MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection

Ziyan Liu, Chunxiao Fan†, Haoran Lou, Yuexin Wu, Kaiwei Deng
Beijing University of Posts and Telecommunications
{liuziyan,cxfan}@bupt.edu.cn

Abstract

The rapid expansion of memes on social media has highlighted the urgent need for effective approaches to detect harmful content. However, traditional data-driven approaches struggle to detect new memes due to their evolving nature and the lack of up-to-date annotated data. To address this issue, we propose MIND, a multi-agent framework for zero-shot harmful meme detection that does not rely on annotated data. MIND implements three key strategies: 1) We retrieve similar memes from an unannotated reference set to provide contextual information. 2) We propose a bi-directional insight derivation mechanism to extract a comprehensive understanding of similar memes. 3) We then employ a multi-agent debate mechanism to ensure robust decision-making through reasoned arbitration. Extensive experiments on three meme datasets demonstrate that our proposed framework not only outperforms existing zero-shot approaches but also shows strong generalization across different model architectures and parameter scales, providing a scalable solution for harmful meme detection. The code is available at https://github.com/destroy-lonely/MIND.

1 Introduction

With the rapid expansion of social media platforms, a new multimodal entity known as the meme has emerged. A meme typically comprises an image combined with a concise textual element, enabling it to spread swiftly across the internet. Memes have become a prevalent form of multimodal content. While often intended to be humorous, memes are increasingly created or manipulated to convey harmful messages, particularly when they are used to exploit political and socio-cultural divides.

Such memes are referred to as harmful memes[1] and are generally defined as "multimodal units consisting of an image and embedded text that have the potential to cause harm to an individual, an organization, a community, or society" (Sharma et al., 2022). These harmful memes can spread misinformation or perpetuate harmful stereotypes, posing risks to individuals, organizations, and communities. For example, during the COVID-19 pandemic, a frequently shared meme shown in Figure 1(a) featured Dr. Fauci and President Trump at White House press briefings, which was often manipulated and repurposed by various groups skeptical of public health measures. This meme not only undermined both figures' public credibility but also potentially damaged broader public health communication efforts during the pandemic. Therefore, automatically understanding and detecting harmful memes becomes increasingly important in maintaining social harmony and integrity on social media.

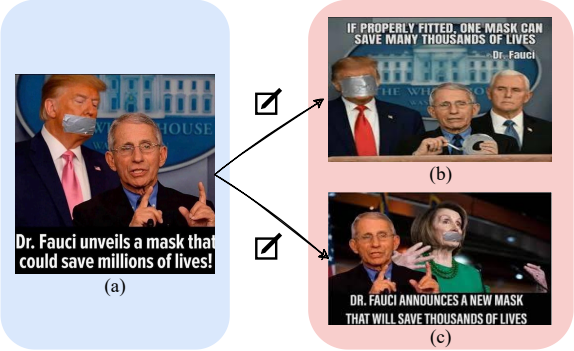


Figure 1: Example of trending memes on social media, where (b) and (c) are modified versions of (a). **Meme text**: (a) "*Dr. Fauci unveils a mask that could save millions of lives!*"; (b) "*IF PROPERLY FITTED, ONE MASK CAN SAVE MANY THOUSANDS OF LIVES Dr. Fauci*"; (c) "*DR. FAUCI ANNOUNCES A NEW MASK THAT WILL SAVE THOUSANDS OF LIVES*"

Previous studies in detecting harmful memes have primarily utilized data-driven multimodal models (Velioglu and Rose, 2020; Lippe et al.,

†Corresponding author.

[1]**Disclaimer:** This paper contains content that may be disturbing to some readers.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 923–947

2020; Muennighoff, 2020; Cao et al., 2022; Lin et al., 2024, 2023; Pramanick et al., 2021b), which depend on large volumes of high-quality annotated data. These models encounter significant challenges in identifying new memes that quickly emerge in response to current events, as rapidly collecting and annotating sufficient data is challenging. While recent research has explored few-shot in-context learning to enhance detection capabilities with minimal annotations (Cao et al., 2024; Huang et al., 2024), even these approaches still depend on pre-existing annotated data, limiting their adaptability to the fast-paced evolution of harmful memes.

This persistent challenge of data annotation underscores the critical need for methods that can operate without labeled data. We advocate that zero-shot approaches merit more attention, as they are crucial for developing detection systems with the adaptability required to effectively navigate the ever-evolving landscape of memes. Our key insight is that despite evolving into new formats, memes often retain core characteristics that can be identified through careful analysis of similar examples. For instance, as depicted in Figure 1, a meme featuring a White House press briefing could be modified and repurposed in various ways, while maintaining the same core elements from the original setting.

Inspired by this observation and the advanced reasoning capabilities of Large Multimodal Models (LMMs), we propose a novel framework: **MIND**, **M**ulti-agent **IN**sight derivation for harmful meme **D**etection. Our approach leverages a multi-agent framework that mimics human collaborative analysis: 1) First, it retrieves similar memes from an unannotated reference set to provide contextual information. 2) Then, through bi-directional insight derivation, two agents collaboratively process these memes to extract comprehensive understanding. 3) Finally, multi-agent debate mechanism enables agents to evaluate the derived insights and resolve potential conflicts through reasoned arbitration. Through these strategies, our approach enables robust harmful meme detection by leveraging multi-agent reasoning on patterns observed across similar memes, effectively adapting to new and evolving content without relying on annotated data. Our contributions can be summarized in three folds:

- To the best of our knowledge, we pioneer the use of a novel multi-agent framework for zero-shot harmful meme detection, which eliminates the need for annotated data.
- We propose MIND, a multi-agent framework that analyzes retrieved similar memes through a novel bidirectional insight derivation and leverages a debate-based reasoning mechanism for robust harm detection.
- Extensive experiments on three meme datasets demonstrate that our proposed framework significantly outperforms existing zero-shot state-of-the-art baselines for harmful meme detection.

2 Related work

2.1 Harmful Meme Detection

Harmful meme detection has emerged as a significant research focus, moving forward with the development of extensive benchmarks (Pramanick et al., 2021b,a; Fersini et al., 2022; Kiela et al., 2020). Due to the inherently multimodal nature of memes, which incorporate both text and imagery, conventional unimodal approaches (Simonyan and Zisserman, 2014; He et al., 2016; Devlin et al., 2019; Raffel et al., 2020) have often proven insufficient. In response, recent research has increasingly adopted multimodal strategies (Dosovitskiy, 2020; Radford et al., 2021), aiming to improve detection efficacy by integrating both textual and visual data.

Previous research has leveraged classical two-stream models that combine textual and visual elements, using text and image encoders to capture these features. These systems often employ attention-based mechanisms and multimodal fusion techniques for classifying harmful memes (Kiela et al., 2019, 2020; Suryawanshi et al., 2020; Pramanick et al., 2021b). Another approach involves fine-tuning pre-trained multimodal models to specifically address the task of harmful meme detection (Lippe et al., 2020; Muennighoff, 2020; Velioglu and Rose, 2020; Hee et al., 2022). Additionally, recent efforts have explored various strategies such as data augmentation (Zhou et al., 2021; Zhu et al., 2022), ensemble methods (Zhu, 2020; Velioglu and Rose, 2020; Sandulescu, 2020), harmful target disentanglement (Lee et al., 2021), and prompt-based tuning (Cao et al., 2022; Ji et al., 2023; Cao et al., 2023). However, most of these approaches rely heavily on large-scale annotated data, which limits their ability to adapt to emerging events and novel harmful content patterns. Although recent studies have explored few-shot in-context learning approaches to address this chal-

lenge in low-resource scenarios (Cao et al., 2024; Huang et al., 2024), these approaches still fundamentally depend on annotated examples, making them insufficient for detecting harmful memes during emerging events where annotated data is scarce. In this work, we leverage LMMs to derive insights from similar memes and subsequently employ a multi-agent debate to arrive at comprehensive judgments on meme harmfulness. Without modifying models' weights or requiring annotated data, our approach offers a significant advantage in adapting to real-world scenarios.

2.2 LLM-Based Multi-Agent Frameworks

The integration of Large Language Models (LLMs) as agents spans various domains, showcasing their robust planning and reasoning capabilities in diverse settings (Wang et al., 2023a; Yao et al., 2022; Shen et al., 2023; Mu et al., 2023; Hong et al., 2023; Liu et al., 2023b; Zhao et al., 2024; Sun et al., 2023; Song et al., 2023; Miao et al., 2023; Madaan et al., 2024). These advancements underscore the ability of LLMs to tackle complex tasks with minimal supervision. Building on the success of single-agent, multi-agent frameworks (Park et al., 2023; Hong et al., 2023; Du et al., 2023; Liang et al., 2023; Wang et al., 2024; Qian et al., 2024; Tao et al., 2024; Zeng et al., 2024; D'Arcy et al., 2024; Huang et al., 2023) facilitate complex interactions and collaborative problem solving, simulating environments where multiple agents work in unison. However, existing frameworks often rely on environmental feedback to iteratively refine their decisions, which becomes impractical for zero-shot binary classification tasks like harmful meme detection, where receiving feedback would effectively reveal the correct answer. In this work, we propose a novel framework that leverages vision-language retrieval enhancement, relevant insight derivation, and multi-agent debate mechanism to enable reliable zero-shot harmful meme detection through collaborative reasoning among specialized agents.

3 Method

3.1 Overview

Problem Statement We define a harmful meme detection dataset as a collection of memes, where each meme M is represented as a tuple $\{\mathcal{V}, \mathcal{T}\}$, consisting of visual component $\mathcal{V}$ and textual component $\mathcal{T}$. In this work, we explore zero-shot harmful meme detection by leveraging LMMs in a natural language generation paradigm, where both visual and textual inputs are collaboratively processed by LMM agents to determine the harmfulness of a given meme.

The scarcity of high-quality labeled data has become a critical challenge in harmful meme detection, particularly given the rapid evolution and emergence of new memes (Sharma et al., 2022). To address this challenge, we propose MIND, a multi-agent framework that utilizes a training set S_{train}, accessing only the images and text content without using their true labels to detect harmful memes in a test set S_{test}. Since our method is gradient-free, we denote S_{train} as the reference set S_{ref}. This approach enables robust detection in real-world scenarios where labeled data is scarce or unavailable.

Our proposed framework coordinates multiple LMM agents to collaboratively analyze memes through three interconnected stages: 1) Similar Sample Retrieval(§3.2), which discovers contextually relevant memes from the unlabeled reference set S_{ref}, 2) Relevant Insight Derivation(§3.3), where agents engage in forward and backward reasoning to derive relevant insight from similar memes, and 3) Insight-Augmented Inference(§3.4), where specialized debater and judge agents deliberate to reach a well-reasoned conclusion about memes' harmfulness. Through this multi-agent collaboration, MIND achieves robust harmful meme detection without relying on annotated data. The overview of our framework is shown in Figure 2.

3.2 Similar Sample Retrieval

Internet memes continuously evolve, yet they often display common underlying patterns (Sharma et al., 2022; Qu et al., 2023). Inspired by retrieval-augmented generation (Guu et al., 2020; Lewis et al., 2020), we focus on retrieving memes similar to the target one, which enables us to utilize these similar memes as a source of insights, providing context and understanding that enhance our assessment of a meme's harmfulness.

For a meme sample $M = \{\mathcal{V}, \mathcal{T}\}$, we derive a multimodal embedding by integrating visual and textual features:

$$\mathbf{E} = \lambda_v \cdot \mathbf{V_{enc}}(\mathcal{V}) + \lambda_t \cdot \mathbf{T_{enc}}(\mathcal{T}), \quad (1)$$

where $\mathbf{E}$ is the multimodal embedding of M, $\mathbf{V_{enc}}(\cdot)$ and $\mathbf{T_{enc}}(\cdot)$ are encoders that extract visual and textual features, respectively. The coefficients λ_v and λ_t are fixed weights for combining

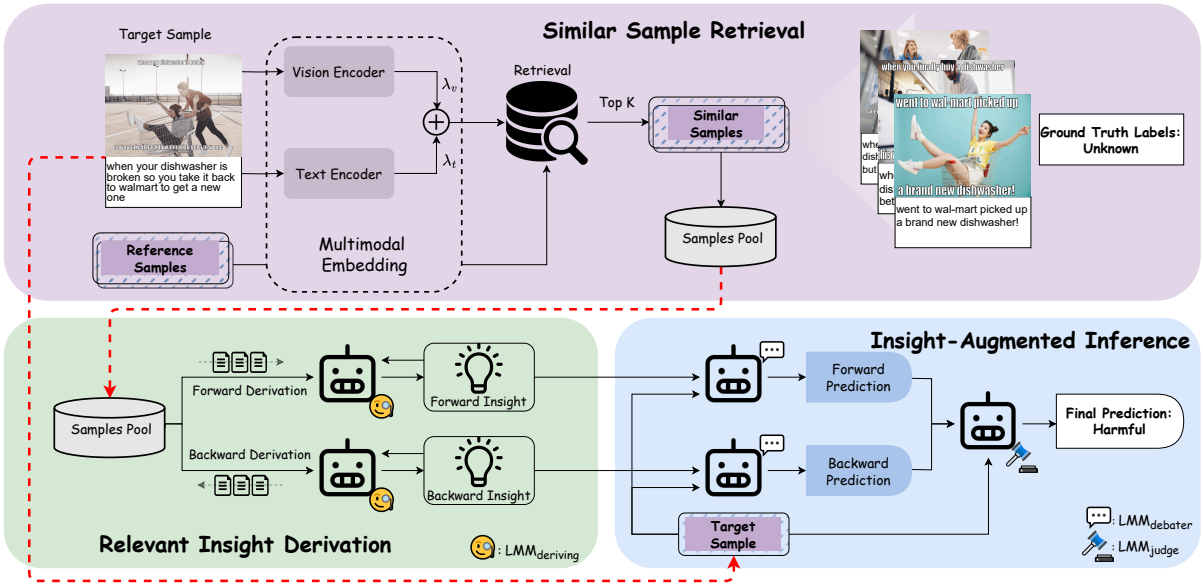


Figure 2: An overview of our framework. MIND, for zero-shot harmful meme detection.

the visual and textual modalities. We then perform the operation as in Equation 1 for all meme samples in both S_{test} and S_{ref} to obtain the embeddings of all meme samples.

To retrieve the most similar samples, we use cosine similarity to match the multimodal embeddings of the target sample and reference samples. The similarity score s is computed as follows:

$$s = \text{sim}(\mathbf{E}_{\text{target}}, \mathbf{E}_{\text{ref}}), \tag{2}$$

$$\text{sim}(\mathbf{E}_{\text{target}}, \mathbf{E}_{\text{ref}}) = \frac{\mathbf{E}_{\text{target}} \cdot \mathbf{E}_{\text{ref}}}{\|\mathbf{E}_{\text{target}}\| \|\mathbf{E}_{\text{ref}}\|}, \tag{3}$$

where $\mathbf{E}_{\text{target}}$ and $\mathbf{E}_{\text{ref}}$ are the embedding for the target sample M_{target} and a reference sample M_{ref}, respectively. $\text{sim}(\cdot)$ denotes the cosine similarity, and $\|\cdot\|$ denotes the vector norm. All the similarity scores of the candidate memes in the reference set to the target meme could form the similarity vector $S = \{s \mid M_{\text{ref}} \in S_{\text{ref}}\} \in \mathbb{R}^N$. Then, the top K reference samples with the highest similarity scores are selected as follows:

$$M_{\text{similar}} = \{M_{\text{ref}} \mid s \in \text{Top}_K(S)\}, \tag{4}$$

where $\text{Top}_K(\cdot)$ is the operation of selecting the K highest values from the similarity vector S. The similar set M_{similar} contains only the K most relevant visual and textual components of the similar reference memes, stored in the samples pool, while maintaining the unannotated nature of the reference set for subsequent analysis.

3.3 Relevant Insight Derivation

The M_{similar}, obtained through the Similar Sample Retrieval process described in Section 3.2, offers valuable context that can assist in detecting harmful memes. However, leveraging these similar memes directly can be challenging since M_{similar} lacks explicit ground truth labels. Without proper handling, they might confuse rather than assist the LMMs in making accurate judgments. To address this issue, we introduce a novel agentic approach designed to derive relevant insights from M_{similar}, ensuring that the information contributes constructively to the harmful meme detection task.

3.3.1 Forward Insight Derivation

For each meme in M_{similar}, the deriving LMM agent processes these memes sequentially. In each iteration i, one meme is input into the agent along with previously derived insights, and a new insight set is generated. This insight set serves two purposes: it captures the understanding of the current meme and provides context for analyzing subsequent memes. This iterative process is formulated as:

$$\mathcal{I}_{\text{fwd},i} = \text{LMM}_{\text{deriving}}(M_{\text{similar},i}, \mathcal{I}_{\text{fwd},i-1}, \mathcal{P}_{\text{deriving}}), \tag{5}$$

where $\mathcal{I}_{\text{fwd},i}$ and $\mathcal{I}_{\text{fwd},i-1}$ represent the derived insight sets at current and previous iterations respectively, $M_{\text{similar},i}$ is the i-th meme from the similar meme set, and $\text{LMM}_{\text{deriving}}$ is an LMM agent equipped with Chain-of-Thought insight derivation prompt $\mathcal{P}_{\text{deriving}}$ as detailed in Appendix §C and Figure 10.

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
GPT-4o (Achiam et al., 2023)	67.51	60.29	68.80	68.25	81.00	81.00
Gemini-1.5-Flash (Team et al., 2024)	66.10	64.18	60.20	58.90	76.40	74.29
LLaVA-1.5-7B (Liu et al., 2024)	59.23	49.44	53.80	45.51	52.90	41.53
InstructBLIP-7B (Dai et al., 2023)	51.13	50.99	52.00	48.85	53.10	46.93
MiniGPT-v2-7B (Chen et al., 2023)	60.12	52.39	51.30	47.88	57.40	52.22
OpenFlamingo-9B (Awadalla et al., 2023)	63.42	54.36	50.50	49.52	54.70	49.88
LLaVA-1.5-13B (Liu et al., 2024)	62.28	50.45	55.20	53.01	60.10	55.52
InstructBLIP-13B (Dai et al., 2023)	64.92	49.61	55.40	51.89	60.00	57.97
LLaVA-1.6-34B (Liu et al., 2024)	67.51	61.59	**64.00**	**63.51**	**71.30**	**71.28**
MIND (LLaVA-1.5-13B)	**68.93**	**65.19**	60.80	60.71	68.90	68.84

Table 1: Zero-shot harmful meme detection results on three datasets. The accuracy and macro-averaged F1 scores (%) are reported as the metrics. All baseline models are equipped with Chain-of-Thought prompt. The best and second best results in open-source setting are in **bold** and underlined, respectively.

Model	HarM				FHM				MAMI			
	Accuracy		Macro-F_1		Accuracy		Macro-F_1		Accuracy		Macro-F_1	
	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND
LLaVA-1.5-7B	59.23	62.71 (+3.48)	49.44	57.22 (+7.78)	53.80	54.00 (+0.20)	45.51	48.28 (+2.77)	52.90	53.90 (+1.00)	41.53	45.45 (+3.92)
LLaVA-1.5-13B	62.28	68.93 (+6.65)	50.45	65.19 (+14.74)	55.20	60.80 (+5.60)	53.01	60.71 (+7.70)	60.10	68.90 (+8.80)	55.52	68.84 (+13.32)
LLaVA-1.6-34B	67.51	69.49 (+1.98)	61.59	66.12 (+4.53)	64.00	66.40 (+2.40)	63.51	68.38 (+4.87)	71.30	73.60 (+2.30)	71.28	75.38 (+4.10)
Gemini-1.5-Flash	66.10	68.36 (+2.26)	64.18	66.92 (+2.74)	60.20	63.80 (+3.60)	58.90	62.50 (+3.60)	76.40	78.00 (+1.60)	74.29	77.89 (+3.60)

Table 2: Performance improvements of our proposed framework across different model scales and datasets for zero-shot harmful meme detection. Numbers in green indicate absolute improvements over original models.

3.3.2 Backward Insight Derivation

A challenge arises from the sequential nature of Forward Insight Derivation—earlier memes in the sequence benefit more from accumulated insights, while later memes might be less thoroughly analyzed. To address this imbalance, we propose an additional Backward Insight Derivation round. While completing the forward pass, we process similar memes in reverse order:

$$\mathcal{I}_{\text{back},i} = \text{LMM}_{\text{deriving}}(M_{\text{similar},K+1-i}, \mathcal{I}_{\text{back},i-1}, \mathcal{P}_{\text{deriving}}), \quad (6)$$

where $\mathcal{I}_{\text{back},i}$ and $\mathcal{I}_{\text{back},i-1}$ represent the derived insight sets at current and previous iterations in backward pass, with K being the total number of similar memes. Through these complementary forward and backward passes, we obtain two sets of insights $\mathcal{I}_{\text{fwd},K}$ and $\mathcal{I}_{\text{back},K}$, ensuring each meme in M_{similar} receives attention from both sequential perspectives.

3.4 Insight-Augmented Inference

To ensure robust decision-making, we employ two debater agents leveraging complementary insights derived from forward and backward analysis in Section 3.3. Each debater processes their respective insight sets along with the target meme to generate reasoning-based judgments:

$$\mathcal{J}_{\text{fwd}} = \text{LMM}_{\text{debater}}(\mathcal{I}_{\text{fwd},K}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}}), \quad (7)$$

$$\mathcal{J}_{\text{back}} = \text{LMM}_{\text{debater}}(\mathcal{I}_{\text{back},K}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}}), \quad (8)$$

where $\text{LMM}_{\text{debater}}$ is an LMM agent that generates judgments based on the derived insights and target meme content, $\mathcal{J}_{\text{fwd}}$ and $\mathcal{J}_{\text{back}}$ represent the judgments from forward and backward debater agents, and $\mathcal{V}_{\text{target}}$ and $\mathcal{T}_{\text{target}}$ denote the visual and textual components of the target meme. Each judgment contains both reasoning process and final decision. While the reasoning processes may differ, the final decisions are binary indicators of harmfulness.

When examining the judgments from two debater agents, our decision-making process follows two paths: for consensus cases, we directly adopt their shared judgment; for disagreements, a judge agent arbitrates by analyzing both debaters' reasoning. This process can be formalized as:

$$\mathcal{J}_{\text{final}} = \begin{cases} \mathcal{J}_{\text{fwd}} & \text{if } \mathcal{J}_{\text{fwd}} = \mathcal{J}_{\text{back}} \\ \text{LMM}_{\text{judge}}(\mathcal{J}_{\text{fwd}}, \mathcal{J}_{\text{back}}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}}) & \text{if } \mathcal{J}_{\text{fwd}} \neq \mathcal{J}_{\text{back}} \end{cases}, \quad (9)$$

where $\text{LMM}_{\text{judge}}$ weighs the competing arguments to reach a final judgment $\mathcal{J}_{\text{final}}$. This multi-agent

debate mechanism enhances reliability through independent assessments and reduces potential biases through diverse perspectives.

In our experiments, we set the number of similar memes K to 3 and use LLaVA-1.5-13B (Liu et al., 2024) as the backbone model for all LMM agents as it offers an optimal balance between computational efficiency and model performance. We also conduct extensive experiments with other LMMs (Liu et al., 2024; Team et al., 2024) of varying model sizes and architectures to demonstrate the generalizability of our framework (see §4.3).

4 Experiment

4.1 Experiment Setup

Datasets We use three publicly available meme datasets for evaluation: (1) HarM (Pramanick et al., 2021a), (2) FHM (Kiela et al., 2020), and (3) MAMI (Fersini et al., 2022). HarM consists of memes related to COVID-19. FHM was released by Facebook as part of a challenge to crowd-source multimodal harmful meme detection in hate speech solutions. MAMI contains memes that are predominantly derogatory towards women, exemplifying typical subjects of online vitriol. Different from FHM and MAMI, where each meme was labeled as *harmful* or *harmless*, HarM was originally labeled with three classes: *very harmful*, *partially harmful*, and *harmless*. For a fair comparison, we merge the *very harmful* and *partially harmful* memes into the *harmful* class, following the setting of recent work (Pramanick et al., 2021b; Cao et al., 2022; Lin et al., 2023; Huang et al., 2024).

Baselines We compare MIND with state-of-the-art (SOTA) approaches for zero-shot harmful meme detection: 1) **GPT-4o** (Achiam et al., 2023); 2) **Gemini-1.5-Flash** (Team et al., 2024); 3) **LLaVA-1.5-7B** (Liu et al., 2024); 4) **InstructBLIP-7B** (Dai et al., 2023); 5) **MiniGPT-v2-7B** (Chen et al., 2023); 6) **OpenFlamingo-9B** (Awadalla et al., 2023); 7) **LLaVA-1.5-13B** (Liu et al., 2024); 8) **InstructBLIP-13B** (Dai et al., 2023); 9)**LLaVA-1.6-34B** (Liu et al., 2024); 10) **MIND (*)**: Our proposed multi-agent approach based on LLaVA-1.5-13B. We use the accuracy and macro-averaged F1 (dominant) scores as the evaluation metrics.

The data statistics, baseline descriptions and model implementation are detailed in the Appendix §A , §B, and §C, respectively.

4.2 Harmful Meme Detection Performance

Table 1 illustrates the performance of our proposed framework MIND versus all the compared baselines for zero-shot harmful meme detection. It is observed that: 1) In the first group of closed-source models, GPT-4o and Gemini-1.5-Flash demonstrate competitive performance across all datasets, with GPT-4o showing generally stronger results, particularly on the MAMI dataset. 2) The second group consists of open-source models with varying parameter sizes. Among them, LLaVA-1.6-34B achieves the best baseline performance in the open-source setting. This superior performance can be attributed to its larger parameter size, as we observe a general trend where models with larger parameters (*e.g.*, LLaVA-1.6-34B) outperform their smaller counterparts (*e.g.*, LLaVA-1.5-7B) in this challenging zero-shot task. 3) Our proposed MIND framework, built upon LLaVA-1.5-13B, demonstrates remarkable improvements. Compared to the base LLaVA-1.5-13B model, MIND improves the macro-averaged-F1 scores by 14.74%, 7.70%, and 13.32% on HarM, FHM, and MAMI respectively. Notably, on the HarM dataset, MIND's performance exceeds LLaVA-34B by 3.60% and surpasses the closed-source GPT-4o by 4.90%. Additionally, MIND achieves comparable performance with Gemini-1.5-Flash on FHM, despite being based on a much smaller open-source model. These results demonstrate the effectiveness of our approach even compared to powerful proprietary models, validating the strength of our proposed multi-agent framework in zero-shot harmful meme detection.

4.3 Improvements Across Model Scales

To further validate the effectiveness and generalization ability of our MIND framework, we apply it to various LMMs with different parameter sizes, including both open-source and closed-source models. As shown in Table 2, MIND consistently brings improvements across all models and datasets. For open-source models, we observe that: 1) The improvement is most significant on LLaVA-1.5-13B, with macro-averaged-F1 scores increasing by 14.74%, 7.70%, and 13.32% on HarM, FHM, and MAMI respectively; 2) Even for the stronger LLaVA-1.6-34B model, MIND still achieves notable gains of 4.53%, 4.87%, and 4.10% across the three datasets. Remarkably, MIND also enhances the performance of closed-source model Gemini-1.5-Flash, improving its performance by 2.74%,

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
MIND (LLaVA-1.5-13B)	68.93	65.19	60.80	60.71	68.90	68.84
w/o SSR	64.97	60.92	60.40	60.38	66.70	66.38
w/o RID	62.67	51.93	57.20	56.02	59.70	56.51
w/o $\text{RID}_{\text{forward}}$	64.97	63.46	60.20	59.81	66.60	66.60
w/o $\text{RID}_{\text{backward}}$	64.41	62.28	59.20	58.94	68.00	67.98
w/o IAI	63.28	60.97	59.00	58.53	68.10	68.10

Table 3: Ablation studies on our proposed framework.

3.60%, and 3.60% on the three datasets. Most notably, as shown in Tables 1 and 2, with our framework, LLaVA-1.5-7B approaches the performance of base LLaVA-1.5-13B, enhanced LLaVA-1.5-13B matches or surpasses base LLaVA-1.6-34B, and both enhanced LLaVA-1.6-34B and Gemini-1.5-Flash achieve competitive performance with GPT-4o, demonstrating the strong capability of our framework in boosting model performance regardless of model scales and architectures.

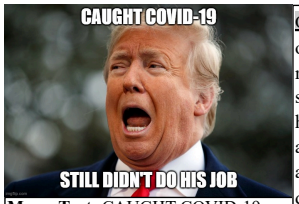
4.4 Ablation Study

To thoroughly evaluate the effectiveness of different strategies in our framework, we conduct ablation studies with several variants of MIND. As shown in Table 3, we examine five variants: 1) *w/o SSR*: replacing Similar Sample Retrieval (SSR) with random selection of three memes from S_{ref} as similar references while keeping other strategies unchanged; 2) *w/o RID*: removing Relevant Insight Derivation (RID), which consequently eliminates Insight-Augmented Inference (IAI) as it relies on derived insights, leaving similar memes to be directly used for reasoning; 3) *w/o* $\textit{RID}_{\text{forward}}$: removing forward insight derivation and its corresponding debate in IAI, where only backward insights are used for reasoning; 4) *w/o* $\textit{RID}_{\text{backward}}$: removing backward insight derivation and its corresponding debate in IAI, where only forward insights are used for reasoning; 5) *w/o IAI*: removing the multi-agent debate mechanism, where both forward and backward insights are directly used together for reasoning.

The ablation results reveal several interesting findings: 1) Random selection of similar memes (*w/o SSR*) leads to significant drops in macro-averaged-F1 scores (4.27%, 0.33%, and 2.46% on three datasets), suggesting that similarity-based retrieval is crucial for finding relevant reference memes; 2) Direct use of similar memes without insight derivation (*w/o RID*) causes the most substantial degradation in F1 scores (13.26%, 4.69%, and 12.33%), indicating that raw similar memes might introduce noise without proper insight derivation; 3) Using single-direction insight derivation (*w/o* $\textit{RID}_{\text{forward}}$ or *w/o* $\textit{RID}_{\text{backward}}$) results in moderate performance drops, with *w/o* $\textit{RID}_{\text{forward}}$ showing slightly better performance than *w/o* $\textit{RID}_{\text{backward}}$ on HarM and FHM datasets, while *w/o* $\textit{RID}_{\text{backward}}$ performs marginally better on MAMI; 4) Without the multi-agent debate mechanism (*w/o IAI*), the F1 scores decrease by 4.22%, 2.18%, and 0.74%, demonstrating that the multi-agent debate mechanism helps reconcile potentially conflicting insights. These results highlight that all strategies in our framework, Similar Sample Retrieval, Relevant Insight Derivation, and Insight-Augmented Inference, play essential and complementary roles in harmful meme detection. SSR retrieves similar memes as references, RID builds upon these retrieved memes to derive bidirectional insights through forward and backward reasoning, and IAI leverages the derived insights to generate robust judgments through multi-agent debate. The synergy among these strategies significantly enhances the framework's robustness in detecting harmful content, as evidenced by the performance degradation when any strategy is disabled.

4.5 Effect of Retrieved Meme Count

To investigate the impact of similar memes in our framework, we conduct experiments examining performance variations with different numbers of retrieved memes K as shown in Figure 4. We observe that: 1) All three datasets show relatively consistent performance trends, with peak macro-averaged-F1 scores generally occurring at lower K values before gradually declining. However, as the K value further increases, this score gradually decreases. This indicates that there exists an optimal number of similar memes that can be effectively utilized by the framework to enhance its discrimi-



Meme Text: CAUGHT COVID-19 STILL DIDN'T DO HIS JOB

Output Thought: The meme contains offensive language and content that may harm or offend individuals who support the president or believe that he handled the pandemic well. It is also disrespectful and may not be appropriate for a general audience due to its explicit nature.
Answer: Harmful.

(a)

Meme Text: enough is enough children are more important than freaks

Output Thought: The meme's message implies that people who identify as LGBTQ+ are "freaks" and are not as important as children. This kind of language is discriminatory and promotes prejudice towards the LGBTQ+ community. The meme's message does not promote equality and understanding among all groups and communities, as it implies that one group is superior to another.
Answer: Harmful.

(b)

Meme Text: I DON'T NEED A TANK TOP TO BE A WIFE BEATER

Output Thought: The meme's language and tone are offensive and derogatory towards women, as it implies that a man can be a "wife beater" without any provocation. The visual elements of the meme, such as the man's appearance and the accompanying text, reinforce negative stereotypes and perpetuate harmful attitudes towards women. The potential audience for this meme may include individuals who are susceptible to its message and could inadvertently contribute to the normalization of abusive behavior or perpetuate harmful attitudes towards women.
Answer: Harmful.

(c)

Figure 3: Examples of correctly predicted harmful memes in (a) HarM, (b) FHM, and (c) MAMI datasets.

native ability. 2) Larger K values don't necessarily lead to better results, as they tend to incorporate less similar memes into the retrieval results, potentially introducing noise rather than beneficial information. Based on these observations, setting K to 3 achieves the optimal balance between performance and efficiency across all datasets.

4.6 Case Study

To better understand how our proposed framework processes and evaluates memes, we analyze several correctly predicted cases, where we show important content in the thought, as shown in Figure 3.

From analyzing the output thought, we observe that: 1) Our proposed framework effectively connects multimodal information between meme text and imagery using commonsense knowledge. For example, in Figure 3(a), it recognizes how "caught COVID-19" and "didn't do his job" in the text directly relates to the presidential image, forming a critical commentary. In Figure 3(b), the framework links the protest imagery with the text's discriminatory comparison between "children" and "freaks", identifying its target toward the LGBTQ+ community shown in the visual elements. For Figure 3(c), it connects the aggressive text about being a "wife beater" with the threatening visual presentation to recognize harmful gender-based messaging. 2) Through deriving insights from similar memes, our proposed framework demonstrates sophisticated analysis of broader implications. In Figure 3(a), the reasoning considers how comparable political memes during the pandemic have contributed to increased social division and undermined public health messaging. The analysis in Figure 3(b) draws parallels with other discriminatory content to show how such messaging creates dangerous hierarchies between social groups and perpetuates harmful stereotypes. For Figure 3(c), by connecting to similar cases of normalized violence, the reasoning reveals how such content can desensitize viewers to domestic abuse and reinforce dangerous attitudes toward women. Through these examples, we can see how our framework provides clear, interpretable reasoning that connects visual and textual elements while considering real-world implications. This analytical transparency can be valuable for human checkers verifying model predictions in zero-shot setting. Additional case study and error analysis of our proposed framework are provided in Appendix §L and §M.

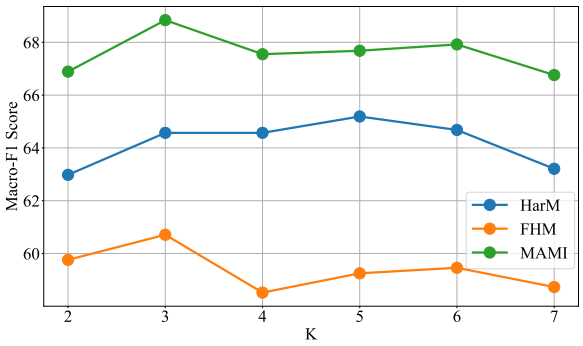


Figure 4: Effect of Top_K in Similar Sample Retrieval

5 Conclusion and Future Work

In this paper, we delved into the zero-shot harmful meme detection task that operates without annotated training data. To this end, we proposed MIND, a novel multi-agent framework that seamlessly integrates Similar Sample Retrieval, Relevant Insight Derivation, and Insight-Augmented Inference to enable robust harmful content detection. Through comprehensive experiments and analyses on three meme datasets, we demonstrated the effectiveness of our proposed framework and the

importance of each strategy. Future efforts aim to enhance our research by exploring the robustness of relevant insight derivation across diverse meme contexts.

Limitations

There are multiple ways to further improve this work:

- The framework's performance partially depends on the quality of retrieved similar memes. While our current similarity-based retrieval mechanism shows effectiveness, selecting truly relevant reference memes remains challenging. Current embedding-based similarity metrics may not fully capture the complex semantic relationships of highly contextual or novel meme formats, suggesting that more sophisticated retrieval strategies could enhance performance.

- Although our bi-directional insight derivation mechanism enables comprehensive analysis, its uniform treatment of all retrieved memes might not optimally capture their varying degrees of relevance. Implementing a more nuanced weighting mechanism that considers relative importance could lead to more refined insights.

- The current multi-agent debate mechanism primarily focuses on achieving consensus through reasoned arbitration. However, quantitatively evaluating the quality and reliability of derived insights remains challenging, particularly when correct conclusions stem from potentially flawed reasoning chains. This limitation makes it difficult to systematically assess the framework's overall robustness.

- While MIND eliminates substantial training and data annotation costs, its inference time can be significant due as it requires numerous LMM calls. This results in an approximate 8x computational overhead compared to simple zero-shot baselines, which is a key consideration for real-time deployment. Nevertheless, our ablation studies demonstrate that specific configurations, such as employing unidirectional RID, can halve these calls, offering a flexible efficiency-performance trade-off.

- Despite efforts to ensure robust decision-making through multi-agent debate, the framework's performance still heavily relies on the base LMM's understanding of social and cultural nuances. It may struggle to identify subtly encoded harmful content or evolving memes requiring deep cultural knowledge that human moderators readily recognize.

Ethics Statement

Our research aims to combat harmful meme content through zero-shot detection methods, contributing to safer online spaces. The harmful content types addressed in our study are well-documented concerns in social media research. Our work focuses on detecting various forms of harmful content including hate speech, misogyny, and misinformation that can negatively impact individuals and communities. However, we are aware of the potential for malicious users to reverse-engineer and create memes that go undetected or misunderstood by AI systems based on MIND. We strongly condemn such practices and emphasize that our research is intended solely for scientific purposes and harmful content prevention. The framework and associated resources are strictly prohibited from commercial use or malicious exploitation. To ensure responsible development and evaluation of our framework, we implemented several protective measures: 1) all experiments were conducted using publicly available research datasets, following their respective usage agreements; 2) no personal user data was collected or utilized in this study. We believe the benefits of advancing harmful meme detection capabilities outweigh the potential risks, particularly given the growing challenge of moderating harmful content on social media. The opinions and content contained in the meme samples should not be interpreted as representing the views of the authors. Our framework is designed to assist, not replace, human moderation efforts in maintaining healthy online communities.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 62376034 and 92467105).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31th ACM international conference on multimedia*.

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.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4575–4584.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

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

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Dong Huang, Qingwen Bu, Jie M Zhang, Michael Luck, and Heming Cui. 2023. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*.

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

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, pages 3868–3872.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, 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. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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*, pages 5138–5147.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 9459–9474.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

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

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

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

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023b. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

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. *arXiv preprint arXiv:2109.05184*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the evolution of (hateful) memes by means of multimodal contrastive learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 293–310. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235*.

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *arXiv preprint arXiv:2403.17927*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. In *Intrinsically-Motivated and Open-Ended Learning Workshop@ NeurIPS2023*.

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *arXiv preprint arXiv:2406.01014*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. 2022. Multimodal zero-shot hateful meme detection. In *14th ACM Web Science Conference 2022*, pages 382–389.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.

Datasets	Test	
	#harmful	#harmless
HarM	124	230
FHM	250	250
MAMI	500	500

Table 4: Statistics of test sets.

A Datasets

The detailed statistics for the original test splits of the three datasets are shown in Table 4.

B Baselines

We compare MIND with several state-of-the-art (SOTA) methods for zero-shot harmful meme detection: **GPT-4o** (Achiam et al., 2023): a proprietary large multimodal model by OpenAI that demonstrates strong zero-shot capabilities in visual-language tasks through in-context learning; **Gemini-1.5-Flash** (Team et al., 2024): Google's latest multimodal model that shows competitive performance in reasoning and visual understanding tasks; **LLaVA-1.5-7B** (Liu et al., 2024): a lightweight multimodal model built on Vicuna-7B, trained on diverse visual instruction data for general vision-language tasks; **InstructBLIP-7B** (Dai et al., 2023): an instruction-tuned vision-language model based on BLIP-2 architecture that leverages Vicuna-7B for language modeling; **MiniGPT-v2-7B** (Chen et al., 2023): a compact yet effective multimodal model that combines visual encoding with instruction-tuned language generation; **OpenFlamingo-9B** (Awadalla et al., 2023): an open-source implementation of Flamingo models that enables frozen language models to process visual inputs through cross-attention; **LLaVA-1.5-13B** (Liu et al., 2024): a medium-sized variant of LLaVA built on Vicuna-13B, with enhanced visual grounding and reasoning capabilities; **InstructBLIP-13B** (Dai et al., 2023): an enhanced version of InstructBLIP using Vicuna-13B as the language model backbone; **LLaVA-1.6-34B** (Liu et al., 2024): the latest and largest version of LLaVA with improved reasoning, OCR, and world knowledge capabilities; **MIND (*)**: our proposed multi-agent approach for zero-shot harmful meme detection, based on LLaVA-1.5-13B. We use the accuracy and macro-averaged F1 (dominant) scores as the evaluation metrics, where the macro-averaged F1 score is the more important metric owing to the imbalanced class prevalence, to capture competitive performance beyond the majority class.

For baseline methods, we implement the following Chain-of-Thought (Kojima et al., 2022) prompt structure, which achieves better performance than direct classification: "*Given the meme, with the Text: "$\{\mathcal{T}\}$" embedded in the image $\{\mathcal{V}\}$, your task is to determine whether this meme is harmful or not, in order to maintain the harmlessness and integrity of information on the Internet. Please leverage your extensive knowledge to deeply analyze and understand this meme, and give your final judgment. Your output should stricly follow the format: "Thought: [Your analysis] Answer: [harmful/harmless].*"

C Implementation Details

For our proposed framework, we conduct extensive experiments with various LMMs including LLaVA series (LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-1.6-34B), and Gemini-1.5-Flash as our LMM agents. For the baseline of the closed-source model, we use the "gpt-4o-2024-11-20" and "gemini-1.5-flash" versions, respectively. To ensure reproducibility and consistent comparison, we set the temperature parameter to 0 without any sampling mechanism for all models. For the multimodal embedding computation in Equation 1, we employ CLIP (Radford et al., 2021) with the specific version "ViT-L/14@336p" as our frozen pretrained vision and text Transformer encoders ($\mathbf{V_{enc}}$ and $\mathbf{T_{enc}}$). The modality fusion weights λ_v and λ_t are set to 0.8 and 0.2 respectively through grid search optimization within [0,1]. In our implementation, we set K to 3 for retrieving the most similar reference samples in Similar Sample Retrieval, which provides a balance between computational efficiency and sufficient context information. For the Relevant Insight Derivation process, we design the prompt template as Figure 10 to guide the $\text{LMM}_{\text{deriving}}$ agent.

In the Insight-Augmented Inference process, for the $\text{LMM}_{\text{debater}}$ agent, we implement the following prompt structure: "*Given the meme, with the Text: "$\{\mathcal{T}\}$" embedded in the image $\{\mathcal{V}\}$. Is it harmful or harmless? Note: [$\{note\}$] (Consider this note above with caution - it may contain biases or subjective interpretations that could influence judgment. Evaluate its reliability and relevance carefully) Your output should strictly follow the for-*

mat: "Thought: [First, analyze the meme's content independently. Then, carefully consider how the provided note may inform or bias your understanding. Weigh the note's credibility and relevance before incorporating it into your final assessment.] Answer: [harmful/harmless].""

When disagreement occurs, the judge agent utilizes this prompt template: "*Given the meme, with the Text: "*$\{\mathcal{T}\}$*" embedded in the image* $\{\mathcal{V}\}$*. Is it harmful or harmless? Regarding this meme, there are two debaters with differing opinions. Debater 1 believes that: The correct answer should be:* $\{debater1answer\}$ *Debater 1's reason: [*$\{debater1reason\}$*] Debater 2 believes that: The correct answer should be:* $\{debater2answer\}$ *Debater 2's Reason: [*$\{debater2reason\}$*] As a judge, you need to determine who is correct based on the viewpoints of the two debaters and provide the correct answer that you believe to be true. Your response should strictly adhere to this format: "Thought: [Based on the meme itself and your careful consideration, whose viewpoint do you think is correct? And why?] Answer: [Your final judgment(harmful/harmless)]."*

All experiments were conducted using four NVIDIA A40 48GiB GPUs. The total processing time varies by dataset size: approximately 1.5 hours for HarM, 3 hours for FHM, and 5 hours for MAMI when using Gemini-1.5-Flash. With LLaVA-1.5-13B, these times extend to 3 hours, 4.5 hours, and 9 hours respectively.

Due to privacy and ongoing research considerations, the code used in this study is not included in the submission. However, we commit to making the code publicly available upon the acceptance of this paper.

D Related work about LMMs

LLMs have recently expanded their capabilities beyond text processing to handle image inputs. Building upon these language models, LMMs have emerged as powerful tools for visual-language understanding. The current LMM landscape features both commercial and open-source solutions. Industry leaders like GPT-4o (Achiam et al., 2023) and Google's Gemini (Team et al., 2024) stand out for their strong zero-shot performance and sophisticated visual reasoning capabilities. These models excel at understanding nuanced visual content, engaging in detailed conversations about images, and providing rich, contextual analysis of visual information. In parallel with these commercial models, the open-source community has achieved significant breakthroughs. LLaVA (Liu et al., 2023a) and its latest version, LLaVA-1.6-34B (Liu et al., 2024), have made substantial progress in matching the capabilities of commercial models while keeping their technology transparent and accessible. These open models (Bai et al., 2023; Chen et al., 2024; Wang et al., 2023b; Chen et al., 2023; Dai et al., 2023; Awadalla et al., 2023) use clever training methods, including visual instruction tuning and streamlined fine-tuning approaches, to achieve strong results without requiring massive computational power. In this work, we utilize LLaVA-1.5-13B (Liu et al., 2024) as our primary backbone model for all LMM agents. To demonstrate the generalizability of our proposed framework, we also conduct experiments with LLaVA-1.6-34B and Gemini-1.5-Flash, which represent the current state-of-the-art in open-source and commercial LMMs respectively.

Algorithm 1 MIND - Similar Sample Retrieval

Initialize:
Modality fusion weights λ_v, λ_t;
Visual Encoder $\mathbf{V_{enc}}(\cdot)$;
Textual Encoder $\mathbf{T_{enc}}(\cdot)$;
Reference set S_{ref}, Test set S_{test};
Target meme M_{target};
Number of similar samples K;
Embedding set $\mathbf{E}_{\text{all}} \leftarrow \emptyset$;
Similarity scores $S \leftarrow \emptyset$;
Similar memes set $M_{\text{similar}} \leftarrow \emptyset$;
Embedding Generation:
for each meme $M \in \{S_{\text{ref}} \cup S_{\text{test}}\}$ **do**
  $\mathbf{V} \leftarrow \mathbf{V_{enc}}(\mathcal{V})$
  $\mathbf{T} \leftarrow \mathbf{T_{enc}}(\mathcal{T})$
  $\mathbf{E} \leftarrow \lambda_v \cdot \mathbf{V} + \lambda_t \cdot \mathbf{T}$
  $\mathbf{E}_{\text{all}} \leftarrow \mathbf{E}_{\text{all}} \cup \{\mathbf{E}\}$
end for
Similar Sample Selection:
for each $M_{\text{ref}} \in S_{\text{ref}}$ **do**
  $s \leftarrow \text{cosine}(\mathbf{E}_{\text{target}}, \mathbf{E}_{\text{ref}})$
  $S \leftarrow S \cup \{s\}$
end for
$M_{\text{similar}} \leftarrow \{M_{\text{ref}} \mid s \in \text{Top}_K(S)\}$
return M_{similar}

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
LLaVA-1.5-13B	62.28	50.45	55.20	53.01	60.10	55.52
zero-shot w/ SSR	62.67	51.93	57.20	56.02	59.70	56.51
zero-shot w/o SSR	59.04	47.72	53.00	52.19	56.20	50.73
3-shot w/ SSR	66.10	59.60	60.60	60.36	66.70	66.66
3-shot w/o SSR	60.45	56.44	56.60	58.42	61.00	60.92
w/ MIND	68.93	65.19	60.80	60.71	68.90	68.84

Table 5: Evaluation results comparing with few-shot methods.

Algorithm 2 MIND - Relevant Insight Derivation

Initialize:
Similar memes M_{similar} from Algorithm 1;
Chain-of-Thought deriving prompt $\mathcal{P}_{\text{deriving}}$;
Deriving agent $\text{LMM}_{\text{deriving}}$;
Forward insight set $\mathcal{I}_{\text{fwd},0} \leftarrow \emptyset$;
Backward insight set $\mathcal{I}_{\text{back},0} \leftarrow \emptyset$;
for $i = 1$ to K **do**
  $\mathcal{I}_{\text{fwd},i} \leftarrow \text{LMM}_{\text{deriving}}(M_{\text{similar},i}, \mathcal{I}_{\text{fwd},i-1}, \mathcal{P}_{\text{deriving}})$
end for
for $i = 1$ to K **do**
  $\mathcal{I}_{\text{back},i} \leftarrow \text{LMM}_{\text{deriving}}(M_{\text{similar},K+1-i}, \mathcal{I}_{\text{back},i-1}, \mathcal{P}_{\text{deriving}})$
end for
return $\mathcal{I}_{\text{fwd},K}, \mathcal{I}_{\text{back},K}$

E MIND Algorithm

Algorithms 1 to 3 detail the multi-agent framework of our approach, outlining the Similar Sample Retrieval, Relevant Insight Derivation, and Insight-Augmented Inference stages, respectively.

While Algorithm 2 provides a formal depiction of the Relevant Insight Derivation process, its iterative and cumulative nature, which is central to leveraging unlabeled reference data effectively, warrants further conceptual explanation. We provide a detailed breakdown of how $\text{LMM}_{\text{deriving}}$ systematically processes memes in two complementary passes to accumulate insights.

Given a set of K similar memes $M_{\text{similar}} = \{M_1, M_2, \ldots, M_K\}$, the process unfolds as follows:

Forward Pass $\text{LMM}_{\text{deriving}}$ processes memes sequentially from M_1 to M_K. In each iteration i, the current meme $M_{\text{similar},i}$ is fed into the $\text{LMM}_{\text{deriving}}$ along with the cumulative insights from previously processed memes. This generates $\mathcal{I}_{\text{fwd},i}$, ensuring $\mathcal{I}_{\text{fwd},K}$ captures insights cumulatively.

For example, when $K = 3$ with similar memes $\{M_1, M_2, M_3\}$, the forward pass progresses as follows:

$$\mathcal{I}_{\text{fwd},1} = \text{LMM}_{\text{deriving}}(M_1, \emptyset, \mathcal{P}_{\text{deriving}})$$
$$\mathcal{I}_{\text{fwd},2} = \text{LMM}_{\text{deriving}}(M_2, \mathcal{I}_{\text{fwd},1}, \mathcal{P}_{\text{deriving}})$$
$$\mathcal{I}_{\text{fwd},3} = \text{LMM}_{\text{deriving}}(M_3, \mathcal{I}_{\text{fwd},2}, \mathcal{P}_{\text{deriving}})$$

The final forward insight set $\mathcal{I}_{\text{fwd},\mathbf{3}}$ is then passed to the Insight-Augmented Inference stage.

Backward Pass To counter potential biases from a single processing order, $\text{LMM}_{\text{deriving}}$ also processes memes in reverse, from M_K down to M_1. In iteration i, $M_{\text{similar},K+1-i}$ is processed by $\text{LMM}_{\text{deriving}}$ alongside insights accumulated from memes already processed in this reverse sequence. This dual-directional approach ensures a comprehensive and robust set of insights.

For example, when $K = 3$ with similar memes $\{M_1, M_2, M_3\}$, the backward pass progresses as follows:

$$\mathcal{I}_{\text{back},1} = \text{LMM}_{\text{deriving}}(M_3, \emptyset, \mathcal{P}_{\text{deriving}})$$
$$\mathcal{I}_{\text{back},2} = \text{LMM}_{\text{deriving}}(M_2, \mathcal{I}_{\text{back},1}, \mathcal{P}_{\text{deriving}})$$
$$\mathcal{I}_{\text{back},3} = \text{LMM}_{\text{deriving}}(M_1, \mathcal{I}_{\text{back},2}, \mathcal{P}_{\text{deriving}})$$

Algorithm 3 MIND - Insight-Augmented Inference

Initialize:
Target meme visual $\mathcal{V}_{\text{target}}$, text $\mathcal{T}_{\text{target}}$;
Forward and backward insight set $\mathcal{I}_{\text{fwd},K}$, $\mathcal{I}_{\text{back},K}$ from Algorithm 2;
Debater agents $\text{LMM}_{\text{debater}}$;
Judge agent $\text{LMM}_{\text{judge}}$;
$\mathcal{J}_{\text{fwd}} \leftarrow \text{LMM}_{\text{debater}}(\mathcal{I}_{\text{fwd},K}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}})$
$\mathcal{J}_{\text{back}} \leftarrow \text{LMM}_{\text{debater}}(\mathcal{I}_{\text{back},K}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}})$
if $\mathcal{J}_{\text{fwd}} = \mathcal{J}_{\text{back}}$ **then**
  $\mathcal{J}_{\text{final}} \leftarrow \mathcal{J}_{\text{fwd}}$
else
  $\mathcal{J}_{\text{final}} \leftarrow \text{LMM}_{\text{judge}}(\mathcal{J}_{\text{fwd}}, \mathcal{J}_{\text{back}}, \mathcal{V}_{\text{target}}, \mathcal{T}_{\text{target}})$
end if
return $\mathcal{J}_{\text{final}}$

Similarly, the final backward insight set $\mathcal{I}_{\text{back},\mathbf{3}}$ is also utilized in the Insight-Augmented Inference stage.

More detailed examples of derived insight sets can be found in Appendix N.

F Discussion about LMM Selection

In our experimental design, we prioritize research reproducibility and transparency in model selection. We primarily adopt LLaVA as our backbone LMM because its training process and data sources are fully transparent, which ensures experimental fairness and reproducibility. Specifically, we use LLaVA-1.5-13B as our main model for its balanced performance and efficiency, while also conducting experiments on LLaVA-1.5-7B and LLaVA-1.6-34B to validate our framework's effectiveness across different model scales. To demonstrate the generalizability of our framework, we also evaluate it using the closed-source model Gemini-1.5-Flash. However, we note that experiments with closed-source models may face two limitations: 1) potential data leakage cannot be completely ruled out due to the undisclosed nature of their training data, and 2) full reproducibility cannot be guaranteed despite setting temperature to 0, as these models may undergo undisclosed updates. Therefore, while we report results on closed-source models for completeness, our main analysis and conclusions are primarily drawn from experiments with open-source models.

G MIND Versus Few-shot

To better understand our proposed framework's effectiveness, we conduct a comprehensive comparison between MIND and few-shot in-context learning approaches. Table 5 presents detailed experimental results across three datasets using LLaVA-1.5-13B as the backbone model.

Several interesting observations emerge from this comparison. First, the introduction of SSR consistently improves performance in both zero-shot and few-shot settings. For instance, on the HarM dataset, adding SSR increases the zero-shot macro-averaged-F1 score from 47.42% to 51.93%, and similarly enhances few-shot performance from 56.44% to 59.60%. This demonstrates SSR's fundamental value in providing relevant contextual information, regardless of the learning paradigm. More significantly, our complete MIND framework achieves superior performance compared to both zero-shot and few-shot variants. Taking the HarM dataset as an example, MIND achieves a macro-averaged-F1 score of 65.19%, substantially outperforming the few-shot with SSR strategy. Similar results are observed across FHM and MAMI datasets, where MIND consistently demonstrates better performance.

The framework's strength comes not just from retrieving similar samples, but from the sophisticated processing of these samples through the RID and IAI strategies. The bidirectional insight derivation and multi-agent debate mechanism appear to capture deeper understanding than what is possible through few-shot learning alone. The practical implications of these results are significant. While few-shot learning requires annotated examples that may need regular updating as meme patterns change, MIND achieves superior performance without requiring any labeled data. This zero-shot capability eliminates the need for maintaining example sets and makes MIND especially valuable in real-world scenarios where obtaining high-quality annotated data is challenging or impractical. Moreover, while few-shot performance might be limited by the quality and representativeness of the few labeled examples available, MIND's zero-shot nature allows it to adapt more flexibly to diverse and evolving harmful content patterns.

H Comparison with Training-based Methods

While MIND operates as a training-free framework, fundamentally differing from traditional data-driven classification methods, we provide a comparison with established training-based approaches for reference. This comparison highlights the distinct advantages of our approach, particularly in scenarios with scarce or evolving harmful meme data. Table 6 presents the performance of MIND alongside two prominent prior works, Late Fusion (Pramanick et al., 2021a) and MOMENTA (Pramanick et al., 2021b), across the HarM, FHM, and MAMI datasets.

As observed from Table 6, traditional data-driven methods, while achieving strong performance on the HarM dataset (which was typically part of their training data), exhibit a noticeable drop in performance on the FHM and MAMI datasets. These datasets represent distributions that were not included in their training, highlighting a funda-

Dataset	HarM		FHM		MAMI	
Method	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
Late Fusion (Pramanick et al., 2021a)	73.24	70.25	59.14	44.81	63.20	59.76
MOMENTA (Pramanick et al., 2021b)	83.82	82.80	61.34	57.45	72.10	66.93
MIND (LLaVA-1.5-13B)	68.93	65.19	60.80	60.71	68.90	68.84

Table 6: Performance comparison between MIND and training-based methods.

mental limitation of supervised approaches: their effectiveness diminishes when encountering new, unseen meme distributions. This is a common challenge in the dynamic and rapidly evolving landscape of online memes. In contrast, MIND demonstrates remarkable generalization ability, maintaining consistent and competitive performance across all three datasets without any task-specific training. For instance, on FHM and MAMI, MIND (LLaVA-1.5-13B) achieves Macro-F_1 scores of 60.71% and 68.84% respectively, which are comparable to or even surpass those of MOMENTA on these out-of-distribution datasets. This stability in performance across diverse datasets underscores MIND's significant advantage in handling the inherently dynamic nature of harmful memes, where continuous collection and annotation of training data for every new meme trend is impractical. The training-free nature of MIND thus offers a robust and adaptable solution for real-world harmful meme detection.

I Discussion about SSR

In our design of Similar Sample Retrieval, several interesting observations emerge regarding its effectiveness in different settings. Unlike traditional approaches that rely heavily on annotated data, SSR demonstrates unique advantages in both zero-shot and few-shot scenarios, as evidenced by our experimental results in Table 5. In the zero-shot setting, SSR significantly improves model performance across all datasets. For instance, on the HarM dataset, introducing SSR increases the macro-averaged-F1 score from 47.72% to 51.93%. This improvement demonstrates that retrieving similar memes helps establish a more comprehensive framework for assessing harmfulness, even without any annotation guidance. More interestingly, SSR's benefits extend to few-shot scenarios, where we observe even more substantial gains. When combined with few-shot learning, SSR boosts the macro-averaged-F1 score from 56.44% to 59.60%. Similar improvements are observed across FHM and MAMI datasets. This suggests that SSR not only provides valuable context in zero-shot settings but also enhances the model's ability to leverage limited labeled data effectively. The consistent performance improvement across different settings highlights SSR's robustness as a fundamental strategy for harmful meme detection. Particularly noteworthy is how SSR helps bridge the gap between zero-shot and few-shot performance, suggesting that retrieved similar samples serve as an effective supplement to limited labeled data. This makes SSR especially valuable in real-world scenarios where obtaining large-scale annotated datasets is challenging or impractical.

J Discussion about RID

In our design of Relevant Insight Derivation, we observe several notable characteristics through empirical analysis. Unlike traditional sequential processing approaches, RID's bidirectional insight derivation mechanism demonstrates unique advantages in harmful meme detection. Through both forward and backward passes, RID effectively addresses the inherent imbalance in sequential processing. Specifically, in the forward pass, early memes in the sequence benefit from repeated refinement of insights, while later memes might receive less attention. The backward pass compensates for this imbalance by approaching the sequence from the opposite direction. This design ensures that insights from all retrieved similar memes contribute equally to the final analysis, regardless of their position in the sequence. The effectiveness of this bidirectional approach is particularly evident in the experimental results. For instance, with RID, the model achieves substantial improvements in harmful content detection compared to models using only unidirectional processing. This suggests that the synthesis of insights from both directions enables a more comprehensive understanding of potential harm. The bidirectional nature of RID also helps mitigate potential biases that might arise from processing memes in a single fixed order. Moreover, RID's iterative refinement process proves especially valuable when dealing with complex or ambiguous cases. In situations where harmful con-

tent is subtly embedded or masked by humor, the repeated processing and refinement of insights helps uncover less obvious harmful elements that might be missed in a single pass. This makes RID particularly robust in handling the diverse and evolving nature of harmful memes in real-world scenarios.

K Discussion about IAI

In our design of Insight-Augmented Inference, the multi-agent debate mechanism presents several interesting characteristics in harmful meme detection. Unlike traditional single-agent approaches that might be prone to biases or incomplete reasoning, IAI's debater-judge framework demonstrates unique advantages in achieving more balanced and reliable decisions. The two-debater design with complementary perspectives proves particularly effective. By leveraging insights from both forward and backward passes, each debater develops a potentially different viewpoint on the meme's harmfulness. This diversity in perspectives is crucial for complex cases where harmfulness might not be immediately apparent. For example, when analyzing memes that appear humorous on the surface but contain subtle discriminatory elements, the contrasting viewpoints of the debaters help surface these nuanced harmful aspects. More interestingly, the introduction of a judge agent for resolving disagreements adds an additional layer of robustness. Rather than simply averaging opinions, the judge agent actively examines the reasoning process of both debaters. This meta-level analysis helps filter out weak arguments and synthesize stronger ones, leading to more reliable final decisions. The effectiveness of this approach is particularly evident in cases where the two debaters reach different conclusions, demonstrating how the judge agent can effectively reconcile conflicting viewpoints.

L Case study of SSR, RID And IAI

We provide a case study of MIND's output thoughts, to investigate the effect of the SSR, RID and IAI strategies on the correctly predicted harmful meme samples, as illustrated in Figure 5 From the output thought in natural text, we observe that: 1) The Similar Sample Retrieval (SSR) mechanism enhances the framework's contextual understanding even without labeled data. For example, in Figure 5(a), without SSR, the framework simply interpreted it as "a satirical representation of a statement made by the US President". However, with

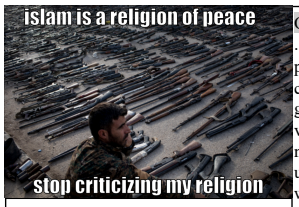
Figure 5: Examples of correctly predicted harmful memes in (a) HarM, (b) MAMI, and (c) FHM datasets.

SSR, the framework recognized broader implications, noting that while the meme may be "intended to be humorous", it's crucial to "consider the potential impact on public perception". This demonstrates how retrieving similar memes helps establish a more comprehensive framework for assessing harmfulness beyond just the immediate content, even without annotation. The framework can leverage patterns and contextual similarities across memes to develop a more nuanced understanding of potential harm. 2) The Relevant Insight Derivation (RID) mechanism significantly improves harm detection through bidirectional analysis. As shown in Figure 5(b), without RID, the framework superficially viewed the content as "a playful and humorous take on a violent situation". In contrast, with RID, the framework identified that the meme "promotes violence against women" and recognized its "harmful and inappropriate behavior". This illustrates how RID's forward and backward insight derivation helps uncover harmful content masked as humor. 3) The Insight-Augmented Inference

Meme Text: i took a shower this morning

Output Thought:
...The meme is offensive and disrespectful, as it makes light of a historical figure who is responsible for the deaths of millions of people. It is not appropriate to trivialize or mock such events or figures, as it can be disrespectful and inappropriate for some audiences. The meme's content is not respectful or considerate of all individuals and groups, and it should be removed from circulation....

(a)



Meme Text: islam is a religion of peace stop criticizing my religion

Output Thought:
...The meme's intent appears to be to promote understanding and peaceful coexistence between different religious groups, and it does not directly incite violence or aggression. The provided note offers additional context and understanding of the meme's intent, which is to challenge stereotypes and promote a more nuanced view of Islam as a religion of peace....

(b)

Figure 6: Examples of wrongly predicted memes by our proposed framework with the ground truth (a) harmless and (b) harmful.

(IAI) mechanism enables more nuanced judgment through multi-agent debate. In Figure 5(c), without IAI, the framework simply categorized it as "a lighthearted and humorous take on feminism". However, with IAI, the framework detected that the meme presents a "misleading and inaccurate representation of the feminist movement and its goals". This shows how IAI's debater-judge framework helps identify subtle forms of discrimination and stereotyping. These three modules work together to create a robust harmful meme detection system. While SSR provides the necessary context through zero-shot similar meme retrieval, RID ensures thorough analysis through bidirectional processing, and IAI guarantees balanced final decisions through multi-agent reasoning. This comprehensive approach enables the framework to effectively identify harmful content across various forms, from public health misinformation to gender discrimination, while maintaining sensitivity to context and nuance.

M Error Analysis

To better understand the behavior of our framework and facilitate future studies, we conduct an error analysis on the wrongly predicted memes. Figure 6 shows two examples of memes incorrectly classified by our framework. In Figure 6(a), which contains the text "I took a shower this morning", our framework incorrectly categorized it as harmful. The output thought suggests that "the meme is offensive and disrespectful, as it makes light of a historical figure who is responsible for the deaths of millions of people." This misjudgment stems from the framework's over-emphasis on the historical implications of the image while failing to properly integrate it with the innocuous shower-related text. The framework exhibited heightened sensitivity to potentially controversial historical content, leading to overly cautious classification. On the other hand, the harmful meme in Figure 6(b), featuring the text "islam is a religion of peace stop criticizing my religion" alongside an image of weapons, was incorrectly classified as harmless. The framework's output thought suggests that "the meme's intent appears to be to promote understanding and peaceful coexistence." This error reveals the framework's failure to recognize the ironic juxtaposition between the peaceful message and the threatening imagery. The framework appears to prioritize the literal meaning of the text while underweighting the visual implications, possibly due to an overly cautious approach to content involving religious themes. Through broader error analysis, we identified several common patterns in misclassification. The framework sometimes struggles with highly ironic or satirical content where the harmful intent is masked by seemingly positive messages. Additionally, in some cases, the framework may overemphasize either the visual or textual component, leading to incomplete context understanding. Complex cultural or historical references can also lead to either overly cautious or overly permissive classifications. These findings suggest potential directions for future improvement, particularly in developing better mechanisms for integrating multimodal information and handling culturally sensitive content.

N Results of Relevant Insight Derivation related to similar memes.

Figure 7 presents similar memes retrieved for a target meme on HarM dataset.

The relevant insight set of this target meme derived on HarM dataset through Forward Derivation are presented as follows.

Forward Insight Set on HarM Dataset.

- Memes should not trivialize or mock serious issues, such as public health crises, as they can cause distress and harm to individuals and communities.
- Memes should not promote discrimination,

harassment, or hate speech, as they can contribute to a toxic online environment and harm individuals.

- Memes should not spread misinformation or false information, as they can perpetuate harmful beliefs and mislead people.

- Memes should respect privacy and personal boundaries, as they can invade people's personal space and cause distress.

- Memes should be respectful and considerate of diverse cultures, beliefs, and experiences, as they can perpetuate stereotypes and offend people.

The relevant insight set of this target meme derived on HarM dataset through Backward Derivation are presented as follows.

Backward Insight Set on HarM Dataset

- Memes that make light of serious situations, such as the COVID-19 pandemic or any other significant health, environmental, or social issue, should be assessed for potential harmfulness and removed from platforms if they are deemed to be harmful or misleading.

- Memes that promote misinformation or false information should be flagged and reported.

- Memes that encourage discrimination, harassment, or violence should be removed and reported.

- Memes that violate copyright or intellectual property laws should be taken down and reported.

- Memes that are shared in a private or closed group should be evaluated based on the context and the audience, as they may have a different impact than if they were shared in a public or open forum.

Figure 8 presents similar memes retrieved for a target meme on FHM dataset.

The relevant insight set of this target meme derived on FHM dataset through Forward Derivation are presented as follows.

Forward Insight Set on FHM Dataset

- Memes containing offensive and derogatory content that has the potential to cause harm, including emotional distress, discrimination, or incitement of violence, may be harmful.

- The impact of such content on individuals and communities, as well as the context of the meme, should be evaluated when assessing the harmfulness of a meme.

- Memes that make light of or promote sensitive and controversial topics, such as terrorism or violence, should be evaluated with additional scrutiny for their potential to cause harm, considering the context in which the meme is shared and the potential for it to be misinterpreted or amplified.

- Memes that have the potential to cause harm to vulnerable individuals or groups, such as those experiencing emotional distress or those who may be more susceptible to harm due to their personal experiences or circumstances, should be evaluated with additional scrutiny for their potential to cause harm.

- Memes that trivialize or promote sensitive and controversial topics should be evaluated with additional scrutiny for their potential to cause harm, considering the context in which the meme is shared and the potential for it to be misinterpreted or amplified.

The relevant insight set of this target meme derived on FHM dataset through Backward Derivation are presented as follows.

Backward Insight Set on FHM Dataset

- Consider the potential emotional impact of the meme on the audience, including the potential for distress or trauma, especially in the context of religious or cultural identity.

- valuate the meme's potential to perpetuate harmful stereotypes or discrimination, including those related to religious or cultural groups.

- Assess the meme's potential to normalize or glorify harmful actions or behaviors, with a focus on the impact on specific groups' identity or self-esteem.

- Consider the potential for harm in the context of religious or cultural identity, as the related meme specifically references a religious group.
- Clarify the existing rule on emotional impact to include the potential for harm to a specific group's identity or self-esteem.

Figure 9 presents similar memes retrieved for a target meme on MAMI dataset.

The relevant insight set of this target meme derived on MAMI dataset through Forward Derivation are presented as follows.

Forward Insight Set on MAMI Dataset

- Avoid making assumptions about a person's gender based on their driving abilities.
- Ensure that any humor or commentary is not offensive or discriminatory towards a specific group of people, including but not limited to gender, race, or any other characteristic.
- Be mindful of the potential harm that memes can cause, and consider the impact on the targeted group, especially when perpetuating harmful stereotypes or generalizations.
- Refrain from spreading memes that promote negative stereotypes or harmful generalizations.
- Encourage a culture of respect and empathy towards all individuals, regardless of their gender, race, or any other characteristic.

The relevant insight set of this target meme derived on MAMI dataset through Backward Derivation are presented as follows.

Backward Insight Set on MAMI Dataset

- Avoid stereotyping or generalizing about women based on their physical appearance or strength.
- Avoid sharing memes that stereotype or generalize about women drivers, as it perpetuates harmful stereotypes and contributes to a culture of sexism and discrimination.
- Be mindful of the context in which the meme is shared and ensure it is appropriate for the audience.
- Consider the potential impact of the meme on individuals and communities, and avoid sharing memes that may cause harm or offense.
- Promote positive and inclusive representations of women in all forms of media, including memes.

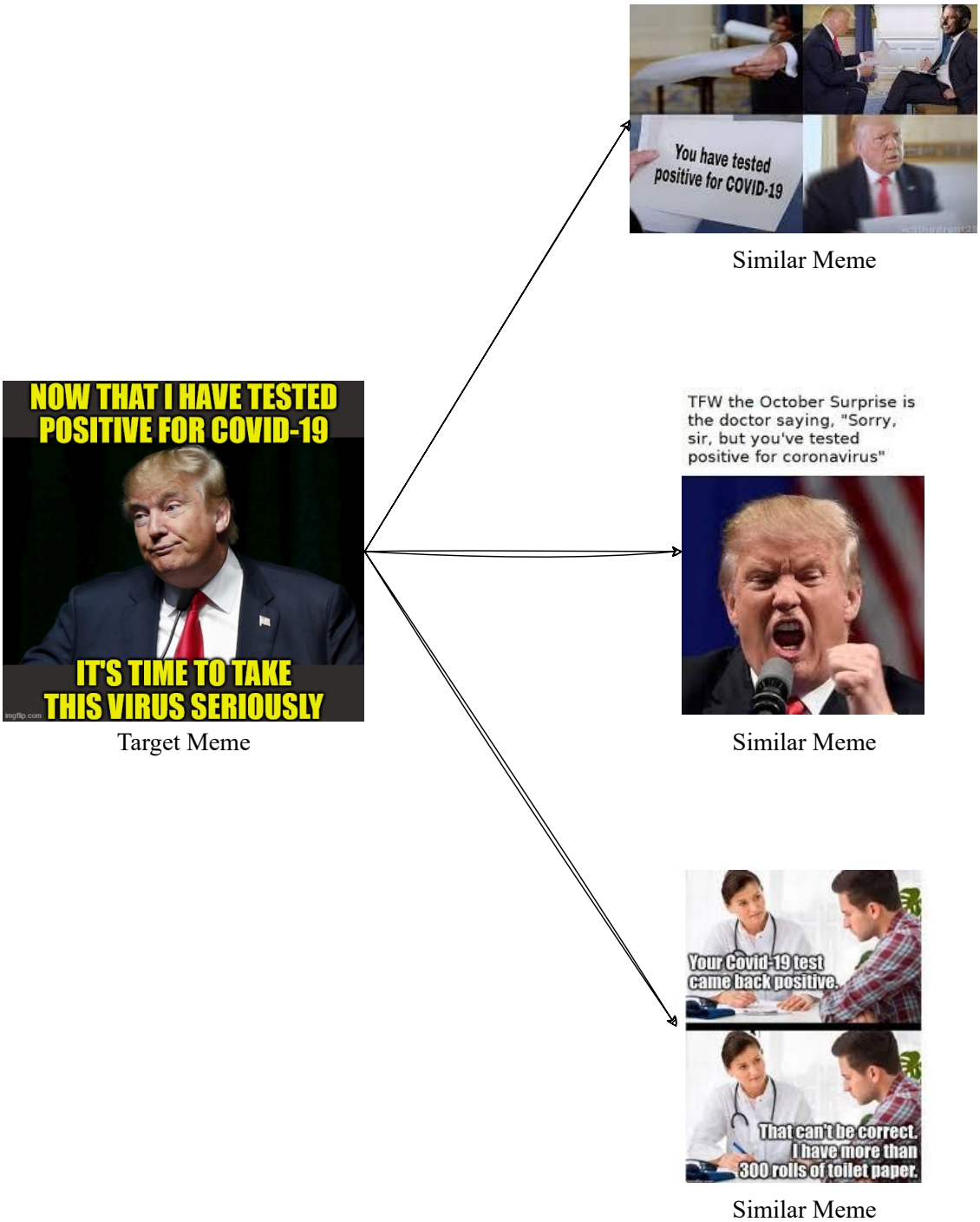


Figure 7: Example of the target meme along with its similar memes from the HarM dataset.

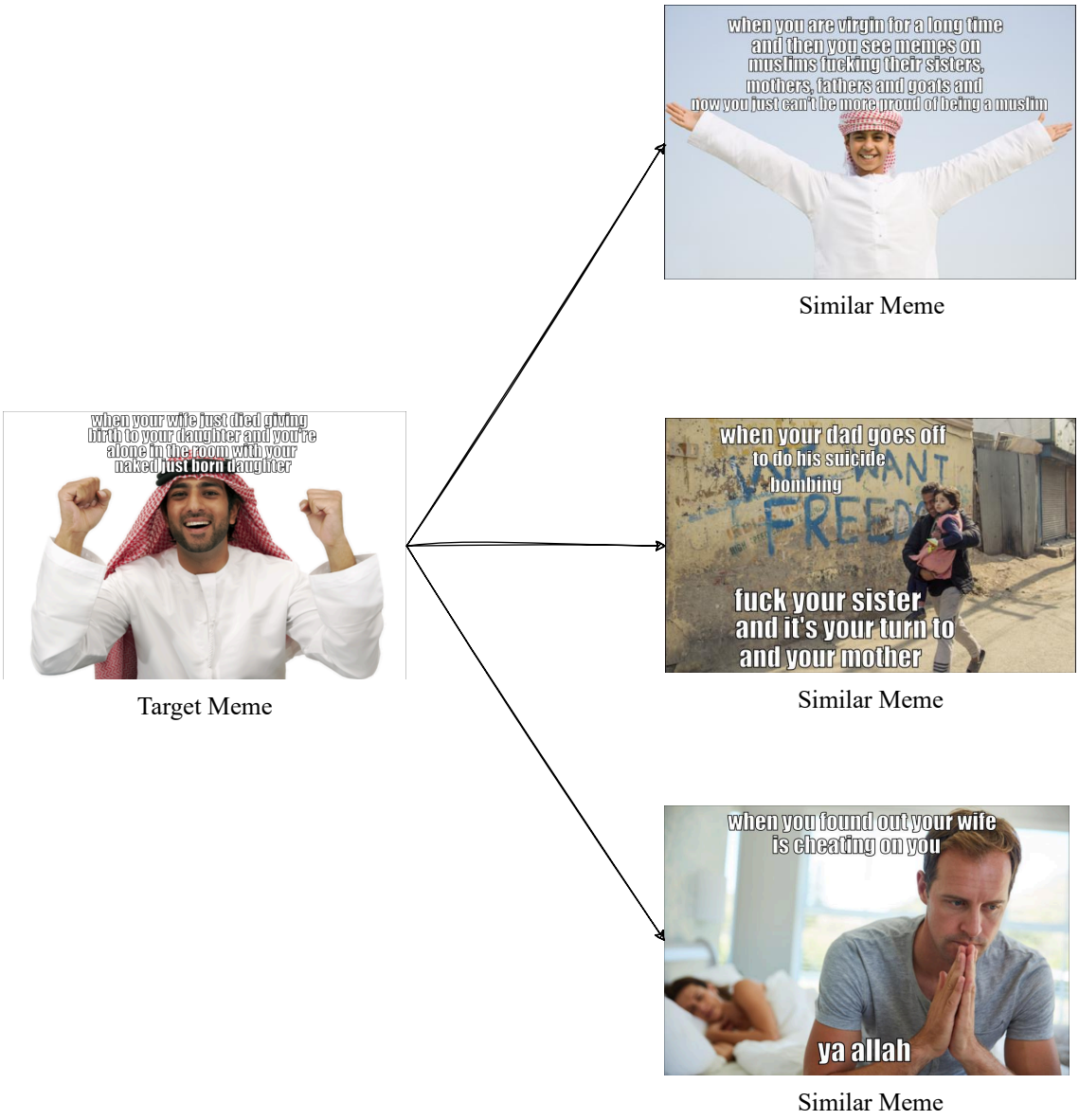


Figure 8: Example of the target meme along with its similar memes from the FHM dataset.

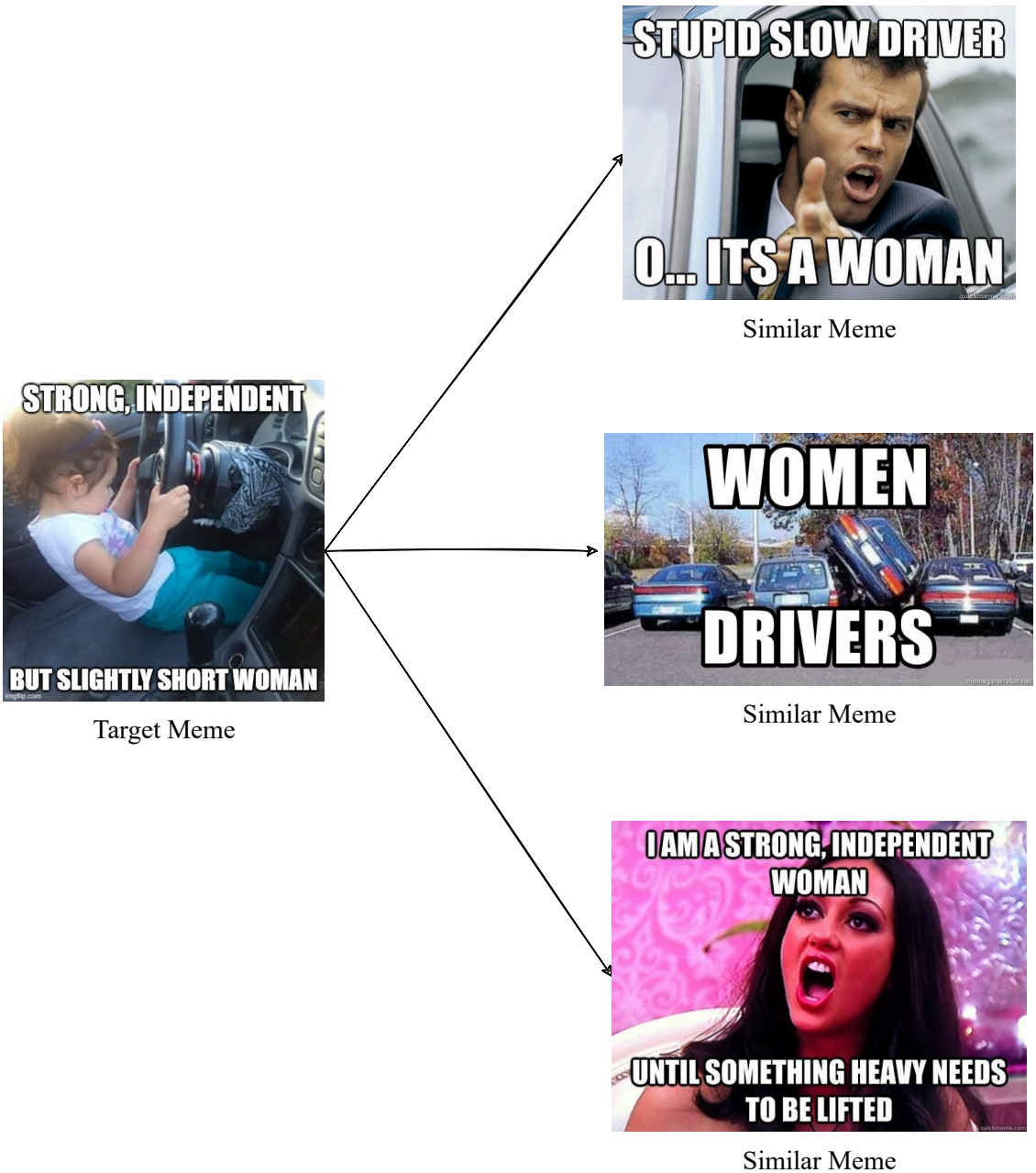


Figure 9: Example of the target meme along with its similar memes from the MAMI dataset.

You are a sophisticated analytical agent tasked with formulating and refining insightful recommendations. Your ultimate objective is to assess whether a meme is harmful. However, the target meme will not be provided directly. Instead, you will receive a meme related to the target meme. Based on this related meme, you are to synthesize advanced insights to determine whether such memes are harmful, reflect on existing insights, and update these reflections into the current rules. Rules you update must be GENERAL and HIGH LEVEL.

Here is the related meme:
[Meme]
Image: "{$\mathcal{V}$}"
Text embedded: "{$\mathcal{T}$}"
[Meme]

Here are the EXISTING RULES:
[EXISTING RULES START]
{rules}
[EXISTING RULES END]

If the existing rules are empty, it indicates that you are the first person to analyse on this related m
You can directly build upon the existing rules by adding new ones, editing them if there is room for enhancement, and removing those that are contradictory or duplicated.
You should focus on distilling insights from the related meme to assess similar memes, rather than determining the harmfulness of the meme currently presented to you.
Each rule should be concise and easy to follow. Make sure there are a maximum of five rules.

Your output should strictly follow the format:
"Thought: [Reflect on whether the current rules are helpful in assessing the harmfulness of this related meme.]
Operations of updating existing rules: [Update your reflections on this meme into the existing rules. You can choose to add, edit, and remove or retain existing rules. You may perform up to two operations. Please specify in detail which rule you want to operate on and how, but do not mention any operations in the rules themselves; your operations regarding the rules should only be mentioned here.]
Updated rules: [Do not mention the operations you have made and the related meme in the rules, directly output all the updated rules, only output the rules that have been determined, and do not leave spaces for rules that have not yet been generated. Rules you update must be GENERAL and HIGH LEVEL.]"

Figure 10: The prompt of $\mathcal{P}_{\text{deriving}}$