
stray dogs

Actual Label	Offensive	Offensive	Offensive	Offensive
CM_CLIP	Non-Offensive	Non-Offensive	Non-Offensive	Non-Offensive
Momenta	Non-Offensive	Non-Offensive	Non-Offensive	Offensive
PromptHate	Non-Offensive	Non-Offensive	Offensive	Offensive
CLIP_MM+Llama	Non-Offensive	Non-Offensive	Non-Offensive	Offensive
CLIP_MM+Mistral	Non-Offensive	Offensive	Non-Offensive	Offensive
M3Hop-CoT+Misteral	Offensive	Offensive	Offensive	Offensive

Figure 10: Predictions from different models on a few test samples from hateful meme dataset.

patterns of offensiveness, as seen in samples (i) and (iii). In such cases, M3Hop-CoT's with CoT-based prompting approach, mimicking human reasoning, empowers it to predict the offensive label accurately.

D.3 Results on Harmful meme dataset

When evaluating our M3Hop-CoT model using the Harmful Memes dataset, we observed a greater difference in the results compared to the MAMI and other datasets. The Harmful Memes predominantly focuses on the domain of U.S. politics and COVID-19, and both the textual and visual modality of these memes differ significantly from those in the MAMI dataset. The image in the Harmful Memes dataset frequently includes scenes of strikes, fires, group gatherings, and riots. Consequently, the emotional tone, target, and contextual background of these memes diverge significantly from the MAMI dataset, which primarily addresses issues related to targeting and hatred against women. This variation in content underscores the unique challenges posed by the Harmful Memes dataset, affecting the model's ability to generalize across different themes and contexts effectively. Although another prompt-based model, such as PromptHate, is delivering good performance on this dataset, our model, when applied in a zero-shot manner, does not outperform the state-of-the-art (SOTA) models. This highlights areas for further refinement and adaptation of our approach to enhance its performance under zero-shot conditions and across diverse datasets.

Dataset	Split	Label	#Memes
MAMI	Train	Misogynous	5,000
		Non-Misogynous	5,000
	Test	Misogynous	500
		Non-Misogynous	500
MIMIC	Train	Misogynous	2,012
		Non-Misogynous	2,032
	Test	Misogynous	503
		Non-Misogynous	507
Memotion2	Train	Offensive	1,933
		Non-Offensive	5,567
	Test	Offensive	557
		Non-Offensive	943
Hateful meme	Train	Offensive	3,050
		Non-Offensive	5,450
	Test	Offensive	500
		Non-Offensive	500
Harmful meme	Train	Harmful	1,064
		Non-harmful	1,949
	Test	Harmful	124
		Non-Harmful	230

Table 7: Class-wise (MAMI, Memotion2, Hateful meme, and HarmMeme dataset) distribution in Train Set and Test Set

E A note on Error Analysis

Among various types of errors, we mostly categorized these errors as (i) Cartoonist image, (ii) Reasoning Failure, (iii) Overgeneralization of identity terms, and (iv) Annotation Error. As depicted in Appendix Figure 11 (a), when the scene graph misidentifies objects and their relationships in the meme image, the LLM struggles to accurately correlate EoRs with the meme's text and context, leading to the generation of irrelevant rationales. In this instance, the scene graph mistakenly identifies the scrub sponge for utensil cleaning as a cake, resulting in altering the contextual interpretation. Despite the meme's intention to insult women, the generated rationale paradoxically praises the woman for her achievements. These errors are particularly prevalent in memes featuring cartoonist illustrations or images with unclear visibility.

Beyond scene graph errors, Appendix Figure 11 (b) showcases another limitation: LLMs can struggle with complex contextual reasoning. Even with accurate EoRs generated by the scene graph, in some scenarios, the LLM fails to understand the overall meaning of the meme. In this example, the meme satirizes the stereotype used to degrade women. However, the LLM misunderstands the context, generating a rationale that targets men for cheating on women. Furthermore, the term "baby" is used figuratively in the meme, but the LLM interprets it literally, resulting in a misguided rationale. This highlights the ongoing challenge of LLM comprehension when dealing with wordplay, sarcasm, and other forms of nuanced communication within memes.

The third category of error observed is the overgeneralization of certain keywords identified as identity terms. In Appendix Figure 11 (c), the LLM misinterprets the context due to this limitation. The presence of the word "sandwich" triggers a bias within the classifier, leading the model to predict the meme as misogynous wrongly. Despite the meme's offensive nature, it does not aim to degrade women; instead, it advocates violence in general. However, the LLM misunderstood the context based on the keyword and generates, *"The use of the word 'sandwich' is also significant, as it is often used to describe women in a derogatory way."* showcasing the overgeneralization of this term without contextual consideration.

The last error category we address involves discrepancies arising from subjectivity inherent in the annotation process. Despite the LLM generating accurate rationales aligned with the meme's context, misclassification occurs when the predicted label fails to align with the actual label (Appendix Figure 11 (d)). Identifying and rectifying such subjectivity is a complex and ongoing area of research. Prior studies on annotation highlight the challenge of effectively mitigating bias and subjectivity despite the implementation of annotation schemes (Davidson et al., 2019). This underscores the need for further exploration and refinement of annotation methodologies to enhance the reliability and objectivity of classification tasks in natural language processing.

E.1 Further Categorization of Errors

In Appendix Figure 4, we performed a comprehensive error analysis to assess the performance variations of the proposed models across the above-

	Emotions	Target	Context
HEY QUEEN YOU DROPPED THIS (a) Error due to cartoonist image	The primary emotions conveyed through the text of this meme are excitement and enthusiasm. The use of the word "queen" suggests a sense of admiration and respect for the person being addressed, while the phrase "you dropped this" implies that the person has done something impressive or noteworthy. .The person with cake suggests a sense of progress and growth, while the dog itself is a symbol of loyalty and companionship. Together, these elements create a sense of warmth and positivity.	The use of the word "queen" is often associated with women and implies that the meme is directed towards them. Additionally, the visual elements of the meme, such as the person with cake, suggest that the meme is related to a celebration. Overall, the combination of the meme text and visual elements suggest that this meme is encouraging woman to achieve their goals and build a successful career.	The primary emotions conveyed through the text of this meme are excitement and enthusiasm. The use of the word "queen" suggests a sense of admiration and respect for the person being addressed, while the phrase "you dropped this" implies that the person has done something impressive or noteworthy. . the person with cake suggests a sense of progress and growth, while the dog itself is a symbol of loyalty and companionship. Together, these elements create a sense of warmth and positivity.
(b) Error due to wrong reasoning	The primary emotions conveyed through the text of this meme are regret and apology. The use of the word "sorry" suggests that the person is expressing remorse for their actions. The phrase "but he had gray sweatpants" may suggest that the person is trying to justify their actions or make excuses for them. The visual elements of the meme also convey emotions. The tree behind the man suggests a sense of nature and calmness, which may be in contrast to the emotions conveyed by the text. The lamp behind the man may suggest hope or a desire for redemption.	The meme suggests that the meme is targeting women. The use of the word "baby" implies that the person cheating is a man, and the phrase "had gray sweatpants" suggests that he is a lazy and unmotivated person. These characteristics are often associated with men who are not taking their relationships seriously, which could be seen as a negative trait by some women. Furthermore, the visual elements of the meme, such as the tree behind the man and the lamp behind the man, could also be interpreted as symbolizing a lack of commitment or effort in the relationship.	The meme is a play on words that refers to a baby who cheated on a test. The phrase "sorry for cheating" is a common apology, but in this case, it is used in a humorous way to acknowledge that the baby cheated. The phrase "but he had gray sweatpants" is a reference to the fact that the baby was wearing gray sweatpants, which is a common uniform for students taking tests. The image with a tree behind the man and a lamp behind the man is not directly related to the meme, but it could be interpreted as a reference to the idea of knowledge being illuminated by a lamp, and the tree representing growth and knowledge.
NO.. YOU MAKE ME A SANDWICH (c) Error due to overgeneralization of a few terms	The primary emotions conveyed through the text of this meme are frustration, disappointment, and anger. The meme is mocking the idea that women have been fighting for equal rights for centuries and that they still have to deal with things like being asked to make a sandwich. The visual elements of the woman with arm, the sign with a leaf, and the red font all contribute to the overall tone of the meme, which is one of frustration and disappointment. The red font of the text adds a sense of urgency and anger, emphasizing the frustration and disappointment that the speaker feels.	The meme text is a play on words that suggests that the meme is not motivational or inspiring, but rather demotivational. The use of the word "sandwich" is also significant, as it is often used to describe women in a derogatory way. The woman with an arm and the sign with a leaf are also important elements in this meme. The woman with an arm suggests a sense of power and control, which could be seen as threatening to women's rights. The sign with a leaf, on the other hand, could be seen as a nod to the environmental movement, which has historically been associated with feminist issues.	The meme is a play on words that refers to the idea of women being treated as objects or commodities, often in a demeaning or disrespectful manner. The image with the bird having an arm and the sign with a leaf is meant to represent the idea of women being seen as inferior or less valuable than men, and being used as a means to an end. The meme is meant to be a commentary on the way that women are often treated in society, and a call to action for people to be more respectful and equitable in their interactions with women.
(d) Error due to problematic annotations	The primary emotions conveyed through the text of this meme are lust and desire. The use of the word "delicious" suggests that the speaker is attracted to the woman's body and finds it appealing. The exclamation marks at the end of the sentence also convey a sense of excitement. The visual elements of the meme also contribute to the emotional tone. The woman's finger pointing to her own buttocks suggests that she is proud of her body and is comfortable with it.	The meme is written in a male-dominated language and is often used to objectify women. The use of the word "fat" is particularly problematic as it is often used to shame and degrade women, particularly those who are overweight or obese. The use of the word "delicious" is also problematic as it implies that the woman's body is something to be consumed or eaten. The visual elements of the meme, such as the woman with her finger, do not provide any additional context. However, the fact that the meme text is written in a male-dominated language and uses problematic language to objectify women suggests that the meme is targeting women.	The meme is a play on words that uses the phrase "dat" (short for "that") to express objetification for a woman's body. The use of the word "fat" is often considered offensive or derogatory and sexual way to describe the woman's body. The phrase "looks so delicious" is also a play on words, suggesting that the woman's body is so appealing that it is like a delicious food. The image of the woman with her finger is not directly related to the meme, but it may contribute to the overall meaning by emphasizing the idea of the woman being objectifying as a food product.

Figure 11: Error analysis of wrong predictions done by our proposed model M3Hop-CoT

mentioned error categories. CLIP_MM exhibited limitations in reasoning and comprehending identity terms, demonstrated by a high rate of overgeneralization errors (23.93%). This suggests potential deficiencies in understanding the diverse identities within the meme text. While CLIP_MM+GPT4 improved reasoning and identity comprehension, it still struggled with annotation errors, pointing towards potential data labeling issues. Conversely, CLIP_MM+ChatGPT achieved enhancements across all the metrics, indicating superior overall performance and improved contextual understanding. CLIP_MM+Llama showed relatively lower overgeneralization and annotation errors, but reasoning failures still exist. Finally, our M3Hop-CoT model achieved the lowest overall error rate, demonstrating significant advancements in reasoning and identity term comprehension. However, annotation errors remain an area for further refinement. These findings highlight the importance of continuous improvement to mitigate errors and enhance the capabilities of these models for tackling complex meme tasks.

F Future Works

While our current zero-shot prompting approach effectively encourages the LLM to generate rationales for misogyny detection, future work could explore fine-tuning the LLM specifically for misogyny detection within memes. Fine-tuning can enhance contextual reasoning by understanding the dataset's pattern, potentially leading to more

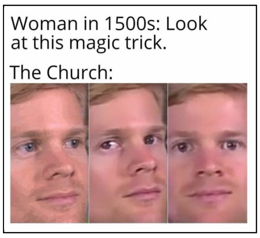
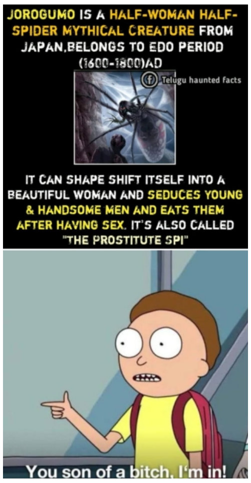
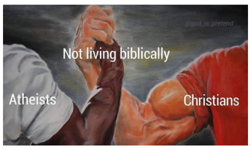
	Emotion	Target	Context
(i)  	Woman in 1500s: Look at this magic trick. The Church: Additionally, analyze the emotions associated with the visual elements man has nose, man has mouth, man has eyes, and man has hair. The primary emotions conveyed through the text of this meme are surprise and curiosity. The woman in the 1500s is looking at something with wonder and amazement, indicating that she is intrigued by the magic trick. The Church's addition of "additionally" suggests that there is more to the trick than just what is being shown, further adding to the sense of mystery and intrigue. The visual elements of the meme also contribute to the emotions conveyed. The man's nose, mouth, eyes, and hair all suggest a sense of realism and authenticity, making the trick seem more impressive and believable.	This meme appears to be targeting women because it uses the image of a woman from the 1500s and the phrase "Look at this magic trick" to make a commentary on the power dynamics between men and women. The visual elements of the meme, such as the man's nose and mouth, are used to emphasize the idea that men have more power and control in society. The use of the phrase "the Church" also suggests that religious institutions and societal norms are contributing to these power imbalances. Overall, the meme is using humor and satire to highlight the ways in which women have been historically marginalized and oppressed.	The meme "Woman in 1500s: Look at this magic trick. The Church: and image with these visual elements" is a play on words that uses the phrase "magic trick" to describe a seemingly impossible or miraculous event. In this case, the "magic trick" is the image of a woman in the 1500s, which is being presented as a feat of modern technology. The use of the phrase "the Church" in the meme adds an additional layer of meaning. The Church is often associated with traditional values and beliefs, and the use of this phrase suggests that the image of the woman is being presented as a challenge to these traditional values. The visual elements of the meme, such as the man with a nose and mouth, are also significant. These elements suggest that the image is being presented as a representation of reality, rather than a fantasy or illusion.
(ii)  JOROGUMO IS A HALF-WOMAN HALF-SPIDER MYTHICAL CREATURE FROM JAPAN.BELONGS TO EDO PERIOD (1600-1800)AD Telugu haunted facts IT CAN SHAPE SHIFT ITSELF INTO A BEAUTIFUL WOMAN AND SEDUCES YOUNG & HANDSOME MEN AND EATS THEM AFTER HAVING SEX. IT'S ALSO CALLED "THE PROSTITUTE SPI" You son of a bitch, I'm in!	The primary emotions conveyed through the text of this meme are confusion and curiosity. The use of the word "half" in describing the Jorogumo creature is likely to create confusion, as it is not a common term in English. Additionally, the use of the word "mythical" may also create curiosity, as people may be unfamiliar with the Jorogumo creature and its significance in Japanese culture. The visual elements of the meme also contribute to the emotions conveyed. The room behind the woman and the girl suggest a sense of mystery and intrigue, as if there is something important or valuable hidden within. The girl in the room may be seen as vulnerable or endangered, as she is in a room that is not fully visible to the viewer. Overall, the combination of the text and visual elements creates a sense of curiosity and intrigue.	The meme text HALF- JOROGUMO IS A HALF-WOMAN SPIDER MYTHICAL CREATURE FROM JAPAN suggests that the meme is targeting women. The visual elements of the meme, such as the room behind the woman, room behind the girl, and room behind the child, further support this conclusion. The room behind the woman is depicted as dark and ominous, which could be interpreted as a representation of the mysterious and dangerous nature of the Jorogumo creature. This could be seen as a metaphor for the challenges and dangers that women face in society. The room behind the girl and child is also depicted as dark and ominous, which could be interpreted as a representation of the potential dangers that children and young women face in a world where they are often objectified and exploited. Overall, the meme text and visual elements suggest that the meme is targeting women and their experiences with gender-based violence and oppression.	The meme "Half-Jorogumo is a Half-Woman Spider Mythical Creature from Japan" is a reference to a popular Japanese mythological creature known as the Jorogumo. The Jorogumo is a spider-like creature that is said to possess magical powers and can transform into a human woman. The meme features an image of a woman with spider-like legs and a room behind her, suggesting that she is in a spider's lair. The use of the term "half-woman" in the meme is significant because it refers to the idea that the Jorogumo is not a fully human creature, but rather a hybrid of human and spider. This idea is central to the mythology surrounding the Jorogumo and is often used to explain her powers and abilities. The use of the term "spider" in the meme is also important because it reinforces the idea that the Jorogumo is a creature that is both beautiful and dangerous. Spiders are often associated with fear and danger, and the use of this imagery in the meme helps
(iii)  	The primary emotions conveyed through the text of this meme are anger, frustration, and disappointment. The text implies that the person is unhappy with the behavior of atheists who do not live according to the Bible, and that they are also unhappy with the behavior of Christians who pretend to be religious but do not actually follow the teachings of the Bible. The visual elements of the meme also convey emotions. The person's head is tilted downward, which could suggest sadness or disappointment. The person's arm is folded across their chest, which could suggest defensiveness or a desire for protection. The person's body is slumped, which could suggest fatigue or exhaustion. The person's face is expressionless, which could suggest detachment or indifference. The person's hair	The meme text "Atheists Not living biblically Christians" suggests that the meme is targeting people who do not live according to the Bible. However, the visual elements of the meme do not provide any specific information about the gender of the people being targeted. Therefore, it is difficult to determine whether or not this meme is targeting women specifically. The person with the head and arm in the meme could represent anyone, regardless of gender. The use of the word "atheists" and "Christians" in the meme text also does not provide any indication of the gender of the people being targeted.	The meme is a commentary on the behavior of some individuals who identify as Christians but do not live up to the teachings of the Bible. The image accompanying the meme features a person with a head and an arm, which could be interpreted as a reference to the idea that people who do not live biblically are not truly following the teachings of God. It also suggests that true faith requires a deep commitment to the teachings of the Bible and a willingness to live in accordance with those teachings, even when it may be difficult or unpopular.

Figure 12: Analysis of rationale generated by the LLMs for the cultural diversity

context-rich rationales and improved misogyny detection. Additionally, it can be particularly beneficial for low-resource domains like misogynous meme identification, offering the potential for superior performance compared to a general-purpose LLM. However, the increased computational cost associated with fine-tuning requires careful consideration, especially when dealing with large datasets or computationally expensive LLM architectures. A future evaluation comparing zero-shot prompting and fine-tuning within our LLM-based model will be crucial for determining the optimal approach for achieving both accurate and efficient misogyny detection in memes.

Another dimension of this work could be related to the scene graph. In the future, we can aim to improve scene graph analysis to mitigate object and relationship recognition errors within memes. This can done by exploring the creation of dynamic scene graphs that adapt in real time to the evolving themes and symbols within memes. This can work better for handling cartoon illustrations and low-visibility images.

Another critical area for future work lies in enhancing the ability of LLMs to perform complex contextual reasoning within the domain of memes. As discussed in the error analysis (c.f. Section 5.5), misogyny in memes often relies on subtle cues, wordplay, sarcasm, and other forms of communication. Current LLMs may struggle to grasp these subtleties, potentially leading to misinterpretations and inaccurate rationale generation. LLMs could benefit from being equipped with pragmatic reasoning techniques that enable them to consider the meme's context, including the speaker's intent, cultural references, and social norms. This would allow the LLM to move beyond the literal meaning of words and understand the underlying message.

G Frequently Asked Questions (FAQs)

Que 1: Why did we choose an emotion, target-aware information, and context as the key factors to generate rationale with LLMs.

Response: Extensive previous research demonstrates the critical role of meme emotions in identifying potential toxicity in memes (Chauhan et al., 2020; Akhtar et al., 2022; Ramamoorthy et al., 2022; Bandyopadhyay et al., 2023; Sharma et al., 2024). These studies also indicate that a major limitation of existing meme identification classifiers is their insufficient contextual understanding (Kumari et al., 2021). Our approach leverages Large Language Models (LLMs) to bridge this gap by integrating comprehensive contextual analysis. Furthermore, targeted information is essential for identifying harmful content, as emphasized by Sharma et al. (2022a). Previous methods primarily relied on supervised learning, which requires extensive data annotation, thereby increasing costs and potential for error. In contrast, our methodology utilizes the capabilities of LLMs to process psycholinguistic features in a cost-effective and error-minimizing manner, thereby enhancing the rationality and effectiveness of meme analysis.

Question 2: Why are the results of the MAMI dataset presented for both the development and test sets?

Response: The MAMI dataset, part of SemEval-2022 Task 5 on Multimedia Automatic Misogyny Identification, aims to explore detecting misogynous memes. This dataset's authors have provided development (dev) and test sets to enable comprehensive evaluation. Presenting results on both sets allows us to assess the model's generalizability and robustness across different subsets of data. This dual-set evaluation strategy ensures that our findings are significant and that the model's performance is robustly demonstrated under varied conditions.

Que 3: Which model serves as the baseline for implementing CoT reasoning?

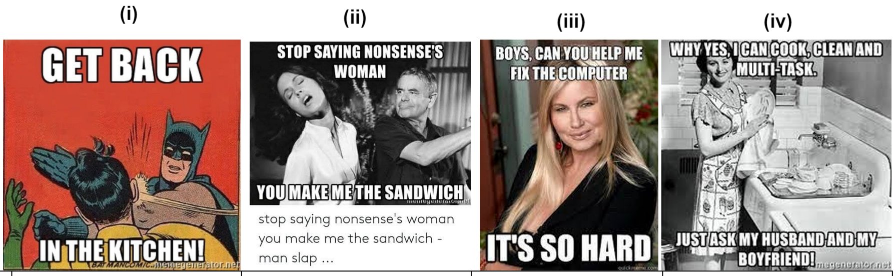
Response: The baseline model for implementing CoT reasoning is our CLIP-based model, *CLIP_MM*. This model integrates the CLIP framework's textual and visual encoders to extract respective features from memes. Feature fusion is achieved using the Multimodal Factorized Bilinear (MFB) pooling technique, with a softmax classification layer with two neurons for label prediction. *CLIP_MM* is optimized using cross-entropy loss and has demonstrated superior performance across all other pre-trained visual-language models for Misogynous meme identification, as evidenced in Table 2. Its effectiveness establishes it as the foundational model for subsequent enhancements with LLM-based techniques.

Que 4: We have written in the results analysis part: "We also observed that multimodal baselines give better results than unimodal ones." Isn't it obvious in a scenario like meme analysis?

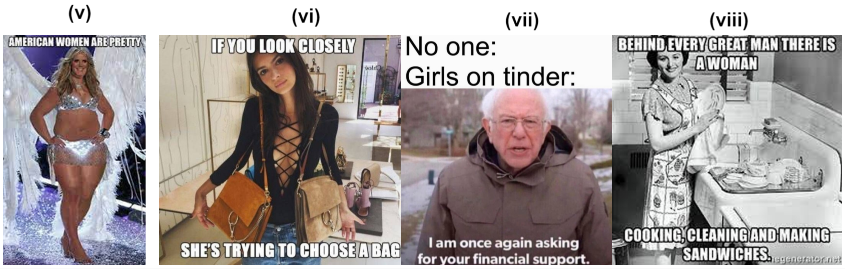
Response: While it might seem intuitive that multimodal approaches would outperform unimodal ones in meme analysis, this is not universally true. Thomason et al. (2019) has shown instances where unimodal inputs surpass multimodal ones, often due to noise and interference in multimodal signals, which can obscure rather than clarify the context. The intention behind highlighting this observation in our study was not to restate the obvious but to provide empirical evidence supporting the efficacy of multimodal systems, specifically in meme analysis tasks within our dataset. This empirical validation emphasizes the practicality and effectiveness of using multimodal techniques for handling memes, strengthening our research's findings. But, yes, we agree that multimodal systems should be better than unimodal systems.

Que 5: Why were uniform metrics not employed across all datasets? This approach could potentially enhance the uniformity of the paper.

Response: The four datasets utilized in our evaluation are publicly available and have been widely adopted in SemEval or other competitions hosted by respective organizations, with the exception of the harmful memes dataset. The authors of the original dataset papers employed specific metrics that have since become standard for these datasets. We opted to use the same


(i)
GET BACK
IN THE KITCHEN!
(ii)
STOP SAYING NONSENSE'S WOMAN
YOU MAKE ME THE SANDWICH
stop saying nonsense's woman you make me the sandwich - man slap ...
(iii)
BOYS, CAN YOU HELP ME FIX THE COMPUTER
IT'S SO HARD
(iv)
WHY YES, I CAN COOK, CLEAN AND MULTI-TASK.
JUST ASK MY HUSBAND AND MY BOYFRIEND!

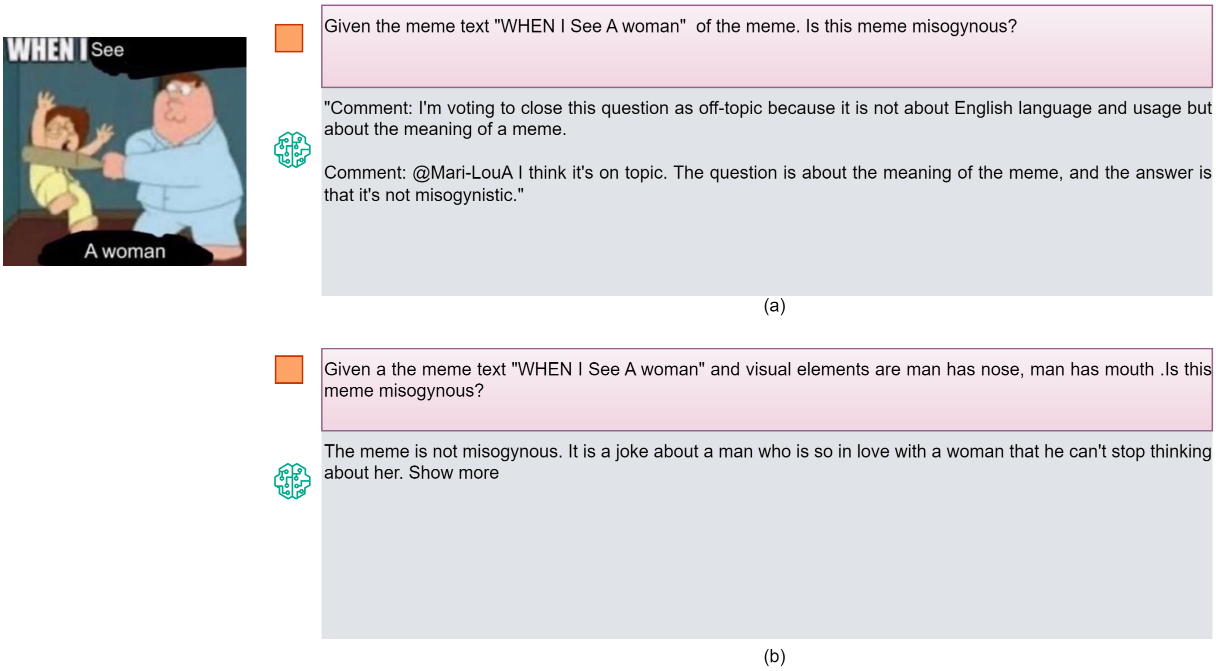
	(i)	(ii)	(iii)	(iv)
Actual Label	Misogynous	Misogynous	Misogynous	Misogynous
CM_CLIP	Non-Misogynous	Non-Misogynous	Non-Misogynous	Non-Misogynous
Momenta	Non-Misogynous	Misogynous	Misogynous	Misogynous
PromptHate	Misogynous	Misogynous	Non-Misogynous	Non-Misogynous
CLIP_MM+ GPT3.5	Misogynous	Non-Misogynous	Non-Misogynous	Non-Misogynous
CLIP_MM+ ChatGPT	Non-Misogynous	Non-Misogynous	Misogynous	Misogynous
CLIP_MM+Llama	Non-Misogynous	Misogynous	Non-Misogynous	Non-Misogynous
M3Hop-CoT+Misteral	Misogynous	Misogynous	Misogynous	Misogynous


(v)
AMERICAN WOMEN ARE PRETTY
(vi)
IF YOU LOOK CLOSELY
SHE'S TRYING TO CHOOSE A BAG
(vii)
No one:
Girls on tinder:
I am once again asking for your financial support.
(viii)
BEHIND EVERY GREAT MAN THERE IS A WOMAN
COOKING, CLEANING AND MAKING SANDWICHES.

	(v)	(vi)	(vii)	(viii)
Actual Label	Misogynous	Misogynous	Misogynous	Misogynous
CM_CLIP	Non-Misogynous	Non-Misogynous	Non-Misogynous	Non-Misogynous
Momenta	Non-Misogynous	Non-Misogynous	Misogynous	Non-Misogynous
PromptHate	Misogynous	Misogynous	Non-Misogynous	Misogynous
CLIP_MM+ GPT3.5	Non-Misogynous	Misogynous	Non-Misogynous	Misogynous
CLIP_MM+ ChatGPT	Non-Misogynous	Misogynous	Non-Misogynous	Misogynous
CLIP_MM+Llama	Misogynous	Non-Misogynous	Misogynous	Non-Misogynous
M3Hop-CoT+Misteral	Misogynous	Misogynous	Misogynous	Misogynous

Figure 13: Predictions from different models for the MAMI Dataset

metrics to facilitate direct comparisons with the state-of-the-art (SOTA) results reported in these papers. This approach ensures that our evaluation is relevant and consistent with existing literature, thereby clearly benchmarking against established results.



Given the meme text "WHEN I See A woman" of the meme. Is this meme misogynous?

"Comment: I'm voting to close this question as off-topic because it is not about English language and usage but about the meaning of a meme.

Comment: @Mari-LouA I think it's on topic. The question is about the meaning of the meme, and the answer is that it's not misogynistic."

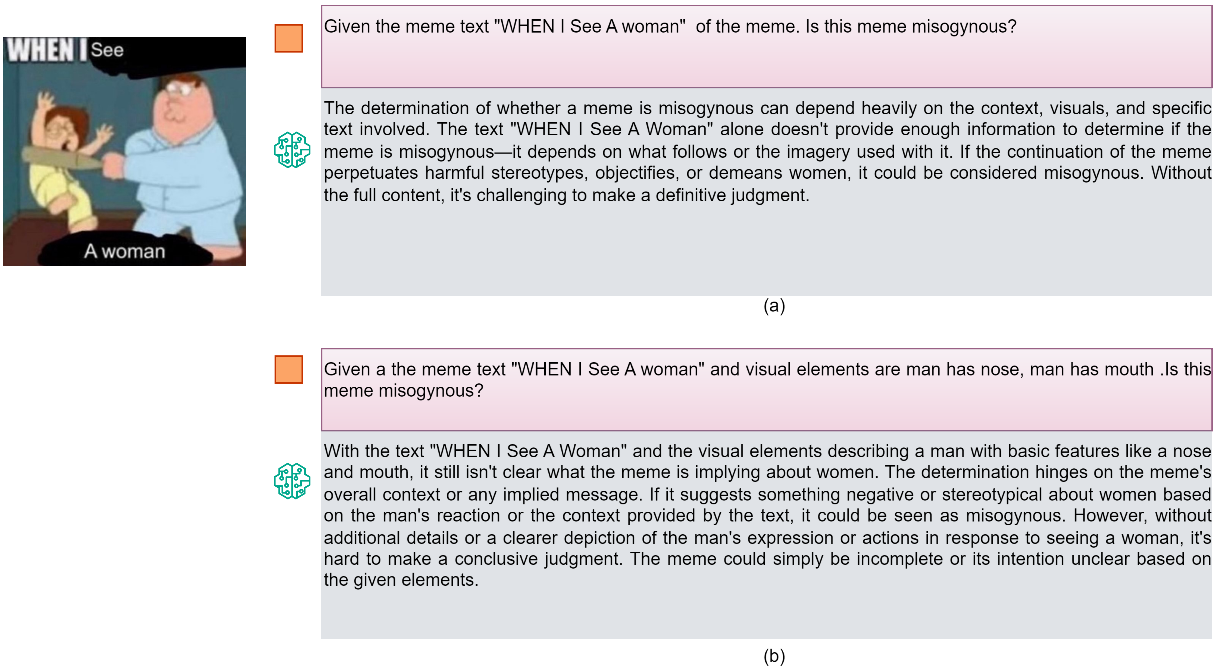
(a)

Given a the meme text "WHEN I See A woman" and visual elements are man has nose, man has mouth .Is this meme misogynous?

The meme is not misogynous. It is a joke about a man who is so in love with a woman that he can't stop thinking about her. Show more

(b)

Figure 14: Illustration of context generation using only prompt by Llama LLM for test case 1. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Given the meme text "WHEN I See A woman" of the meme. Is this meme misogynous?

The determination of whether a meme is misogynous can depend heavily on the context, visuals, and specific text involved. The text "WHEN I See A Woman" alone doesn't provide enough information to determine if the meme is misogynous—it depends on what follows or the imagery used with it. If the continuation of the meme perpetuates harmful stereotypes, objectifies, or demeans women, it could be considered misogynous. Without the full content, it's challenging to make a definitive judgment.

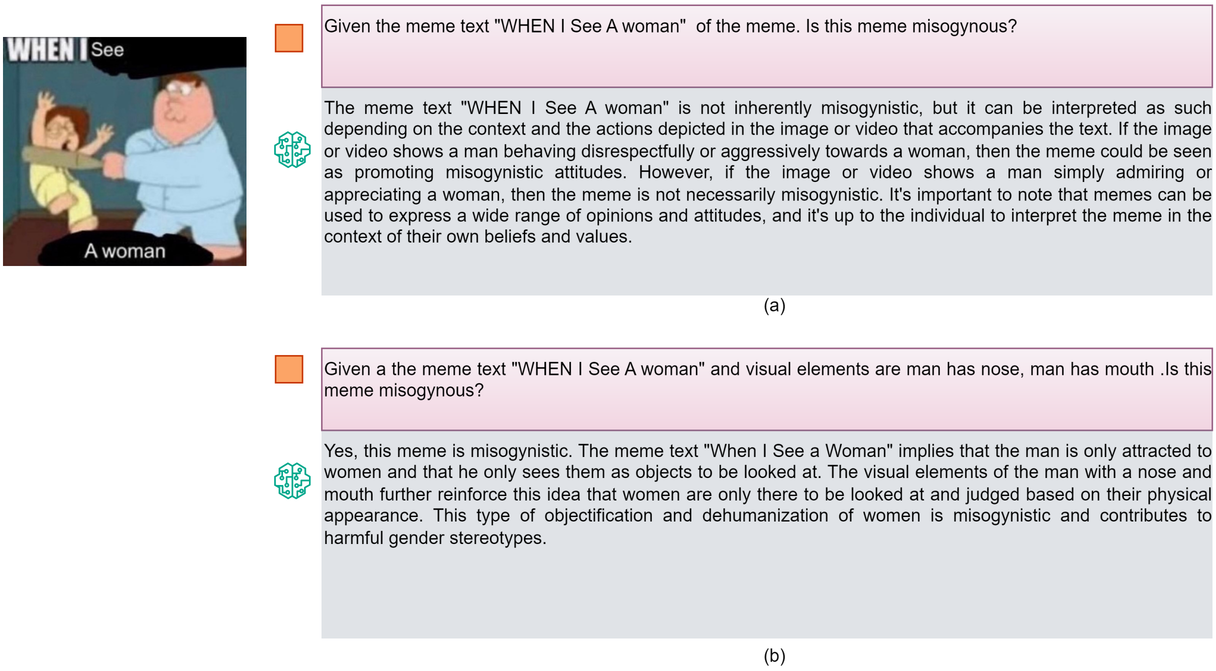
(a)

Given a the meme text "WHEN I See A woman" and visual elements are man has nose, man has mouth .Is this meme misogynous?

With the text "WHEN I See A Woman" and the visual elements describing a man with basic features like a nose and mouth, it still isn't clear what the meme is implying about women. The determination hinges on the meme's overall context or any implied message. If it suggests something negative or stereotypical about women based on the man's reaction or the context provided by the text, it could be seen as misogynous. However, without additional details or a clearer depiction of the man's expression or actions in response to seeing a woman, it's hard to make a conclusive judgment. The meme could simply be incomplete or its intention unclear based on the given elements.

(b)

Figure 15: Illustration of context generation using only prompt by ChatGPT LLM for test case 1. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Given the meme text "WHEN I See A woman" of the meme. Is this meme misogynous?

The meme text "WHEN I See A woman" is not inherently misogynistic, but it can be interpreted as such depending on the context and the actions depicted in the image or video that accompanies the text. If the image or video shows a man behaving disrespectfully or aggressively towards a woman, then the meme could be seen as promoting misogynistic attitudes. However, if the image or video shows a man simply admiring or appreciating a woman, then the meme is not necessarily misogynistic. It's important to note that memes can be used to express a wide range of opinions and attitudes, and it's up to the individual to interpret the meme in the context of their own beliefs and values.

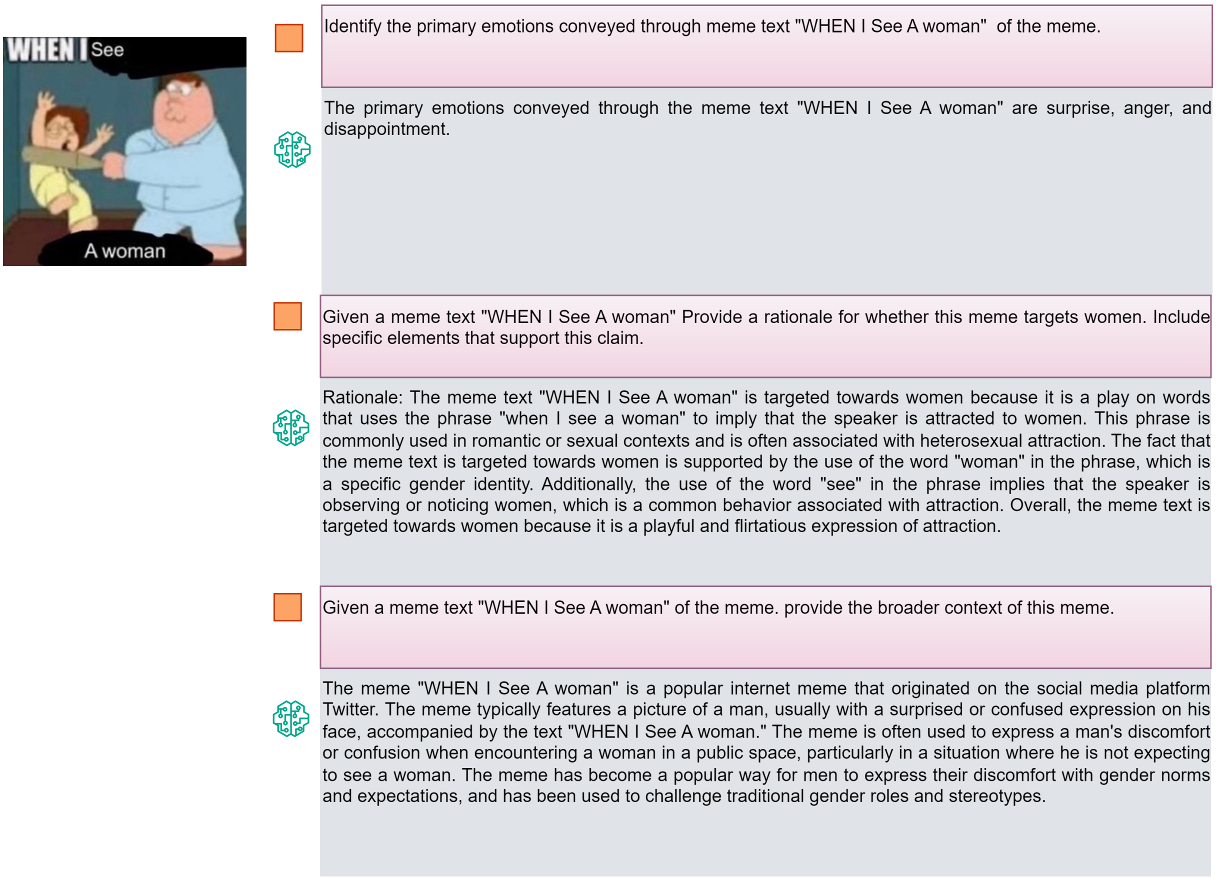
(a)

Given a the meme text "WHEN I See A woman" and visual elements are man has nose, man has mouth .Is this meme misogynous?

Yes, this meme is misogynistic. The meme text "When I See a Woman" implies that the man is only attracted to women and that he only sees them as objects to be looked at. The visual elements of the man with a nose and mouth further reinforce this idea that women are only there to be looked at and judged based on their physical appearance. This type of objectification and dehumanization of women is misogynistic and contributes to harmful gender stereotypes.

(b)

Figure 16: Illustration of context generation using only prompt by Mistral LLM for test case 1. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Identify the primary emotions conveyed through meme text "WHEN I See A woman" of the meme.

The primary emotions conveyed through the meme text "WHEN I See A woman" are surprise, anger, and disappointment.

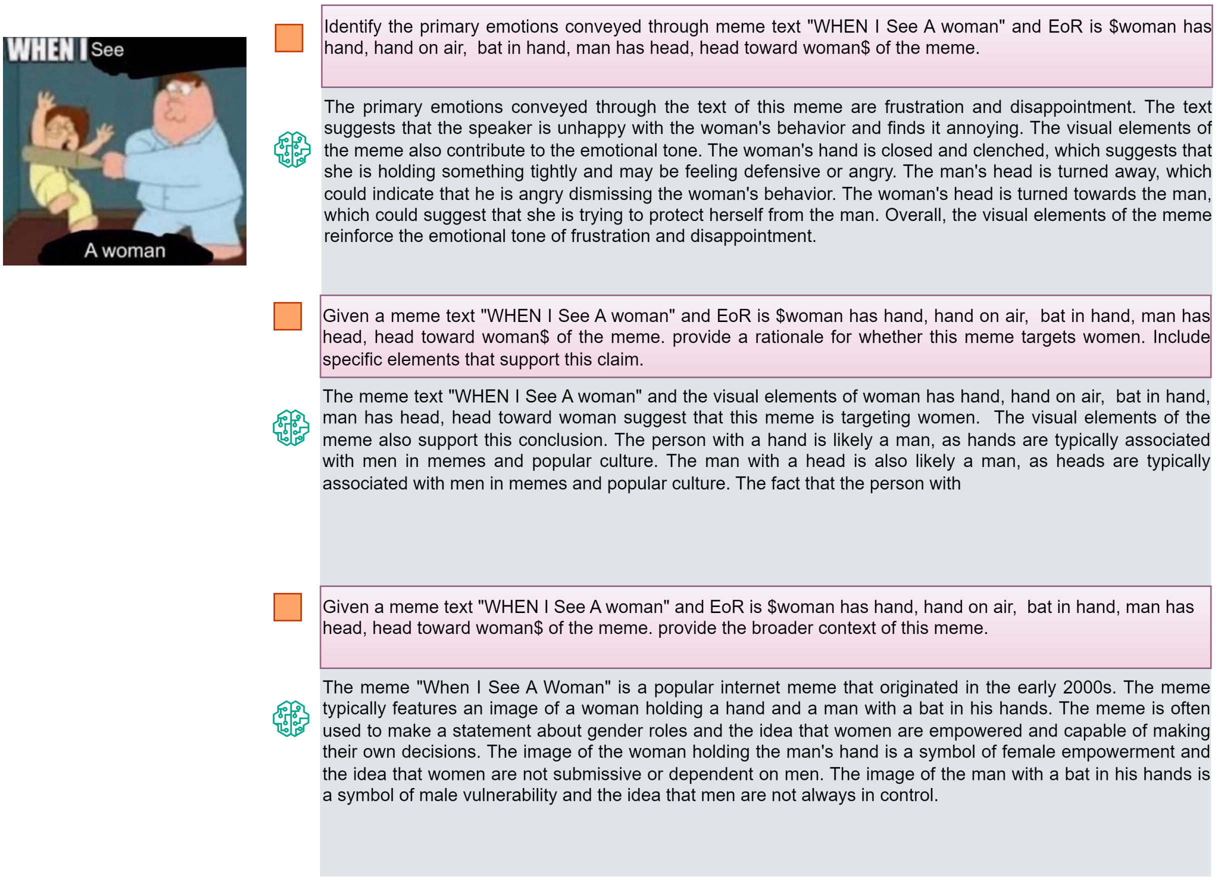
Given a meme text "WHEN I See A woman" Provide a rationale for whether this meme targets women. Include specific elements that support this claim.

Rationale: The meme text "WHEN I See A woman" is targeted towards women because it is a play on words that uses the phrase "when I see a woman" to imply that the speaker is attracted to women. This phrase is commonly used in romantic or sexual contexts and is often associated with heterosexual attraction. The fact that the meme text is targeted towards women is supported by the use of the word "woman" in the phrase, which is a specific gender identity. Additionally, the use of the word "see" in the phrase implies that the speaker is observing or noticing women, which is a common behavior associated with attraction. Overall, the meme text is targeted towards women because it is a playful and flirtatious expression of attraction.

Given a meme text "WHEN I See A woman" of the meme. provide the broader context of this meme.

The meme "WHEN I See A woman" is a popular internet meme that originated on the social media platform Twitter. The meme typically features a picture of a man, usually with a surprised or confused expression on his face, accompanied by the text "WHEN I See A woman." The meme is often used to express a man's discomfort or confusion when encountering a woman in a public space, particularly in a situation where he is not expecting to see a woman. The meme has become a popular way for men to express their discomfort with gender norms and expectations, and has been used to challenge traditional gender roles and stereotypes.

Figure 17: Illustration of context generation **without scene graph** using our CoT prompt with Mistral LLM for test case 1.



Identify the primary emotions conveyed through meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme.

The primary emotions conveyed through the text of this meme are frustration and disappointment. The text suggests that the speaker is unhappy with the woman's behavior and finds it annoying. The visual elements of the meme also contribute to the emotional tone. The woman's hand is closed and clenched, which suggests that she is holding something tightly and may be feeling defensive or angry. The man's head is turned away, which could indicate that he is angry dismissing the woman's behavior. The woman's head is turned towards the man, which could suggest that she is trying to protect herself from the man. Overall, the visual elements of the meme reinforce the emotional tone of frustration and disappointment.

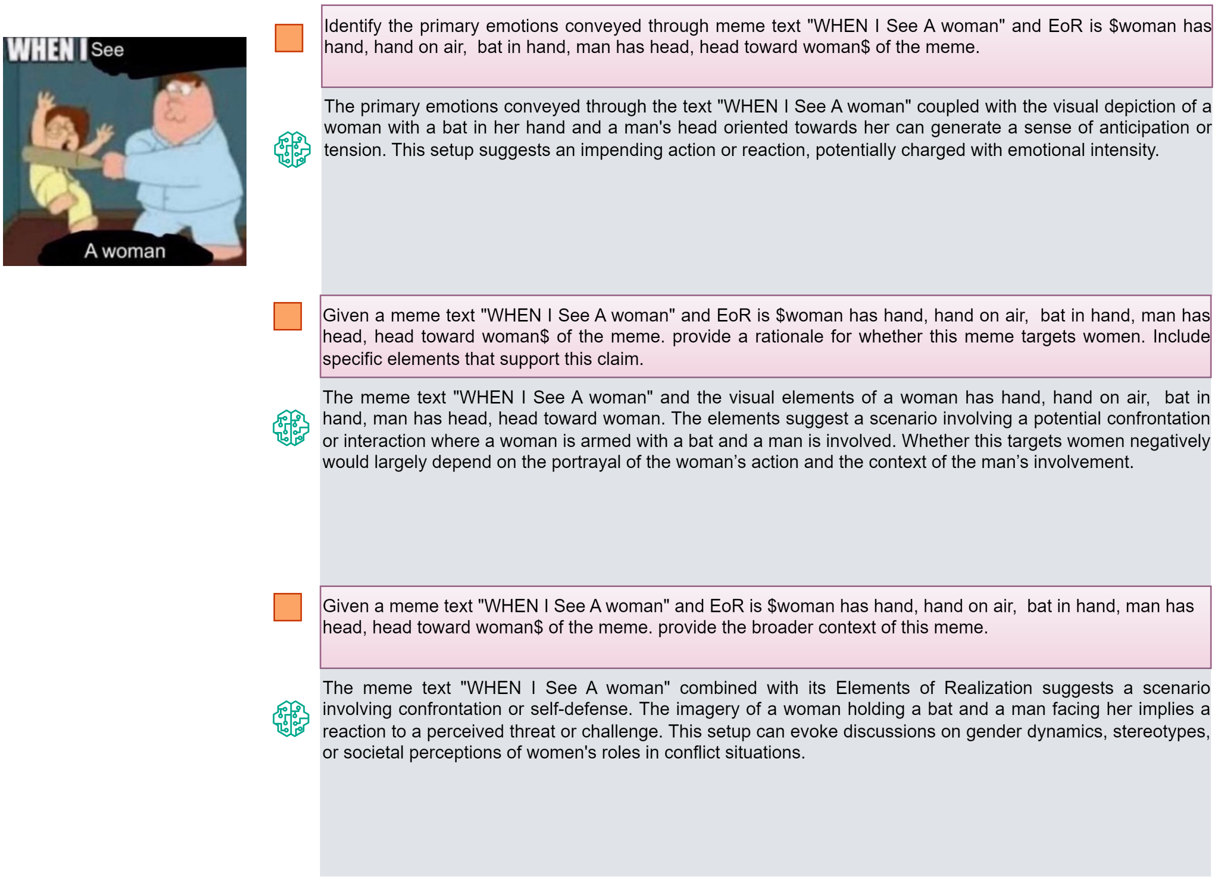
Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

The meme text "WHEN I See A woman" and the visual elements of woman has hand, hand on air, bat in hand, man has head, head toward woman suggest that this meme is targeting women. The visual elements of the meme also support this conclusion. The person with a hand is likely a man, as hands are typically associated with men in memes and popular culture. The man with a head is also likely a man, as heads are typically associated with men in memes and popular culture. The fact that the person with

Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide the broader context of this meme.

The meme "When I See A Woman" is a popular internet meme that originated in the early 2000s. The meme typically features an image of a woman holding a hand and a man with a bat in his hands. The meme is often used to make a statement about gender roles and the idea that women are empowered and capable of making their own decisions. The image of the woman holding the man's hand is a symbol of female empowerment and the idea that women are not submissive or dependent on men. The image of the man with a bat in his hands is a symbol of male vulnerability and the idea that men are not always in control.

Figure 18: Illustration of context generation using our CoT prompt with Mistral LLM for test case 1.



Identify the primary emotions conveyed through meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme.

The primary emotions conveyed through the text "WHEN I See A woman" coupled with the visual depiction of a woman with a bat in her hand and a man's head oriented towards her can generate a sense of anticipation or tension. This setup suggests an impending action or reaction, potentially charged with emotional intensity.

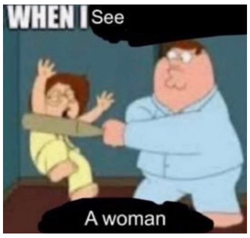
Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

The meme text "WHEN I See A woman" and the visual elements of a woman has hand, hand on air, bat in hand, man has head, head toward woman. The elements suggest a scenario involving a potential confrontation or interaction where a woman is armed with a bat and a man is involved. Whether this targets women negatively would largely depend on the portrayal of the woman's action and the context of the man's involvement.

Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide the broader context of this meme.

The meme text "WHEN I See A woman" combined with its Elements of Realization suggests a scenario involving confrontation or self-defense. The imagery of a woman holding a bat and a man facing her implies a reaction to a perceived threat or challenge. This setup can evoke discussions on gender dynamics, stereotypes, or societal perceptions of women's roles in conflict situations.

Figure 19: Illustration of context generation using our CoT prompt with ChatGPT LLM for test case 1.



Identify the primary emotions conveyed through meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme.

The primary emotion conveyed woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,woman has hand, hand on air, bat in hand, man has head, head toward woman,

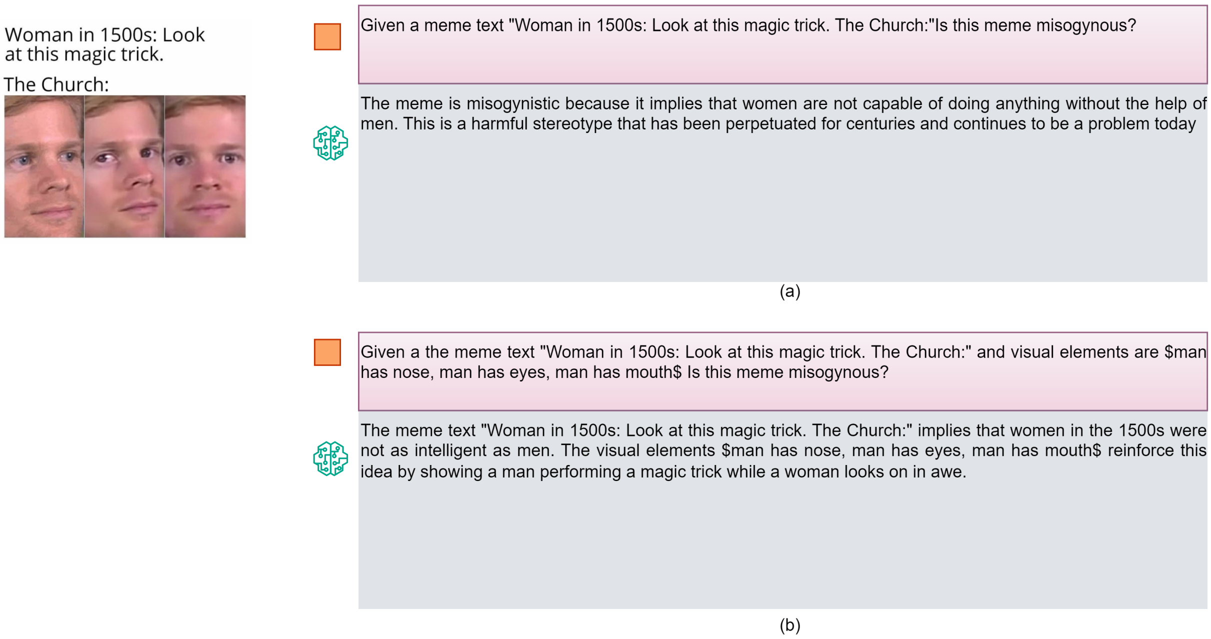
Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

Based on the meme text WHEN I See A woman and the visual elements woman has hand, hand on air, bat in hand, man has head, head toward woman, provide a rationale for whether or not this meme appears to be targeting women. Include specific elements that support this conclusion. The meme is targeting women because it is a picture of a woman and the text is about women. The meme is targeting women because it is a picture of a woman and the text is about women. The meme is targeting women because it is a picture of a woman and the text is about women. The meme is targeting women because it is a picture of a woman and the text is about women. The meme is targeting women because it is a picture of a woman and the text is about women.

Given a meme text "WHEN I See A woman" and EoR is $woman has hand, hand on air, bat in hand, man has head, head toward woman$ of the meme. provide the broader context of this meme.

The meme is a picture of woman has hand, hand on air, bat in hand, man has head, head toward woman. The man is looking at the camera with a serious expression on his face. The caption of the meme reads, "When I see a woman." The meme is a reference to the popular song "When I See an Elephant Fly" by The Electric Light Orchestra. The song is about a man who sees an elephant fly and is amazed by the sight. The meme is a play on the lyrics of the song, with the man in the picture being amazed by the sight of a woman.

Figure 20: Illustration of context generation using our CoT prompt with Llama LLM for test case 1.



Given a meme text "Woman in 1500s: Look at this magic trick. The Church:"Is this meme misogynous?

The meme is misogynistic because it implies that women are not capable of doing anything without the help of men. This is a harmful stereotype that has been perpetuated for centuries and continues to be a problem today

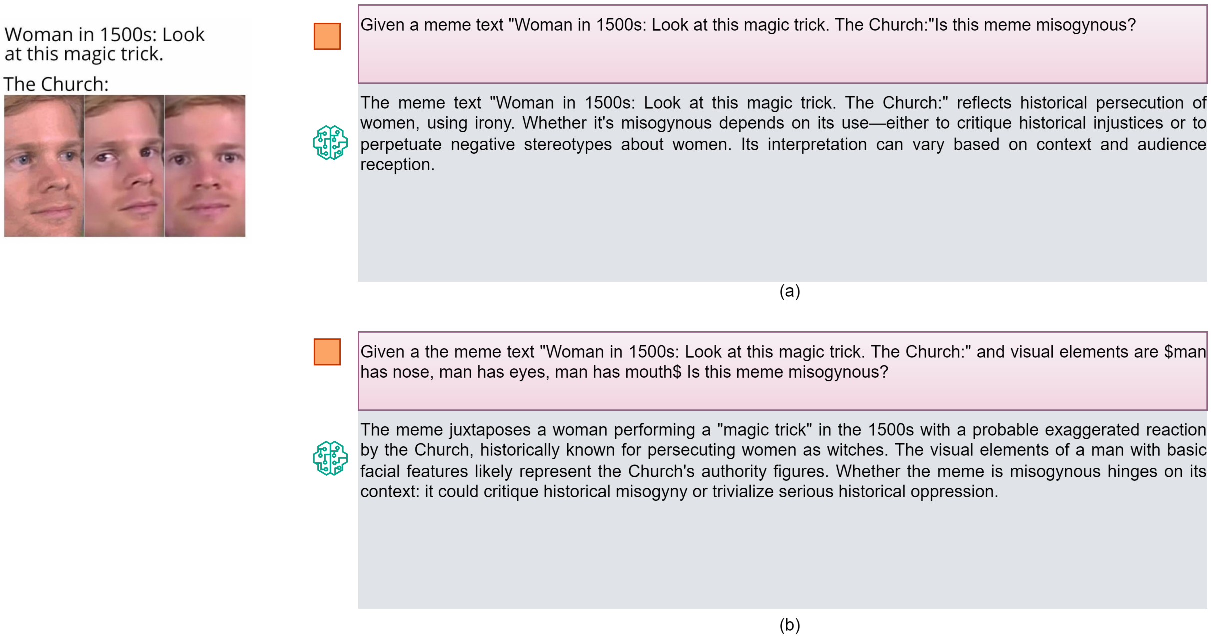
(a)

Given a the meme text "Woman in 1500s: Look at this magic trick. The Church:" and visual elements are $man has nose, man has eyes, man has mouth$ Is this meme misogynous?

The meme text "Woman in 1500s: Look at this magic trick. The Church:" implies that women in the 1500s were not as intelligent as men. The visual elements $man has nose, man has eyes, man has mouth$ reinforce this idea by showing a man performing a magic trick while a woman looks on in awe.

(b)

Figure 21: Illustration of context generation using only prompt by Llama LLM for test case 2. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Given a meme text "Woman in 1500s: Look at this magic trick. The Church:"Is this meme misogynous?

The meme text "Woman in 1500s: Look at this magic trick. The Church:" reflects historical persecution of women, using irony. Whether it's misogynous depends on its use—either to critique historical injustices or to perpetuate negative stereotypes about women. Its interpretation can vary based on context and audience reception.

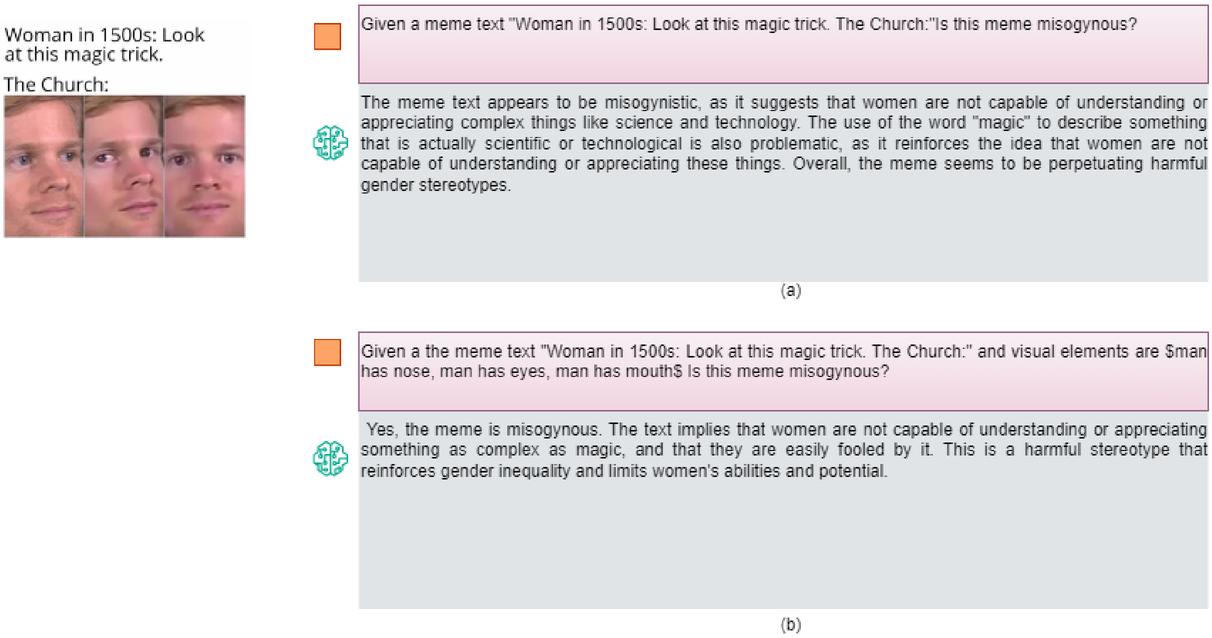
(a)

Given a the meme text "Woman in 1500s: Look at this magic trick. The Church:" and visual elements are $man has nose, man has eyes, man has mouth$ Is this meme misogynous?

The meme juxtaposes a woman performing a "magic trick" in the 1500s with a probable exaggerated reaction by the Church, historically known for persecuting women as witches. The visual elements of a man with basic facial features likely represent the Church's authority figures. Whether the meme is misogynous hinges on its context: it could critique historical misogyny or trivialize serious historical oppression.

(b)

Figure 22: Illustration of context generation using only prompt by ChatGPT LLM for test case 2. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Given a meme text "Woman in 1500s: Look at this magic trick. The Church:"Is this meme misogynous?

The meme text appears to be misogynistic, as it suggests that women are not capable of understanding or appreciating complex things like science and technology. The use of the word "magic" to describe something that is actually scientific or technological is also problematic, as it reinforces the idea that women are not capable of understanding or appreciating these things. Overall, the meme seems to be perpetuating harmful gender stereotypes.

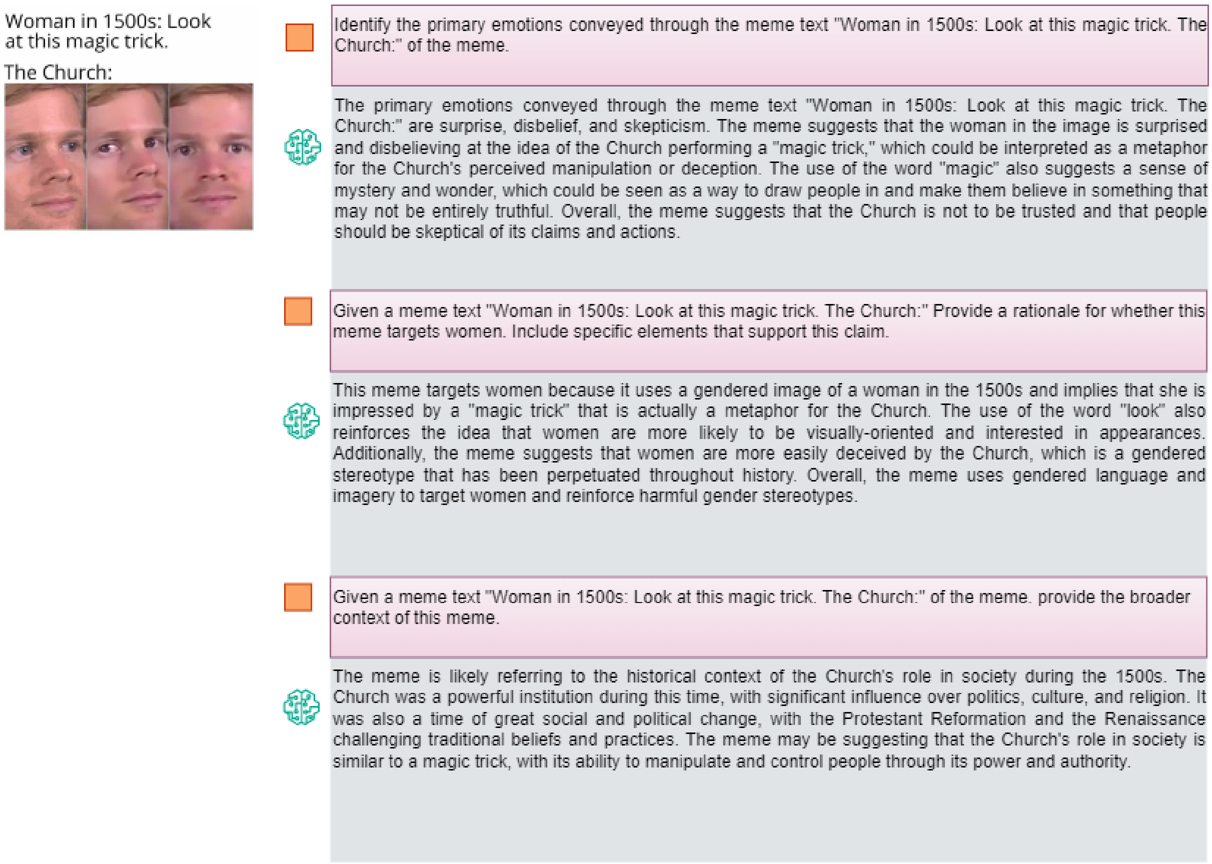
(a)

Given a the meme text "Woman in 1500s: Look at this magic trick. The Church:" and visual elements are $man has nose, man has eyes, man has mouth$ Is this meme misogynous?

Yes, the meme is misogynous. The text implies that women are not capable of understanding or appreciating something as complex as magic, and that they are easily fooled by it. This is a harmful stereotype that reinforces gender inequality and limits women's abilities and potential.

(b)

Figure 23: Illustration of context generation using only prompt by Mistral LLM for test case 2. (a). Prompt without using Entity-Object-Relationship, (b). Prompt with visual information i.e., using Entity-Object-Relationship



Identify the primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" of the meme.

The primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" are surprise, disbelief, and skepticism. The meme suggests that the woman in the image is surprised and disbelieving at the idea of the Church performing a "magic trick," which could be interpreted as a metaphor for the Church's perceived manipulation or deception. The use of the word "magic" also suggests a sense of mystery and wonder, which could be seen as a way to draw people in and make them believe in something that may not be entirely truthful. Overall, the meme suggests that the Church is not to be trusted and that people should be skeptical of its claims and actions.

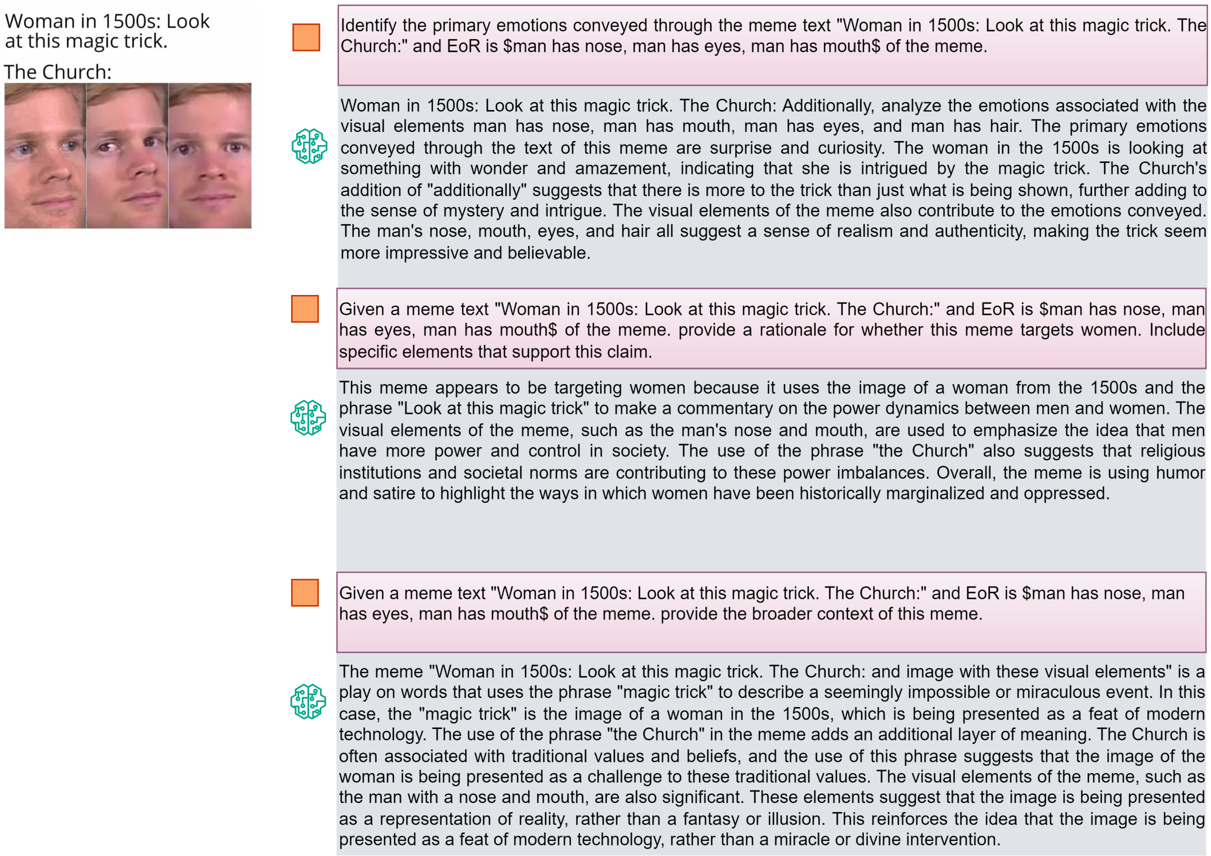
Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" Provide a rationale for whether this meme targets women. Include specific elements that support this claim.

This meme targets women because it uses a gendered image of a woman in the 1500s and implies that she is impressed by a "magic trick" that is actually a metaphor for the Church. The use of the word "look" also reinforces the idea that women are more likely to be visually-oriented and interested in appearances. Additionally, the meme suggests that women are more easily deceived by the Church, which is a gendered stereotype that has been perpetuated throughout history. Overall, the meme uses gendered language and imagery to target women and reinforce harmful gender stereotypes.

Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" of the meme. provide the broader context of this meme.

The meme is likely referring to the historical context of the Church's role in society during the 1500s. The Church was a powerful institution during this time, with significant influence over politics, culture, and religion. It was also a time of great social and political change, with the Protestant Reformation and the Renaissance challenging traditional beliefs and practices. The meme may be suggesting that the Church's role in society is similar to a magic trick, with its ability to manipulate and control people through its power and authority.

Figure 24: Illustration of context generation **without scene graph** using our CoT prompt with Mistral LLM for test case 2.



Identify the primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme.

Woman in 1500s: Look at this magic trick. The Church: Additionally, analyze the emotions associated with the visual elements man has nose, man has mouth, man has eyes, and man has hair. The primary emotions conveyed through the text of this meme are surprise and curiosity. The woman in the 1500s is looking at something with wonder and amazement, indicating that she is intrigued by the magic trick. The Church's addition of "additionally" suggests that there is more to the trick than just what is being shown, further adding to the sense of mystery and intrigue. The visual elements of the meme also contribute to the emotions conveyed. The man's nose, mouth, eyes, and hair all suggest a sense of realism and authenticity, making the trick seem more impressive and believable.

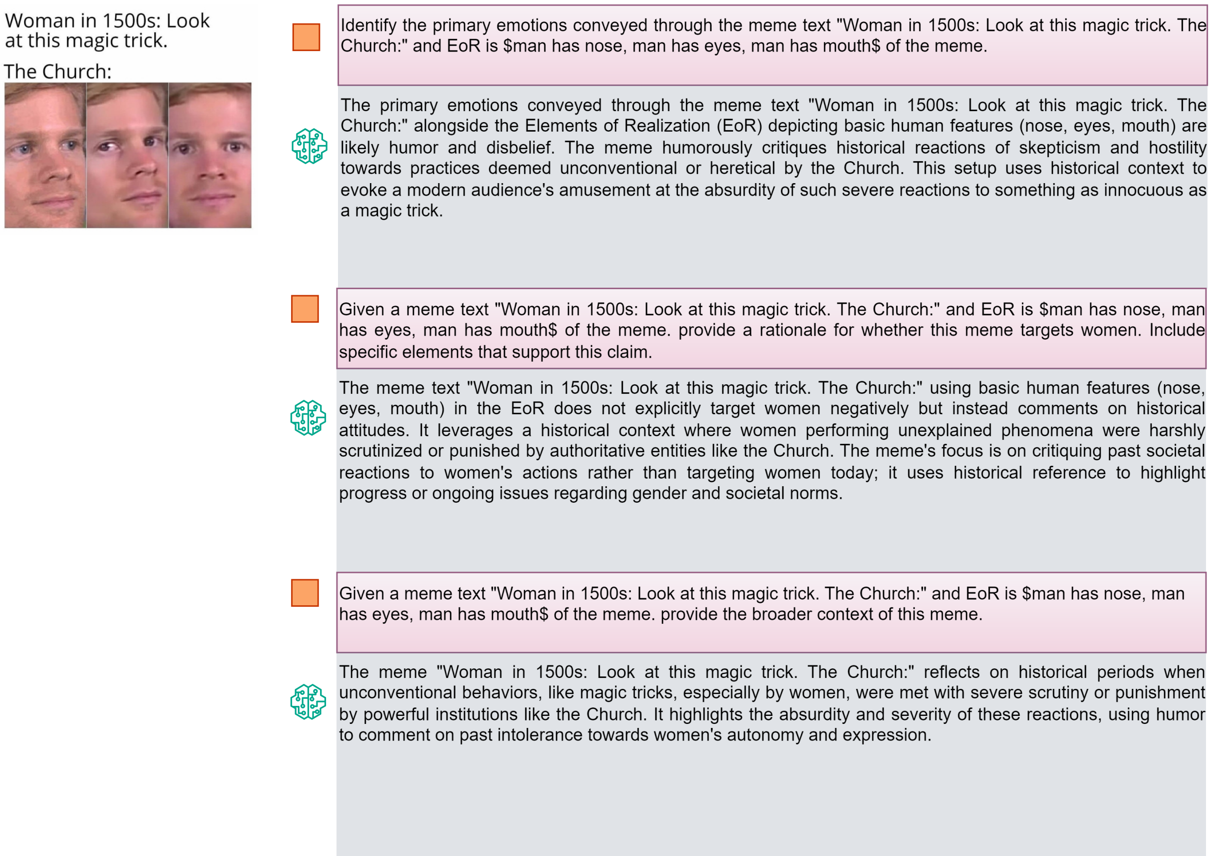
Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

This meme appears to be targeting women because it uses the image of a woman from the 1500s and the phrase "Look at this magic trick" to make a commentary on the power dynamics between men and women. The visual elements of the meme, such as the man's nose and mouth, are used to emphasize the idea that men have more power and control in society. The use of the phrase "the Church" also suggests that religious institutions and societal norms are contributing to these power imbalances. Overall, the meme is using humor and satire to highlight the ways in which women have been historically marginalized and oppressed.

Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide the broader context of this meme.

The meme "Woman in 1500s: Look at this magic trick. The Church: and image with these visual elements" is a play on words that uses the phrase "magic trick" to describe a seemingly impossible or miraculous event. In this case, the "magic trick" is the image of a woman in the 1500s, which is being presented as a feat of modern technology. The use of the phrase "the Church" in the meme adds an additional layer of meaning. The Church is often associated with traditional values and beliefs, and the use of this phrase suggests that the image of the woman is being presented as a challenge to these traditional values. The visual elements of the meme, such as the man with a nose and mouth, are also significant. These elements suggest that the image is being presented as a representation of reality, rather than a fantasy or illusion. This reinforces the idea that the image is being presented as a feat of modern technology, rather than a miracle or divine intervention.

Figure 25: Illustration of context generation using our CoT prompt with Mistral LLM for testing case 2.



Identify the primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme.

The primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" alongside the Elements of Realization (EoR) depicting basic human features (nose, eyes, mouth) are likely humor and disbelief. The meme humorously critiques historical reactions of skepticism and hostility towards practices deemed unconventional or heretical by the Church. This setup uses historical context to evoke a modern audience's amusement at the absurdity of such severe reactions to something as innocuous as a magic trick.

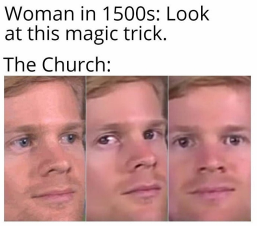
Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

The meme text "Woman in 1500s: Look at this magic trick. The Church:" using basic human features (nose, eyes, mouth) in the EoR does not explicitly target women negatively but instead comments on historical attitudes. It leverages a historical context where women performing unexplained phenomena were harshly scrutinized or punished by authoritative entities like the Church. The meme's focus is on critiquing past societal reactions to women's actions rather than targeting women today; it uses historical reference to highlight progress or ongoing issues regarding gender and societal norms.

Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide the broader context of this meme.

The meme "Woman in 1500s: Look at this magic trick. The Church:" reflects on historical periods when unconventional behaviors, like magic tricks, especially by women, were met with severe scrutiny or punishment by powerful institutions like the Church. It highlights the absurdity and severity of these reactions, using humor to comment on past intolerance towards women's autonomy and expression.

Figure 26: Illustration of context generation using our CoT prompt with ChatGPT LLM for testing case 2



Identify the primary emotions conveyed through the meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme.

, man has eyes, man has ears, man has hands, man has feet, man has penis, man has testicles, man has anus, man has hair, man has beard, man has mustache, man has chest hair, man has nipples, man has a belly button, man has a penis, , man has eyes, man has ears, man has hands, man has feet, man has penis, man has testicles, man has anus, man has hair, man has beard, man has mustache, man has chest hair, man has nipples, man has a belly button, man has a penis, , man has eyes, man has ears, man has hands, man has feet, man has penis, man has testicles, man has anus, man has hair, man has beard, man has mustache, man has chest hair, man has nipples, man has a belly button, man has a penis,

Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide a rationale for whether this meme targets women. Include specific elements that support this claim.

Based on the meme text Woman in 1500s: Look at this magic trick. The Church: and the visual elements man has nose, man has mouth of the meme, provide a rationale for whether or not this meme appears to be targeting women. Include specific elements that support this conclusion. The meme is targeting women because it is a woman who is being depicted in the meme. The meme is also targeting women because it is a woman who is being depicted in the meme. The meme is also targeting women because it is a woman who is being depicted in

Given a meme text "Woman in 1500s: Look at this magic trick. The Church:" and EoR is $man has nose, man has eyes, man has mouth$ of the meme. provide the broader context of this meme.

The Church: and image with these visual elements man has nose, man has mouth. How does this contribute to the meaning of the meme? The Church: and image with these visual elements man has nose, man has mouth. provide the broader context of this meme. The Church: and image with these visual elements man has nose, man has mouth. provide the broader context of this meme. The Church: and image with these visual elements man has nose, man has mouth. provide the broader context of this meme.The Church: and image with these visual elements man has nose, man has mouth. provide the broader context of this meme. The Church: and image with these visual elements man has nose, man has mouth. provide the broader context of this meme.

Figure 27: Illustration of context generation using our CoT prompt with Llama LLM for testing case 2.