DR-HM: Distill-then-Reinforce Training with Cognition-Aware Data Synthesis for Harmful Meme Detection

Zihan Cheng[*,1] **Jianxiang Ma**[*,1] **Xiaocui Yang**[†1] **Peidong Wang**[1]
Wen Zhang[1] **Shi Feng**[†1] **Daling Wang**[1] **Yifei Zhang**[1] **Mingfu Zhang**[2]
[1]School of Computer Science and Engineering,
Northeastern University, Shenyang 110819, China
[2]OranAI Ltd., California 91748, USA
{chengzh1, majx6, zhangw6}@mails.neu.edu.cn
{yangxiaocui, wangdaling, fengshi, zhangyifei}@cse.neu.edu.cn
pdongwang@163.com, mingfu.zhang@mail.polimi.it

Abstract

Harmful memes convey offensive intent through implicit associations between visual symbols and text, requiring a broad understanding of cultural stereotypes and visual metaphors. Small-scale Multimodal Large Language Models (MLLMs) often lack the knowledge required to identify such implicit hate, whereas Large-scale MLLMs, despite their broader knowledge, exhibit systematic labeling bias. To address these challenges, we propose **DR-HM**, a Distill-then-Reinforce training framework with cognition-aware data synthesis for harmful meme detection, which aims to transfer knowledge from closed-source models while mitigating their biases. DR-HM introduces a six-step structured data synthesis scheme with self-refinement that decomposes meme analysis into a progressive, human-inspired reasoning process from entity recognition to harmfulness judgment. Based on the synthesized reasoning data, we further adopt a Distill-then-Reinforce training strategy. This approach combines a two-stage Supervised Fine-Tuning (SFT) with an Adaptive Group Relative Policy Optimization (A-GRPO) algorithm, which incorporates class-ratio-aware reward weighting and dynamic KL coefficients. Experiments[1] on three benchmark datasets show that the proposed approach consistently outperforms existing methods and achieves an accuracy of 84.7% on the FHM dataset, approaching the reported performance of human annotators.

1 Introduction

Multimodal memes have become a dominant medium for online expression. Beyond their humorous surface, they frequently encode implicit attacks by deliberately pairing images and text to target specific groups (Kiela et al., 2020b; Fersini et al., 2022; Pramanick et al., 2021b). Such offensive intent relies on cultural stereotypes, visual metaphors, and cross-modal interactions that remain invisible when either modality is considered in isolation.

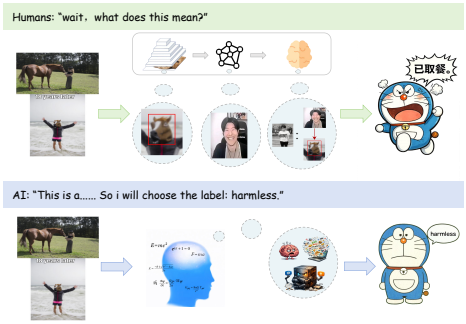

Figure 1: Comparison of human vs. AI interpretation of a meme. Humans follow a three-step reasoning chain: detect a horse-head cue, link it to esports player TheShy's nickname, and infer dehumanizing intent—thus judging the meme harmful and signaling they will report it. AI, relying mainly on surface cues, predicts a harmless label.

Current detection approaches face a trade-off between interpretability and efficiency. Some methods adopt end-to-end direct classification without explicit reasoning (Cao et al., 2022; Shah et al., 2024; Zheng et al., 2025), making it difficult to interpret the implicit attack logic in memes. Another line of work adopts a two-stage pipeline (Ji et al., 2024; Lin et al., 2024), where a large model first generates reasoning and a smaller model then performs prediction. However, this approach is cumbersome to deploy and incurs higher computational cost. In contrast, Group Relative Policy Optimization (GRPO) follows the Explain-then-Detect paradigm, generating reasoning prior to classifi-

[*] Equal contribution.
[†] Corresponding authors.
[1]Code available at https://github.com/newczh/DR-HM

Findings of the Association for Computational Linguistics: ACL 2026, pages 42975–42993
July 2-7, 2026

cation, which better matches the step-by-step nature of meme understanding (Wei et al., 2022; Pan et al., 2026). Nevertheless, recent studies show that small-scale MLLMs struggle to benefit from this paradigm due to insufficient world knowledge, often underperforming baselines trained solely with label supervision (Mei et al., 2026).

As shown in Figure 1, human interpretation of memes relies heavily on shared cultural knowledge, historical context, and nuanced visual cues. Due to the lack of this crucial cultural grounding, current small-scale MLLMs frequently miss implicit offensive meanings, rendering them unreliable for subtle toxicity detection.

This gap motivates knowledge transfer from larger closed-source models, such as the Gemini series, which encode broader commonsense and cultural knowledge and produce coherent multi-step reasoning. Mr.Harm (Lin et al., 2023) provides an initial exploration of distillation-based methods, but reports that jointly training explanation generation and harmfulness prediction can hurt performance. It even performs worse than direct label-based fine-tuning. A likely reason is its reliance on simple or free-form prompts to elicit reasoning, which often produces unstructured and noisy outputs. To mitigate this issue, Mr.Harm(Lin et al., 2023) adopts a two-stage training strategy, first learning to generate reasoning and then learning to predict labels. However, this design requires two separate models during inference, one for reasoning generation and the other for classification, which increases system complexity. A more natural alternative is to integrate reasoning and prediction within a single framework, which can improve efficiency while maintaining interpretability.

Despite closed-source models ability to provide high-quality reasoning signals, their training objectives are not fully aligned with the judgment criteria in meme harmfulness analysis, which heavily rely on subjective context, resulting in systematic biases in harmfulness judgments. We therefore need a method that exploits their reasoning quality while correcting their classification errors.

At the same time, standard SFT focuses on learning direct input–output mappings and thus struggles to capture implicit associations beyond surface patterns. Reinforcement learning approaches (Wang et al., 2025; Chen et al., 2025), such as GRPO (Shao et al., 2024) and related methods, offer a way to enhance reasoning through post-training. However, their direct application is hindered by practical challenges, including severe class imbalance in harmful meme datasets and the difficulty of designing suitable task-specific reward functions.

To address these challenges, we propose Distill-then-Reinforce Training with Cognition-Aware Data Synthesis for harmful meme detection, collectively referred to as **DR-HM**. We design a six-step structured data synthesis scheme that decomposes meme analysis into progressive cognitive stages, mirroring how human annotators reason about memes. To mitigate label bias in closed-source models, we introduce a Teacher–Reviewer–Proxy self-refinement mechanism that validates and revises generated reasoning when predictions are incorrect. Building on the synthesized data, we further propose a Distill-then-Reinforce Training strategy. In this framework, a two-stage SFT process first injects domain knowledge through complete reasoning traces and then learns compressed inference representations. Finally, we introduce an improved reinforcement learning algorithm, A-GRPO, which incorporates class-ratio–aware reward weighting to address label imbalance and sample-level dynamic KL coefficients to balance exploration based on the reference model's confidence. Our main contributions are as follows:

- We propose a six-step structured data synthesis scheme with a refinement mechanism for harmful meme detection, releasing related reasoning datasets.
- We design the Distill-then-Reinforce Training strategy combining two-stage SFT with A-GRPO, enabling open-source MLLMs with 7B parameters to perform explicit reasoning for detecting harmful memes.
- Experiments on three datasets show that DR-HM achieves 84.7% accuracy on FHM, approaching human-level performance (Kiela et al., 2020b), 84.8% on MAMI and 81.3% on PrideMM.

2 Related Work

2.1 Harmful Meme Detection

Harmful meme detection seeks to classify online memes as harmful or harmless. Due to the complex and multimodal nature of memes, which rely on contextual factors (Pramanick et al., 2021a), relying solely on unimodal detection methods often yields suboptimal performance. Consequently, researchers have shifted towards sophisticated mul-

timodal approaches. To address the inherent alignment problem between different modalities, early solutions typically employed classical two-stream architectures (Lu et al., 2019). These methods encode text and image features separately, letting deep neural networks learn joint representations through either early fusion (Kiela et al., 2020a) or late fusion (Kiela et al., 2020b). To improve cross-modal alignment, attention-based mechanisms are widely used (Kumari et al., 2024; Pramanick et al., 2021b). For example, MOMENTA leverages cross-modal attention to model interactions between images and background text at both global and local levels. Subsequently, a major line of work explored fine-tuning pre-trained multimodal models specifically for this task (Lippe et al., 2020; Muennighoff, 2020; Velioglu and Rose, 2020). Parallel to this, other studies leveraged contrastive learning frameworks based on CLIP (Radford et al., 2021) to fuse cross-modal features within a unified embedding space (Burbi et al., 2023; Kumar and Nandakumar, 2022; Mei et al., 2024; Tzelepi and Mezaris, 2025; Yang et al., 2024). In addition to architectural advancements, efforts such as data augmentation (Zhou et al., 2021; Zhu et al., 2022) and harmful target disentanglement (Lee et al., 2021) were introduced to improve classification accuracy.

However, most methods optimize classification objectives without explicitly modeling step-by-step reasoning, limiting interpretability and generalization to novel meme formats.

2.2 Explainable Prediction

As LLMs have demonstrated remarkable capabilities in complex reasoning (Brown et al., 2020; Rae et al., 2022), explainable prediction has become an important direction in multimodal learning, aiming to improve interpretability and reliability by providing reasoning alongside model predictions. In harmful meme detection, it has been explored through various approaches, including multimodal debate (Lin et al., 2024), informative captioning (Ji et al., 2024), evolution prompting (Huang et al., 2025), and prompt engineering (Cao et al., 2022).

ExPO-HM (Mei et al., 2026) further investigates incorporating reasoning into the prediction process for open-source VLMs, showing that models can benefit from training with SFT and reinforcement learning. However, its SFT stage still fails to surpass standard binary classification training in performance.

3 Method

3.1 Problem Formulation

We formulate harmful meme detection as a multimodal binary classification problem. A dataset is represented as $\mathcal{M} = \{(m_i, y_i)\}_{i=1}^{N}$, where $m_i = (v_i, t_i)$, v_i denotes the image, t_i denotes the text, $y_i \in \{harmful, harmless\}$ indicates whether the meme attacks a protected group, and N is the total number of memes in the dataset. Following ExPO-HM (Mei et al., 2026), we adopt the Explain-then-Detect paradigm where the model f_θ first generates structured reasoning $\mathbf{E}$ before producing the final prediction $\hat{y}$:

$$f_\theta(m) = \{\text{think} : \mathbf{E}, \text{answer} : \hat{y}\}. \quad (1)$$

3.2 Cognition-Aware Data Synthesis with Self-Refinement

Knowledge distillation provides an effective way to transfer reasoning capabilities to smaller models (Ho et al., 2023; Hsieh et al., 2023). While closed-source models can generate high-quality reasoning data, their training objectives are not fully aligned with the subjective standards used in meme harmfulness assessment (see Appendix B.1), which may introduce systematic biases. To mitigate this, we design a Cognition-Aware Data Synthesis framework with a Self-Refinement mechanism that leverages the strong text generation ability while ensuring alignment with ground-truth labels.

3.2.1 Structured Question Design

Two properties of meme detection shape our question design. First, unlike mathematical reasoning tasks that rely on deep logical chains, meme detection requires broad world knowledge, including an understanding of cultural stereotypes, group-specific references, and visual metaphors. Such knowledge is inherently limited in small open-source models. Second, human annotators do not assess memes through instantaneous pattern matching. Instead, they follow a systematic cognitive process that progresses from surface perception to deeper semantic judgment, as illustrated in Figure 1. Accordingly, we design structured questions that mirror this human cognition process, enabling the distillation of broader world knowledge from large-scale MLLMs pretrained on vast corpora.

Following the cognitive trajectory of perception, interpretation, cultural association, deep semantics and final judgment, we decompose meme rea-

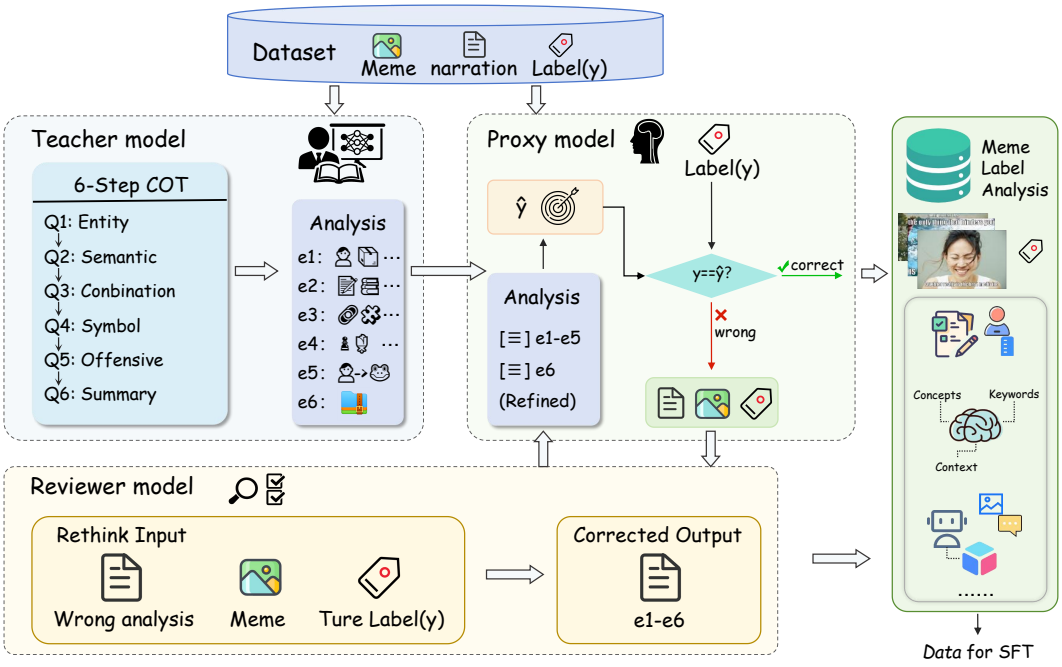

Figure 2: The overview of Cognition-Aware Data Synthesis with Self-Refinement. The Teacher generates reasoning without label hints, the Proxy Model verifies predictions, if incorrect, the Reviewer corrects the reasoning based on the ground-truth label.

soning into five sequential questions. q^1: Identify entities and objects appearing in the image and text. q^2: Describe salient visual elements. q^3: Detect biased or offensive associations and explicitly identify targeted groups. q^4: Analyze metaphors emerging from the image–text interaction. q^5: Based on the preceding analysis, determine whether the meme exhibits harmful behaviors toward specific targets. This progression follows the cognitive trajectory of human annotators and is consistent with chain-of-thought prompting (Wei et al., 2022). Notably, $\mathbf{Q}^{2:4} = (q^2, q^3, q^4)$ explicitly elicit cultural stereotypes, group-specific knowledge, and symbolic interpretations that are difficult for smaller models to infer independently. Since $\mathbf{Q}^{1:5} = (q^1, q^2, q^3, q^4, q^5)$ produce verbose output unsuitable for direct training, q_6 performs lossless compression of $\mathbf{Q}^{1:5}$, condensing core evidence without revealing classification tendency. Detailed prompt templates are provided in Appendix A.

3.2.2 Self-Refinement Mechanism

A key challenge in harmful meme annotation is subjectivity, that is, different annotators or models may have varying thresholds for what constitutes harmful content. To address this while preserving generation quality, we propose the Self-Refinement mechanism, as illustrated in Figure 2, including (1) a **Teacher Model**, such as Gemini-3-Flash (Google DeepMind, 2025), that generates initial reasoning under label-free prompting, preserving natural generation without label-induced bias; (2) a **Proxy Model**, such as Gemini-2.5-flash-lite, that evaluates whether the reasoning leads to correct predictions, simulating capability of the target student model; and (3) a **Reviewer Model**, such as Gemini-3-Flash, that corrects erroneous reasoning when necessary.

Self-Refinement mechanism decouples the strong text generation capability of closed-source LLMs from their classification bias.

For each meme $m_i \in \mathcal{M}$, the teacher model $\mathcal{T}$ first generates reasoning $\mathbf{E}_i^{(0)}$ through our six progressive questions without label hints, preserving the fluency and naturalness of the reasoning chain.

$$\mathbf{E}_i^{(0)} = \mathcal{T}(m_i, \mathbf{Q}^{1:6}), \tag{2}$$

where $\mathbf{Q}^{1:6}$ denotes the prompt containing all six questions. A proxy model $\mathcal{P}$ then predicts the label $\hat{y}_i$ conditioned on both the meme and the generated reasoning.

$$\hat{y}_i = \mathcal{P}(m_i, \mathbf{E}_i^{(0)}). \tag{3}$$

If $\hat{y}_i \neq y_i$, a reviewer model $\mathcal{R}$ revises the reasoning by incorporating the ground-truth label as guidance.

$$\mathbf{E}_i = \mathcal{R}(m_i, \mathbf{E}_i^{(0)}, y_i). \tag{4}$$

Otherwise, the original reasoning is retained: $\mathbf{E}_i = \mathbf{E}_i^{(0)}$. This design ensures that reasoning is generated without label leakage while still achieving alignment with ground-truth labels through targeted revision. Last, we construct a reasoning-augmented dataset $\mathcal{D}^* = \{(m_i, \mathbf{E}_i^*, y_i)\}_{i=1}^N$, where each sample contains the meme, high-quality reasoning text involving multi-step answers $\mathbf{E}^* = (e^1, e^2, e^3, e^4, e^5, e^6)$, and the corresponding label. This dataset serves as the foundation for subsequent supervised fine-tuning stages.

By decoupling reasoning generation from label verification, self-refinement enforces strong consistency between the synthesized rationales and ground-truth labels. This separation also prevents the propagation of teacher classification errors, overcoming an inherent flaw in standard single-pass distillation.

3.3 Distill-then-Reinforce Training

Figure 3 illustrates our three-stage Distill-then-Reinforce training framework. The three stages build capabilities progressively: Stage 1 injects reasoning knowledge from distilled data; Stage 2 compresses this knowledge into efficient inference patterns; and Stage 3 moves beyond the imitation ceiling of SFT through reinforcement learning.

Stage 1: Full Knowledge Injection In the first stage, the model is trained to generate complete responses to questions $\mathbf{Q}^{1:5}$, fully transferring the high-quality reasoning knowledge from the teacher model. This stage encourages the model to learn comprehensive, fine-grained reasoning patterns rather than directly mapping inputs to labels, thereby establishing a solid reasoning foundation for subsequent compression and optimization. Given an input meme $m \in M$ and the corresponding five questions $\mathbf{Q}^{1:5}$, the training objective is defined as:

$$\mathcal{L}_{\text{SFT-1}} = -\sum_{t=1}^{|\mathbf{e}^{1:5}|} \log p\big(\mathbf{e}^{1:5,t} \mid m, \mathbf{Q}^{1:5}, \mathbf{e}^{1:5,<t}\big), \quad (5)$$

where $\mathbf{e}^{1:5} = (e^1, e^2, e^3, e^4, e^5)$ from $\mathbf{E}^*$ in $\mathcal{D}^*$ denotes the multi-step answers corresponding to $\mathbf{Q}^{1:5}$. Here $|\mathbf{e}^{1:5}|$ is the total token count, $\mathbf{e}^{1:5,t}$ is the t-th token, and $\mathbf{e}^{1:5,<t}$ denotes all preceding tokens.

Stage 2: Reasoning Compression At this stage, the model compresses the reasoning information contained in answers e^1 to e^5 and leverages it for holistic reasoning and judgment of meme harmfulness. We design the question q^{judge} to guide the model to first generate a complete reasoning process before producing the harmfulness judgment. By training exclusively on this question, the model is encouraged to aggregate essential evidence in a compact and effective form, enabling efficient and reliable classification. The training loss for this stage is defined as:

$$\mathcal{L}_{\text{SFT-2}} = -\sum_{t=1}^{|e^6|} \log P_\theta\big(e^{6,t} \mid m, \mathbf{q}^{judge}, e^{6,<t}\big), \quad (6)$$

where e^6 denotes the response corresponding to q^{judge}. Here $|e^6|$ is the total token count, $e^{6,t}$ is the t-th token, and $e^{6,<t}$ denotes all preceding tokens.

This two-stage design allows the model to first learn full reasoning knowledge and then compress its outputs, thereby mitigating information loss that may occur when directly learning compressed reasoning.

Stage 3: Adaptive Group Relative Policy Optimization SFT is essentially behavioral cloning: the model imitates input-output mappings without exploring alternative reasoning paths (Shen et al., 2025). To enhance output regularity and classification accuracy, we introduce GRPO (Shao et al., 2024) for post-training after SFT, and propose two improvements tailored for the classification characteristics of harmful meme detection, forming A-GRPO.

Given input meme m, A-GRPO samples a group of candidate outputs $\{s_1, \ldots, s_j, \ldots, s_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, where each $s_j = (e_j, \hat{y}_j)$ comprises generated reasoning and prediction. For the t-th token of the j-th output sequence, the importance sampling ratio is defined as:

$$\rho_{j,t} = \frac{\pi_\theta(s_{j,t} \mid m, s_{j,<t})}{\pi_{\theta_{\text{old}}}(s_{j,t} \mid m, s_{j,<t})}. \quad (7)$$

Our A-GRPO objective function is:

$$\mathcal{J}_{\text{A-GRPO}}(\theta) = \mathbb{E}_{(m,y),\{s_j\}\sim\pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G} \sum_{j=1}^{G} \frac{1}{|s_j|} \sum_{t=1}^{|s_j|} \Big(\min\big(\rho_{j,t}\hat{A}_j,\ \text{clip}(\rho_{j,t}, 1-\epsilon, 1+\epsilon)\hat{A}_j\big) - \hat{\beta}_m D_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big] \Big) \Bigg], \quad (8)$$

where ϵ is the clipping threshold and $\hat{\beta}_m$ is the sample-level KL coefficient. $\hat{A}_j = R_j - mean(R)$,

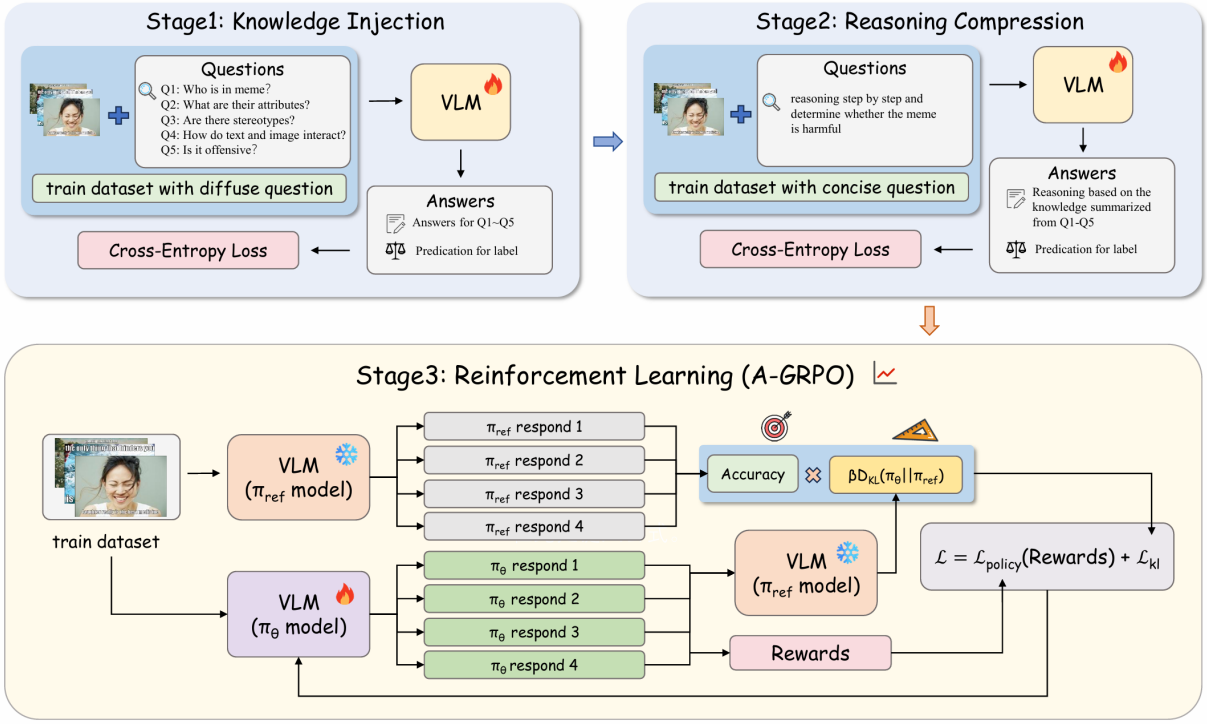

Figure 3: Overview of the Distill-then-Reinforce training framework. Stage 1 injects full reasoning knowledge ($\mathbf{Q}^{1:5}$); Stage 2 compresses reasoning into concise outputs (q^6); Stage 3 further enhances classification accuracy through reinforcement learning.

where $R = \{R_1, ..., R_G\}$ denotes the set of rewards corresponding to each candidate output s_j in the group. Below, we introduce the two improvements of A-GRPO over standard GRPO.

• **Class-Ratio Reward Weighting.** Harmful meme datasets exhibit significant class imbalance. For instance, in the FHM dataset, there are 3019 harmful samples and 5481 harmless samples. Using uniform rewards causes the model to favor predicting the majority class for higher expected returns. To address this, we design a class-aware reward function including format reward and classification reward. The classification reward is:

$$R_{\text{cls}}(\hat{y}_j, y) = \begin{cases} w, & \hat{y}_j = y \wedge y \in \mathcal{C}_{\text{min}} \\ 1, & \hat{y}_j = y \wedge y \in \mathcal{C}_{\text{maj}} \\ 0, & \hat{y}_j \neq y \end{cases}, \tag{9}$$

where $\mathcal{C}_{\text{min}}$ and $\mathcal{C}_{\text{maj}}$ denote the minority and majority classes, respectively, containing p_{min} and p_{maj} memes. $w = p_{\text{maj}}/p_{\text{min}}$ is the class ratio weight.

The format reward is used to verify whether the output adheres to the template defined in Equation 1, assigning a score of 1 for compliance and 0 otherwise. The final reward R_j for the j-th sampled output combines format compliance and classification accuracy:

$$R_j = R_{\text{fmt}}(s_j) + R_{\text{cls}}(\hat{y}_j, y). \tag{10}$$

This design provides higher incentives for correct predictions of the minority class, mitigating policy bias caused by class imbalance.

• **Sample-Level Dynamic KL Coefficient.** Standard GRPO uses a uniform KL coefficient β for all samples, making it difficult to balance preventing forgetting for easy samples and encouraging exploration for difficult samples. Before training, we perform multiple inference runs on each sample using π_{ref}, calculate the accuracy p_m, and define the sample-level KL coefficient:

$$\hat{\beta}_m = \beta \cdot \max(0.1, p_m) \tag{11}$$

where β is the base coefficient and p_m is the precomputed accuracy of the reference model π_{ref} on meme m. Samples with higher p_m receive stronger KL constraints to prevent policy deviation, while those with lower p_m receive weaker constraints, encouraging exploration of new reasoning paths.

4 Experiments

4.1 Experimental Setup

We evaluate on three mainstream harmful meme detection datasets: FHM (Kiela et al., 2020b), MAMI

Model	FHM		MAMI		PrideMM		Avg.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Closed-Source Baselines								
Gemini-3-Flash	80.4	81.8	**89.5**	**89.6**	70.6	62.8	80.2	78.1
Gemini-3-Pro	80.9	79.9	87.6	86.9	69.6	60.1	79.4	75.6
Gemini-2.5-flash-lite	68.6	68.2	80.1	79.1	62.7	70.7	70.5	72.7
GPT-5.2	71.3	68.1	85.0	85.8	65.3	70.8	73.9	74.9
GPT-5.1	70.1	66.8	85.1	85.9	64.7	64.4	73.3	72.4
Claude-Opus-4.5	80.1	81.9	89.0	89.4	69.8	64.3	79.6	78.5
Open-Source Baselines								
Qwen2.5-VL-7B (Zero-shot)	64.7	66.5	72.6	69.8	63.1	62.7	66.8	66.3
Pro-Cap	72.3	-	73.1	-	-	-	-	-
LoReHM	70.2	70.1	83.0	83.0	-	-	-	-
RA-HMD (Qwen2-VL-7B)	82.1	79.7	79.9	81.2	78.1	78.4	80.0	79.8
RA-HMD (Qwen2.5-VL-7B)	80.8	80.1	81.0	81.0	78.0	77.8	79.9	79.6
ExPO-HM (Qwen2.5-VL-7B)	-	81.1	-	82.3	-	78.7	-	80.7
Ours								
DR-HM (InternVL3-8B)	82.3	81.3	83.4	84.3	**81.5**	**82.2**	82.4	82.6
DR-HM (Qwen2.5-VL-7B)	**84.7**	**84.2**	84.8	85.5	81.3	81.1	**83.6**	**83.6**

Table 1: Main experimental results on three harmful meme detection benchmarks. "-" indicates the model was not evaluated on the corresponding task or metric. Best results are in **bold** and second-best results are underlined.

(Fersini et al., 2022), PrideMM (Shah et al., 2024). The detailed description of datasets and implementation details can be found in Appendix B. We use accuracy (Acc) and F1 score as primary metrics.

Closed-source models We evaluate several closed-source models via API using zero-shot prompts. These include Gemini-3-Flash and Gemini-3-Pro (Google DeepMind, 2025), GPT-5.1 and GPT-5.2 (OpenAI, 2025a,b), as well as Claude-Opus-4.5 (Anthropic, 2025).

Open-source baselines We compare Qwen2.5-VL-7B in a zero-shot setting, performing direct inference without fine-tuning, and Label-SFT, which is fine-tuned using only classification labels without reasoning data. We also include previous state-of-the-art methods, RA-HMD (Mei et al., 2025), a retrieval-augmented approach with contrastive learning, Pro-Cap (Cao et al., 2023), a method that generates probing-based captions, combines them with meme text to train BERT and PromptHate for hateful meme detection, LoReHM (Huang et al., 2024), which uses an agent-based LMM framework leveraging few-shot in-context learning and self-improvement for low-resource hateful meme detection, and ExPO-HM (Mei et al., 2026), which combines SFT with GRPO for an explain-then-detect strategy.

4.2 Main Results

Table 1 reports the performance of all models across three benchmarks. Our DR-HM achieves the strongest overall performance among open-source methods and remains competitive with closed-source models. With the Qwen2.5-VL-7B backbone, it reaches an average accuracy of 83.6%, outperforming strong closed-source models such as Gemini-3-Flash with 80.2% accuracy and Claude-Opus-4.5 with 79.6%. Compared with prior open-source methods, DR-HM shows improvements across all datasets. On FHM, it achieves 84.7% accuracy, matching reported human annotator agreement to the best of our knowledge. The performance gain is also robust across different backbones. Using InternVL3-8B, DR-HM achieves competitive results and obtains the best performance on PrideMM, reaching 81.5% accuracy and 82.2% F1.

4.3 Evaluation of Reasoning Quality

We evaluate reasoning quality using both subjective and objective metrics. For subjective evaluation, we employ an LLM (Gemini-3-Flash) to score generated reasoning from multiple perspectives, including completeness, coherence, and conciseness. For objective evaluation, we adopt Reasoning-Assisted Prediction Gain, RAPG =

$\text{Acc}(\mathcal{P}(m, \mathbf{E})) - \text{Acc}(\mathcal{P}(m))$, where $\mathbf{E}$ is generated by the teacher without label access and supplied to proxy $\mathcal{P}$ as reference-only evidence. The distilled data score above 9.1 in both completeness and coherence. Compared with $\mathbf{e}^{1:5}$, e^6 shows improved conciseness with only a minor completeness drop, confirming the compression design. The trained model's reasoning quality exceeds the zero-shot baseline by over 2 points in completeness (8.19 vs. 5.86) and approaches that of the distilled data. Other results can be found in Table 14.

Model	FHM			
	Comp	Coh	Conc	RAPG
Our data (Gemini-3-Flash)				
$\mathbf{e}^{1:6}$	9.25	9.38	8.58	15.4
$\mathbf{e}^{1:5}$	9.14	9.32	8.79	14.6
e^6	8.66	9.21	9.26	12.7
Closed-source				
Claude-Opus-4.5	8.61	9.21	9.19	11.28
Gemini-3-Flash	7.90	8.97	9.54	10.60
GPT-5.2	7.27	8.85	9.65	3.50
Ours (Qwen2.5-VL-7B)				
zero-shot	5.86	7.39	8.72	-3.6
DR-HM	8.19	8.74	9.07	15.0

Table 2: Reasoning quality evaluation on FHM. Comp: Completeness, Coh: Coherence, Conc: Conciseness (LLM-scored 1-10).

4.4 Ablation Study

Training Pipeline Ablation Table 3 evaluates the different stages of our training pipeline. We compare against two baselines: Zero-shot, which directly uses the base model, and Label-SFT, which fine-tunes solely on classification labels. We also include SFT EasyR, a simple distillation approach that trains on single free-form rationales generated by the teacher. For our proposed method, SFT 1 denotes Stage 1 training, +SFT 2 adds Stage 2 compression, and +A-GRPO further introduces Stage 3 reinforcement learning. We additionally evaluate Only-SFT 2, which skips Stage 1.

SFT EasyR and SFT 1 do not consistently beat the Label-SFT baseline. They show improvements on MAMI but actually perform worse on FHM and PrideMM. This inconsistency implies that generating unconstrained or very long reasoning traces introduces noise and weakens the main classification signal.

Adding the compression stage in SFT 2 alleviates this issue and improves performance across all datasets. This indicates that concise, decision-focused reasoning is more suitable for classification. However, compression alone is not sufficient. When we test Only-SFT 2, which skips Stage 1 entirely, performance drops significantly, suggesting that the initial knowledge injection is necessary for effective compression. Finally, applying A-GRPO achieves the best overall performance, showing the benefit of reinforcement learning.

Model	FHM		MAMI		PrideMM	
	Acc	F1	Acc	F1	Acc	F1
Zero-shot	64.7	66.5	72.6	69.8	63.9	68.2
Label-SFT	79.8	80.3	75.7	79.4	75.3	75.1
SFT EasyR	77.4	73.5	78.8	78.6	70.8	70.4
SFT 1	77.0	74.9	79.2	79.6	70.6	70.1
+ SFT 2	82.4	82.4	82.0	83.5	77.3	77.2
+ A-GRPO	**84.7**	**84.2**	**84.8**	**85.5**	**81.3**	**81.1**
Only-SFT 2	80.8	80.7	80.4	81.9	73.0	72.8

Table 3: Ablation results for training pipeline.

GRPO Configuration Ablation Table 4 examines the components of A-GRPO. We evaluate two variants to isolate their effects. Removing the class-ratio reward weighting (w/o crw) eliminates the additional incentives for minority samples, while removing the dynamic KL coefficient (w/o dKL) replaces the adaptive constraint with a fixed value. Both changes lead to performance drops. On the imbalanced FHM dataset, accuracy decreases from 84.7% to 82.8% without class-ratio reward weighting. Similarly, removing the dynamic KL coefficient reduces both accuracy and F1 across all datasets. These results indicate that the two mechanisms have distinct but complementary effects.

Model	FHM		MAMI		PrideMM	
	Acc	F1	Acc	F1	Acc	F1
w/o crw	82.8	82.6	-	-	-	-
w/o dKL.	83.5	82.9	84.1	84.1	80.5	80.9
Full	**84.7**	**84.2**	**84.8**	**85.5**	**81.3**	**81.1**

Table 4: Ablation results for GRPO configurations. MAMI and PrideMM are excluded from the w/o crw as their training sets are nearly balanced. Detailed dataset statistics can be found in Appendix B.2.

MAMI and PrideMM are not included in the class-ratio reward weighting ablation study because their training sets are nearly balanced. Detailed dataset statistics can be found in Appendix B.

Rollouts (m)	4	8	12
Accuracy	81.3	81.9	82.4

Table 5: Effect of the number of inference rollouts (m) on PrideMM.

Impact of Reference Rollouts and Efficiency The dynamic KL coefficient requires running the reference model for m inference rollouts to estimate the sample-level accuracy. To evaluate how m affects both performance and computational cost, we test $m \in \{4, 8, 12\}$ on the PrideMM dataset. As shown in Table 5, the accuracy steadily improves from 81.3% to 82.4% as the number of rollouts increases.

However, this pre-computation introduces a efficiency trade-off. For context, our full training pipeline takes approximately 7 hours on 4× NVIDIA RTX PRO 6000 GPUs, while a single round of reference inference requires about 1 hour on a single GPU. Although $m = 12$ yields the highest accuracy, $m = 4$ strikes a practical balance between computational overhead and performance. Therefore, we adopt $m = 4$ as the default setting for our main experiments, leaving larger rollout sizes as a reliable option to further boost performance when sufficient computational resources are available.

Teacher Model	FHM		MAMI		PrideMM	
	Acc	F1	Acc	F1	Acc	F1
Gemini-3 -Flash	82.4	82.4	82.0	83.5	77.3	77.2
Qwen3-VL -235B	78.9	78.1	79.3	81.2	76.7	74.3

Table 6: Impact of closed-source and open-source teacher models on Qwen2.5-VL-7B at the SFT stage.

Framework Generalization We further examine the robustness of our framework from two perspectives: cross-cultural transfer and reliance on closed-source teacher models.

Since the primary benchmarks are limited to English and Western contexts, we additionally evaluate DR-HM on ToxiCN_MM (Lu et al., 2024), a Chinese harmful meme dataset grounded in a distinct cultural setting. Without requiring structural changes to the training pipeline, DR-HM achieves an F1 of 83.3% and an $F1_{Harm}$ of 76.5%, outperforming the best baseline reported in the original dataset paper, as shown in Table 7. This result suggests that the framework transfers well across languages and cultural contexts.

Model	F1	$F1_{Harm}$
CLIP + MKE (Lu et al., 2024)	80.2	72.4
DR-HM (Ours)	**83.3**	**76.5**

Table 7: Cross-cultural generalization on ToxiCN_MM, a Chinese harmful meme benchmark.

We also investigate whether our pipeline depends on closed-source APIs for effective distillation. To this end, we replace the Gemini-3-Flash teacher with an open-source alternative, Qwen3-VL-235B-A22B. As shown in Table 6, although the performance slightly decreases, the student model remains competitive across all benchmarks in both accuracy and F1. These findings indicate that strong performance can still be achieved without relying on closed-source models, offering a practical path toward lower-cost and more reproducible training.

5 Conclusions

To distill rich knowledge for harmful meme detection, we propose DR-HM, a Distill-then-Reinforce training framework with cognition-aware data synthesis. Our approach first simulates the human perceptual process of meme interpretation by decomposing meme analysis into six progressive steps, ranging from entity identification to harmfulness recognition, and introduces a self-refinement mechanism to improve the quality of the synthesized data. Building on the resulting inference-enhanced dataset, we further adopt a Distill-then-Reinforce training strategy that employs a two-stage fine-tuning procedure and incorporates Adaptive GRPO to enhance both reasoning and detection capabilities. On three benchmarks, DR-HM with a 7B-parameter model reaches 83.6% average accuracy, matching or exceeding closed-source models an order of magnitude larger. Future work will focus on improving the model's cross-domain detection capability, particularly in multilingual and cross-cultural settings where culturally specific knowledge plays a more critical role.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N25XQD004).

Limitations

Our approach relies on closed-source models for initial knowledge distillation, inheriting potential biases in their world knowledge. The dynamic KL coefficient requires pre-computing sample difficulties, adding computational overhead.

Ethical Considerations

This work constructs an inference-enhanced multimodal dataset based on publicly available datasets using closed-source large multimodal language models (MLLMs) accessed through official APIs. All source datasets used in this study are publicly available and were collected in compliance with their original licenses and terms of use. The closed-source MLLMs employed in our framework are used solely for knowledge distillation and reasoning generation, without accessing private or sensitive user data. No new personal data are collected, and no attempt is made to identify or infer personal attributes beyond those explicitly annotated in the original datasets.

Given that harmful meme detection inherently involves sensitive and potentially offensive content, we take care to use the data exclusively for research purposes aimed at improving content moderation systems.

References

Anthropic. 2025. System card: Claude opus 4.5. https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf. Accessed: January 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2824–2828.

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 31st ACM International Conference on Multimedia*, MM '23, page 5244–5252, New York, NY, USA. Association for Computing Machinery.

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

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. 2025. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *Preprint*, arXiv:2506.16141.

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

Google DeepMind. 2025. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/. Accessed: January 2026.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with LMM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2025. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7321–7330, Abu Dhabi, UAE. Association for Computational Linguistics.

Junhui Ji, Xuanrui Lin, and Usman Naseem. 2024. Capalign: Improving cross modal alignment via informative captioning for harmful meme detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4585–4594, New York, NY, USA. Association for Computing Machinery.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *Preprint*, arXiv:2210.05916.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3Hop-CoT: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138, Miami, Florida, USA. Association for Computational Linguistics.

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

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2359–2370, New York, NY, USA. Association for Computing Machinery.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2012.12871.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. In *Advances in Neural Information Processing Systems*, volume 37, pages 13302–13320. Curran Associates, Inc.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand. Association for Computational Linguistics.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23806–23828, Suzhou, China. Association for Computational Linguistics.

Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Pengda Qin, Yuhong Li, Da Chen, and Bill Byrne. 2026. Expo-hm: Learning to explain-then-detect for hateful meme detection. *Preprint*, arXiv:2510.08630.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visiolinguistic models applied to hateful memes. *Preprint*, arXiv:2012.07788.

OpenAI. 2025a. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. `https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf`. Accessed: January 2026.

OpenAI. 2025b. Update to gpt-5 system card: Gpt-5.2. `https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf`. Accessed: January 2026.

Fengjun Pan, Xiaobao Wu, Tho Quan, and Anh Tuan Luu. 2026. Read as you see: Guiding unimodal llms for low-resource explainable harmful meme detection. *Preprint*, arXiv:2506.08477.

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

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

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

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2022. Scaling language models: Methods, analysis & insights from training gopher. *Preprint*, arXiv:2112.11446.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *Preprint*, arXiv:2504.07615.

Maria Tzelepi and Vasileios Mezaris. 2025. Improving multimodal hateful meme detection exploiting lmm-generated knowledge. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 202–211.

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

Zihan Wang, Xingle Xu, Hao Wang, Yiwen Ye, Yuchen Li, Linhao Wang, Hongze Tan, Peidong Wang, Shi Feng, Guoqing Chen, and 1 others. 2025. A survey on entropy mechanism in large reasoning models. *Authorea Preprints*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Chuanpeng Yang, Yaxin Liu, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2024. Uncertainty-guided modal rebalance for hateful memes detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4361–4371, Bangkok, Thailand. Association for Computational Linguistics.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025. Multi-granular multimodal clue fusion for meme understanding. *Preprint*, arXiv:2503.12560.

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.

Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. 2022. Multimodal zero-shot hateful meme detection. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, page 382–389, New York, NY, USA. Association for Computing Machinery.

A Prompts

A.1 Detailed Prompt Templates

We design six progressive questions to guide the teacher model in generating structured reasoning text. The question design follows a cognitive pathway of "perception → description → cultural association → deep semantics → judgment → compression."

FHM Prompts Q1–Q5

> Q1. Identify all the objects and people in the image and the text. Note that if pronouns refer to a person, they should be counted as people; if pronouns refer to an object, they should be counted as objects.
>
> Q2. Please provide a detailed description of the people or objects, paying close attention to the following attributes: skin color, race, ethnicity, nationality, religion, sex, sexual orientation, disability, and disease. If certain attributes are not discernible or applicable, you may omit them from your explanation.
>
> Q3. Please explain whether any people or objects are associated with stereotypes or problematic representations. Consider behaviors such as satire or insult, violent or dehumanizing speech (e.g., comparing people to animals), statements of inferiority, calls for exclusion or segregation, or mocking of hate crimes. Pay special attention to how people and objects are referenced, including direct mentions, pronouns, or implied references within the image or text. Describe whether any of these references contribute to or reflect harmful stereotypes.
>
> Q4. From now on, you need to identify what each person or object is doing. Pay attention to the cross-modal connections between people and objects in both the text and the image, as well as the referents of pronouns. Based on the above, explain whether the people or objects carry any implied metaphors or symbolic references. Pay attention to the connections between the behaviors of the people and objects, as well as any associated stereotypes.
>
> Q5. Based on the above analysis, please explain whether this meme contains any of the following related behaviors toward the listed people, objects, or the persons or things implied by them: satire or insult, violent or dehumanizing speech (such as comparing people to non-human things like animals), statements of inferiority, calls for exclusion or segregation, or mocking hate crimes. Please note that all of these behaviors are considered harmful. A direct or indirect attack on criminals, terrorists, or ongoing or completed criminal activities is considered harmless.

MAMI Prompts Q1–Q5

> Q1. This meme is accompanied by the text 'text'. Identify all the objects and people in the image and the text. Note that if pronouns refer to a person, they should be counted as people; if pronouns refer to an object, they should be counted as objects.
>
> Q2. Please provide a detailed description of the people or objects, paying close attention to the following attributes: skin color, race, ethnicity, nationality, religion, sex, sexual orientation, disability, and disease. If certain attributes are not discernible or applicable, you may omit them from your explanation.
>
> Q3. Please explain whether any people or objects are associated with stereotypes or problematic representations. Consider behaviors or content such as shaming, stereotype, objectification, or violence. Pay special attention to how people and objects are referenced, including direct mentions, pronouns, or implied references within the image or text. Describe whether any of these references contribute to or reflect harmful stereotypes.
>
> Q4. From now on, you need to identify what each person or object is doing. Pay attention to the cross-modal connections between people and objects in both the text and the image, as well as the referents of pronouns. Based on the above, explain whether the people or objects carry any implied metaphors or symbolic references. Pay attention to the connections between the behaviors of the people and objects, as well as any associated stereotypes.
>
> Q5. Based on the above analysis, please explain whether this meme contains any of the following related behaviors toward the listed people, objects, or the persons or things implied by them: whether it depicts an offensive, sexist, or hateful scene (implicitly or explicitly, weak or strong) targeting a woman or a group of women. Please note that all of these behaviors are considered harmful.

PrideMM Prompts Q1–Q5

> Q1. This meme is accompanied by the text 'text'. Identify all objects and any individuals, communities, or organizations present in the image and the text. Note that if pronouns refer to an individual, a community, or an organization, they should be counted as such; if pronouns refer to objects, they should be counted as objects.
>
> Q2. Please provide a detailed description of all objects, individuals, communities, and organizations present. For individuals, pay close attention to the following attributes: skin color, race, ethnicity, nationality, religion, sex, sexual orientation, disability, and disease. If certain attributes are not discernible or applicable, you may omit them from your explanation.
>
> Q3. Please explain whether any objects, individuals, communities, or organizations present are associated with stereotypes or problematic representations. Consider behaviors such as satire or insult, violent or dehumanizing speech (e.g., comparing people to animals), statements of inferiority, calls for exclusion or segregation, or mocking of hate crimes. Pay close attention to how these entities are referenced, including direct mentions, pronouns, or implied references within the image or text. Describe whether any of these references contribute to or reflect harmful stereotypes.
>
> Q4. From now on, you need to identify what each objects, individuals, communities, or organizations present is doing. Pay attention to the cross-modal connections between these entities in both the text and the image, as well as the referents of pronouns. Based on the above, explain whether these objects, individuals, communities, or organizations present carry any implied metaphors or symbolic references. Pay attention to the connections between their behaviors, as well as any associated stereotypes.
>
> Q5. Based on the above analysis, please explain whether this meme contains any of the following related behaviors toward the listed objects, individuals, communities, organizations present, or any entities implied by them: satire or insult, violent or dehumanizing speech (such as comparing people to non-human things like animals), statements of inferiority, calls for exclusion or segregation, or mocking hate crimes. Please note that all of these behaviors are considered harmful.

Q6 for all datasets Perform lossless compression of Q1–Q5 to generate a chain-of-thought. Output format specification: Use subheadings with concise phrases, objectively narrating without revealing classification tendencies.

> Q6. Please summarize your analysis by integrating insights from all previous responses, with particular emphasis on your answer to Question 5, which highlights potential concerns. Based on this, construct a clear and logically sound reasoning process to assess whether the meme might contain content that violates the criteria for problematic material. Focus exclusively on the reasoning process; do not offer any explicit final judgment or use terms like 'harmful' or 'harmless'

.

A.2 Prompt Templates for the SFT EasyR Baseline

To construct the SFT EasyR baseline, we prompt the teacher model to generate a single, free-form rationale before outputting the final classification label. Unlike our proposed structured questions (Q1–Q6), which strictly enforce a multi-step cognitive pathway, these baseline prompts simply ask the model to analyze the meme organically. Below are the exact prompt templates used for each dataset.

Prompt for the FHM Dataset

> This image includes text with 'text',you need to analyze the image to determine whether the meme is harmful. The following is the definition of harmful. A direct or indirect attack on

people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech. Attacking groups perpetrating hate (e.g. terrorist groups) is not considered hate.

Prompt for the MAMI Dataset

This image includes text with 'text',you need to analyze the image to determine whether the meme is harmful. A meme is harmful if it describes an offensive, sexist, or hateful scene (weak or strong, implicitly or explicitly) and attack a woman or a group of women. The attack can be expressed through shaming, stereotype, objectification, or violence.

Prompt for the PrideMM Dataset

This image includes text with 'text',you need to analyze the image to determine whether the meme is harmful. The following is the definition of harmful: A direct or indirect attack targeting an individual, a community, an organization, or an undirected target. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech. Attacking groups perpetrating hate (e.g. terrorist groups) is not considered hate.

A.3 LLM-as-Judge Prompt

We employ Gemini-3-Flash as the judge model to score reasoning quality on a 1–10 scale.

Evaluation Prompt

You are a strict and critical evaluator. Please evaluate the quality of the following reasoning text for harmful meme detection task.
Reasoning to evaluate:
<reasoning_content>
reasoning_content ...
</reasoning_content>
Dimensions and Scoring Rubric (1–10 scale):
Completeness - Whether it covers key information needed to judge harmfulness (entities, context, stereotypes, symbolic meanings, etc.)
1–2: Missing most analytical components;
5–6: Basic structure present but lacks depth;
9–10: Comprehensive analytical framework (rare)
Coherence - Whether the reasoning is clear, logical, and well-organized with smooth transitions between points
1–2: Disorganized, contradictory, or incoherent;
5–6: Acceptable structure but some logical jumps;
9–10: Flawless logical chain (rare)
Conciseness - Whether it avoids redundancy and delivers information efficiently without unnecessary repetition
1–2: Extremely verbose or repetitive;
5–6: Some unnecessary content;
9–10: Perfectly concise (rare)
Please output strictly in the following format:
<completeness>score</completeness>
<coherence>score</coherence>
<conciseness>score</conciseness>
<tendency>harmful or harmless</tendency>

B Experiments

B.1 Closed-Source Model Analysis

We analyze the precision-recall trade-offs exhibited by closed-source models to understand their underlying classification biases, which informed the design of our Self-Refinement mechanism.

Over-Alignment Phenomenon. Several safety-aligned models show a better safe than sorry tendency: high recall at the cost of precision. On FHM, Claude-Opus-4.5 achieves the highest recall of 90.7% among all models but with precision of only 74.7%, indicating a tendency to over-predict harmfulness. Similarly, Gemini-2.5-flash-lite on PrideMM shows extreme over-caution with 92.3% recall but merely 57.3% precision, misclassifying nearly half of harmless memes as harmful.

While this conservative stance is desirable for production safety systems, it introduces systematic bias when using these models for annotation or knowledge distillation.

Under-Detection Pattern. Conversely, GPT-series models on FHM exhibit the opposite bias: GPT-5.1 and GPT-5.2 achieve relatively higher precision (73.1% and 74.8%) but notably lower recall (61.6% and 62.4%), suggesting a more permissive threshold for harmfulness judgment. On PrideMM, Gemini-3-Flash and Gemini-3-Pro show high precision (81.8% and 83.5%) but extremely low recall (51.0% and 47.0%), potentially reflecting insufficient sensitivity to LGBTQ+-targeted content in their training.

Connection to Self-Refinement Statistics. These biases are precisely what our Self-Refinement mechanism targets. As shown in Table 15, 17.9% of samples required correction through the Reviewer, corresponding to cases where the Teacher model's reasoning led to predictions inconsistent with ground-truth labels. The 0.2% failure rate represents cases where systematic alignment biases, such as extreme sensitivity to historically sensitive terms like gas chambers (see Section C), prevent correction even with label guidance.

These statistics confirm that closed-source models generate high-quality reasoning text with 9.25

Model	Acc	Prec	Rec	F1
Gemini-2.5-flash-lite	68.6	67.7	68.7	68.2
Gemini-3-Flash	80.4	74.8	80.4	77.5
Gemini-3-Pro	80.9	82.4	77.5	79.9
GPT-5.1	70.1	73.1	61.6	66.9
GPT-5.2	71.3	74.8	62.4	68.0
Claude-Opus-4.5	80.1	74.7	90.7	81.9

Table 8: Complete results of closed-source models on the FHM dataset.

Completeness and 9.38 Coherence in Table 2, yet their classification judgments require calibration against dataset-specific annotation standards, highlighting the gap addressed by our Self-Refinement mechanism, which decouples reasoning generation from label alignment.

Model	Acc	Prec	Rec	F1
Gemini-3-Flash	89.5	88.9	90.2	89.5
Gemini-3-Pro	87.6	92.2	82.2	86.9
Gemini-2.5-flash-lite	80.1	83.2	75.4	79.1
GPT-5.1	85.1	84.9	86.9	85.9
GPT-5.2	85.0	84.1	87.5	85.8
Claude-Opus-4.5	89.0	88.6	90.2	89.4

Table 9: Complete results of closed-source models on the MAMI dataset.

Model	Acc	Prec	Rec	F1
Gemini-3-Flash	70.6	81.8	51.0	62.8
Gemini-3-Pro	69.6	83.5	47.0	60.1
Gemini-2.5-flash-lite	62.7	57.3	92.3	70.7
GPT-5.1	64.7	63.3	65.6	64.4
GPT-5.2	65.3	59.9	86.6	70.8
Claude-Opus-4.5	69.8	75.8	55.9	64.3

Table 10: Complete results of closed-source models on the PrideMM dataset.

B.2 Experimental Details

B.2.1 Dataset Statistics

We conduct experiments on three datasets for harmful meme detection, including FHM (Kiela et al., 2020b), MAMI (Fersini et al., 2022), and PrideMM (Shah et al., 2024), with dataset statistics summarized in Table 11.

The class distribution across datasets shaped our A-GRPO design. FHM exhibits significant class imbalance with only 35.52% harmful samples in the training set, creating a 1:1.82 ratio between minority and majority classes. This imbalance causes standard RL algorithms to favor predicting the majority class for higher expected returns.

In contrast, MAMI is fully balanced, and PrideMM is nearly balanced. This motivates our ablation design in Table 4, where the class-ratio reward weighting ablation is conducted only on FHM, since this component degenerates to uniform weighting when $w = p_{\text{maj}}/p_{\text{min}} \approx 1$ in balanced datasets such as MAMI and PrideMM.

Dataset	Train		Valid		Test	
	H	NH	H	NH	H	NH
FHM	3,019	5,481	247	253	490	510
MAMI	4,500	4,500	500	500	500	500
PrideMM	2,120	2,208	115	113	247	260

Table 11: Label distribution statistics across the train, validation, and test splits for the FHM, MAMI, and PrideMM datasets. Notably, while MAMI and PrideMM maintain a nearly balanced class ratio across all splits, FHM exhibits severe class imbalance in its training set, with harmful memes constituting only 35.52%.

B.2.2 Implementation Details

We employ LLaMA-Factory for the SFT stage and VLM-R1 for the GRPO stage. All experiments are conducted on 4× NVIDIA RTX PRO 6000 GPUs (96GB each). SFT hyperparameter settings can be found in Table 12, and GRPO hyperparameter settings can be found in Table 13.

Parameter	Stage 1	Stage 2
Learning Rate	1e-5	1e-5
Batch Size	24	24
Epochs	3	3
Max New Tokens	4096	4096
Freeze Vision Tower	true	true
Freeze Multimodal projector	true	true
Freeze Language Model	false	false

Table 12: SFT hyperparameter settings.

Parameter	Value
Base KL Coefficient (β)	0.04
Temperature	0.9
Top_p	1
Top_k	50
Number Generations	4
Learning Rate	5e-6
Per Device Train Batch Size	24
Reference Inference Runs (K)	4

Table 13: GRPO hyperparameter settings.

B.3 Additional Experiments

Table 14 reports reasoning quality on MAMI and PrideMM. Our cognition-aware synthesis produces consistently strong reasoning on both datasets. On MAMI, the full reasoning chain reaches 9.11 in completeness and 9.22 in coherence. On PrideMM, the scores are 9.18 and 9.27. These results show that the six-step design can reliably produce reasoning that is both comprehensive and well-structured across different domains.

When detailed reasoning from e^1 to e^5 is compressed into e^6, conciseness improves from 8.58 to 9.16 on MAMI and from 8.63 to 9.08 on PrideMM. This compression leads to a moderate drop in completeness, but RAPG only decreases slightly, from 10.10 to 9.80 on MAMI and from 7.50 to 6.89 on PrideMM. This suggests that most decision-relevant information is retained after compression.

Among closed-source baselines, GPT-5.2 performs notably worse on PrideMM, with completeness at 4.83 and RAPG at 1.60. In contrast, Claude-Opus-4.5 remains stable across datasets, with completeness scores of 8.72 and 8.45. This difference indicates that Claude provides more balanced coverage.

Finally, Distill-then-Reinforce training brings clear improvements over zero-shot performance. RAPG increases from -6.92 to 4.97 on MAMI and from 0.39 to 14.83 on PrideMM.

B.4 RAPG Additional Details

Formally, $\text{RAPG} = \text{Acc}(\mathcal{P}(m, \mathbf{E})) - \text{Acc}(\mathcal{P}(m))$, where the teacher generates $\mathbf{E}$ label-free and $\mathcal{P}$ is explicitly instructed to treat it as reference only. Empirically, RAPG approaches the teacher–proxy accuracy gap and is slightly higher, confirming that structured reasoning transfers useful evidence beyond $\mathcal{P}$'s intrinsic capability.

C Case Study of the Distillation Process

We analyze the role of Self-Refinement in the data synthesis process through concrete examples. Cases are categorized into three types based on the number of Teacher-Proxy-Reviewer correction iterations. Table 15 summarizes the distribution of correction attempts during reasoning data generation on the FHM training set.

C.1 Direct Success Case (attempts = 1)

Meme Text: its their character not their color that matters

Figure 4: Direct success case. The meme promotes character-based judgment over racial prejudice, with clear positive messaging that requires no correction.

Ground Truth: harmless

Reasoning Highlights:

- e^3 **(Stereotype Analysis):** The text echoes Martin Luther King Jr.'s famous ideal—judging people by the content of their character rather than the color of their skin—representing an anti-racial prejudice expression.
- e^5 **(Harm Judgment):** The content advocates evaluation standards based on character rather than race, containing no derogatory, dehumanizing, or exclusionary language.
- e^6 **(Summary):** The meme employs a common social commentary format, advocating equality rather than hatred.

Analysis When the semantic intent is unambiguous and the cultural reference is well-known, a single reasoning pass suffices. The text directly expresses positive racial equality ideals, and the visual elements contain no negative stereotypes.

C.2 Rethink Correction Case (attempts > 1)

Meme Text: woman badly injured in fight
Ground Truth: harmless

Teacher's Initial Reasoning (Incorrect Tendency): The initial reasoning over-interpreted the violent elements in the image, associating them with violence against women as a group. The Proxy Model predicted *harmful* based on this reasoning.

Reviewer-Corrected Reasoning:

Model	MAMI				PrideMM			
	Comp	Coh	Conc	RAPG	Comp	Coh	Conc	RAPG
Our data (Gemini-3-Flash)								
$\mathbf{e}^{1:6}$	9.11	9.22	8.40	10.60	9.18	9.27	8.34	8.20
$\mathbf{e}^{1:5}$	9.09	9.25	8.58	10.10	8.49	9.03	8.63	7.50
e^6	8.55	9.13	9.16	9.80	8.71	9.16	9.08	6.89
Closed-source								
Claude-Opus-4.5	8.72	9.28	9.25	8.87	8.45	9.15	9.12	6.34
Gemini-3-Flash	8.12	9.05	9.58	9.34	7.78	8.92	9.48	6.81
GPT-5.2	7.45	8.92	9.68	5.02	4.83	8.78	9.62	1.60
Ours								
zero-shot	6.15	7.62	8.85	-6.92	6.29	8.01	9.03	0.39
Qwen2.5-VL-7B	8.35	8.85	9.18	4.97	8.19	8.83	8.86	14.83

Table 14: Reasoning quality evaluation on MAMI and PrideMM. Comp: Completeness, Coh: Coherence, Conc: Conciseness (LLM-scored 1-10). RAPG: Reasoning-Assisted Prediction Gain (proxy model accuracy improvement with reasoning).

Category	Count	Ratio
Direct Success	6959	81.9%
Refine Corrected	1523	17.9%
Refine Failed	18	0.2%

Table 15: Distribution of Self-Refinement outcomes on FHM training set.

Figure 5: Rethink correction case. Initial reasoning over-interpreted violent imagery; Reviewer corrected by emphasizing the neutral, descriptive nature of the text.

- e^3 **(Stereotype Analysis):** The text provides an objective description of the image content, without implying that violence is targeted at women as a group or that it is being justified.
- e^5 **(Harm Judgment):** Characterizing the event as a "fight" rather than domestic violence or gender-based violence constitutes a neutral factual statement.
- e^6 **(Summary):** The meme functions as a literal description, lacking typical markers of hate speech.

Analysis This case illustrates the value of Self-Refinement. The Teacher model's initial judgment may have been overly sensitive due to the violent visual elements in the image. The Reviewer corrected the reasoning by re-examining the neutral descriptive nature of the text. This illustrates how the mechanism recovers from initial misclassifications.

C.3 Failure Case

Figure 6: Failure case. The meme contains the historically sensitive phrase gas chambers without an explicit protected-group target, which leads models to over-predict harmfulness.

Meme Text: rev up those gas chambers
Ground Truth: harmless

Teacher's Incorrect Reasoning:

- e^3 **(Stereotype Analysis):** Identified the his-

torical association between gas chambers and the Holocaust, interpreting this as mockery of genocide.

- e^5 **(Harm Judgment):** Determined that the content trivializes mass murder through a pop culture meme template.
- e^6 **(Summary):** The meme exploits historical tragedy for comedic effect.

Post-Rethink Incorrect Reasoning: The Reviewer attempted to reanalyze from the perspective of no specific attack target:

- **Argument:** The image is a generic cartoon fish, and the text does not explicitly target any protected group.
- **Outcome:** The Proxy Model still predicted *harmful* based on the inherent sensitivity of "gas chambers."

Failure Analysis After RLHF alignment, closed-source models adopt conservative stances toward content involving historically sensitive events and are difficult to persuade through reasoning text to change their judgments. The harmfulness of gas chambers in the absence of a specific attack target involves subjective judgment. The FHM dataset may employ more lenient harmless criteria, labeling borderline content without explicit attack targets as harmless.

D Case Study of the DR-HM Method

D.1 Harmful Case

Figure 7: Examples of harmful case

Meme Text: girl goes out with him blind date
Ground Truth: harmful
Prediction: harmful
Reasoning: Summary and Reasoning Process: The analysis of this meme centers on the intersection of a specific celebrity's physical appearance and the cultural context surrounding his identity.

1. **Identification of Subject:** The image identifies Stevie Wonder, a prominent figure known for his blindness.

2. **Linguistic Analysis:** The text uses the word b̈lindïn two distinct senses: literally (as a description of blindness) and figuratively (as part of the idiom b̈lind datë).

3. **Synthesis of Meaning:** The humor is derived from the literal interpretation of a b̈lind datewẗhen applied to a person who is physically blind. This creates a pun that relies on the subject's disability.

4. **Assessment of Tone:** The tone is satirical. It uses a protected attribute–disability–as the central mechanism of its joke.

5. **Reasoning on Criteria:** The meme targets an individual based on their disability. In the context of content safety, the use of a protected attribute as a punchline or a means of mockery can be viewed through the lens of satire directed at a group. The reasoning process suggests that because the humor is inextricably linked to a disability, it functions as a joke about that protected characteristic, which aligns with the criteria for material that mocks or demeans individuals based on their physical attributes.

D.2 Harmless Case

Figure 8: Examples of harmless case

Meme Text: like and share if you want loretta lynch investigated
Ground Truth: harmless
Prediction: harmless
Reasoning: Summary and Reasoning Process: The analysis of this meme centers on the intersection of a specific political figure's identity and her public role.

1. **Identification of Target:** The meme tar-

gets Loretta Lynch, a Black woman and former U.S. Attorney General.

2. **Contextual Analysis:** The text advocates for her ïnvestigation,ẅhich is a standard procedure in political discourse regarding public officials.

3. **Attribute Check:** While the subject belongs to protected groups, the text does not reference those attributes. There is no evidence in the image or text that the subject is being attacked because of her race or sex.

4. **Behavioral Assessment:** The meme uses a call to action (l̈ike and shar̈e) to mobilize support for political accountability. It does not utilize dehumanizing metaphors, incite violence, or promote stereotypes associated with her protected status.

5. **Reasoning:** Because the content focuses on the professional conduct of a public official and lacks any derogatory or exclusionary language based on protected characteristics, it functions as a piece of political speech. The reasoning process suggests that the meme is a form of expression regarding a public figure's actions rather than an attack on their inherent identity.