# LEARN BEFORE YOU JUDGE: PROGRESSIVE KNOWLEDGE-TO-DECISION ALIGNMENT FOR EXPLAINABLE HATEFUL MEME DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hateful memes spread abusive content through implicit interactions between images and text, posing serious threats to the safety of online communities. In recent years, multimodal large language models have been widely used for hateful meme detection and are increasingly adopted to generate explainable detection results. However, we find that existing explain-then-detect methods often couple explanation generation and label prediction within the same training process. This coupling causes interference between task objectives, leading to limited detection performance and even worse results than simple SFT baselines. To address these challenges, we propose ProKDA, a progressive knowledge-to-decision alignment method for explainable hateful meme detection. Inspired by the human annotation training process, ProKDA first uses an agentic background knowledge construction pipeline to obtain external knowledge related to meme understanding. It then adopts a three-stage training strategy that sequentially performs background knowledge learning, hatefulness detection learning, and hatefulness boundary alignment. Unlike prior explain-then-detect methods that jointly optimize both tasks, ProKDA focuses on a single training objective at each stage. This design reduces interference between the two tasks and progressively transforms background knowledge into robust detection decisions. Experiments on three public hateful meme benchmarks show that ProKDA achieves state-of-the-art detection performance and provides accurate, explainable, and evidence-supported decisions for hateful meme moderation.

## 1 INTRODUCTION

With the rapid development of the Internet and social media, memes have become a prevalent medium on social media platforms, usually consisting of images and text (Dancygier & Vandelanotte, 2017). Their concise, humorous, and highly shareable form makes them an important way for users to express opinions and emotions (Lestari et al., 2024; Castaño Díaz, 2013). However, memes can also be used to incite hatred, manipulate public opinion, and spread prejudice or harmful values (Davidson et al., 2019; Martinez Pandiani et al., 2025; Kennedy et al., 2022). Hateful memes not only harm individuals and targeted groups, but also threaten social stability and the health of online communities (Alafnan, 2025; Sharma et al., 2022). Therefore, accurate and automatic hateful meme detection is crucial.

In recent years, the release of several benchmarks has advanced research on hateful meme detection, such as the Hateful Memes Challenge dataset (HMC) (Kiela et al., 2020) and the Multimedia Automatic Misogyny Identification dataset (MAMI) (Fersini et al., 2022). Early studies mainly focus on binary classification, i.e., determining whether a meme is hateful or benign. These methods usually use textual and visual features from memes and improve detection performance through attention mechanisms or other multimodal fusion strategies (Shang et al., 2021; Suryawanshi et al., 2020; Kumar & Nandakumar, 2022; Burbi et al., 2023). However, this direct detection paradigm provides limited explainability, which limits its use in real-world scenarios.

The development of multimodal large language models (MLLMs) (Liu et al., 2024; Chen et al., 2023; Lu et al., 2024; Bai et al., 2023; Bavishi et al., 2023; Li et al., 2024) provides new opportu-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

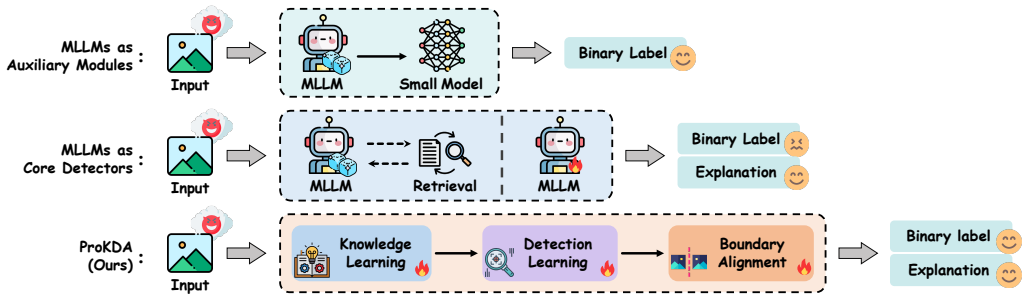



Figure 1: Comparison of ProKDA with existing MLLM-based methods.

nities for hateful meme detection. Their visual understanding and cross-modal reasoning abilities enable them to generate explanations for hateful memes, which can assist human moderation and support human understanding of hateful content. However, memes often express hatred and abuse through metaphors (Xu et al., 2022; Hwang & Shwartz, 2023), sarcasm (Chang et al., 2024), and other implicit forms, making both detection and explainability challenging. Recent studies attempt to build an explain-then-detect reasoning paradigm through chain-of-thought reasoning (Mei et al., 2026; Kmainasi et al., 2026; Pan et al., 2026), in which models generate both explanations and detection labels. These methods further optimize classification accuracy and explanation quality through supervised fine-tuning (SFT) (Ouyang et al., 2022) and reinforcement learning (RL) techniques such as Group Relative Policy Optimization (GRPO) (DeepSeek-AI, 2025). However, this paradigm essentially couples explanation generation and label prediction within the same training process. Such coupling causes interference between the two task objectives (Xu et al., 2026b; Li et al., 2025), thereby affecting both detection performance and explanation quality.

To address these issues, we propose ProKDA, a **Pro**gressive **K**nowledge-to-**D**ecision **A**lignment method. The method is inspired by the human annotation training process: annotators usually first study annotation guidelines and necessary background knowledge, then annotate samples, and further review ambiguous or difficult cases (Zhang et al., 2021). Based on this idea, ProKDA first uses an agentic background knowledge construction pipeline to generate relevant background knowledge for each training meme. We then design a three-stage progressive training strategy. The first stage learns background knowledge and its relation to meme expressions, the second stage adapts the model to hateful meme classification through SFT, and the third stage applies confidence-based sample filtering and Direct Preference Optimization (DPO) (Rafailov et al., 2023) to refine the detection boundary. Unlike existing explain-then-detect training paradigms, our strategy assigns a single training objective to each stage, reducing interference between explanation generation and label prediction. During DPO, we construct preference samples directly from binary labels without complex reward design, while confidence-based filtering reduces training samples and focuses the model on boundary cases. In summary, our contributions are as follows:

- We propose a progressive knowledge-to-decision alignment strategy for explainable hateful meme detection, improving both detection performance and explanation quality.
- Our method simulates the training process of human annotators by progressively performing background knowledge learning, hatefulness detection learning, and DPO-based boundary alignment on confidence-filtered samples, mitigating task interference between explanation generation and label prediction.
- Experimental results show that our method outperforms both direct detection and explain-then-detect paradigms, achieving state-of-the-art performance on three hateful meme benchmarks.

## 2 Related Work

Existing MLLM-based methods can be grouped into two categories: auxiliary-module methods that provide explanations or knowledge for classifiers, and core-detector methods that directly perform reasoning and prediction. Their comparison with ProKDA is illustrated in Figure 1.

**MLLMs as Auxiliary Modules.** These methods typically use MLLMs as interpreters or knowledge generators to support downstream classifiers (Mei et al., 2024; 2025; Tzelepi & Mezaris, 2025). For example, Lin et al. (2023) distill multimodal reasoning knowledge from large models, while Lin et al. (2024) use MLLM-based debate to generate harmfulness rationales. Kumari et al. (2024) obtain multi-level meme analyses through multi-hop prompting, and Garg et al. (2025) convert MLLM-generated explanations into a knowledge graph for classifier enhancement. Hee & Lee (2025) further use a frozen MLLM to generate explanatory analyses and combine them with meme representations for hateful meme detection. Although these studies improve detection performance, they do not provide direct explanations for meme predictions, leaving their explainability limited (Lin et al., 2026).

**MLLMs as Auxiliary Modules.** Recent studies use MLLMs as core units for hateful meme reasoning through prompting, agent frameworks, or parameter-efficient fine-tuning (Rizwan et al., 2026; Liu et al., 2026; Giorgis et al., 2026). For example, Jha et al. (2024) combines an MLLM and an LLM for harmful meme understanding and intervention generation. Cao et al. (2024) use task-specific LoRA modules (Hu et al., 2021) for few-shot detection, while Huang et al. (2024) and Lang et al. (2026) improve MLLM-based detection through retrieval, self-reflection, contrastive samples, or agent experience summarization. Recent studies further introduce RL for joint detection and explanation optimization. Kmainasi et al. (2026) use SFT followed by GRPO to optimize label prediction and explanation quality, while Mei et al. (2026) combine SFT warm-up, curriculum-based GRPO, and CDE rewards to further improve performance.

Although these methods use MLLMs as core detectors, they still have limitations. Prompting- and agent-based methods rely heavily on the native reasoning ability of MLLMs, making performance sensitive to model capacity and prompt design. SFT- and RL-based methods usually optimize explanation generation and label prediction together, which can cause task interference and make it difficult to balance detection performance and explanation quality (Xu et al., 2026a). To address these issues, we propose a progressive knowledge-to-decision alignment strategy that separates knowledge learning, detection learning, and boundary alignment into different stages, reducing task interference while improving both detection performance and explainability.

## 3 PRELIMINARIES

**Problem Statement.** Given a hateful meme dataset $\mathcal{D} = \{(I_i, y_i^*)\}_{i=1}^N$ where $I_i$ denotes the $i$-th meme image and $y_i^*$ denotes the ground-truth label, we define hateful meme detection as a binary classification task. The model is required to generate a textual label $y_i^* \in \{\texttt{yes}, \texttt{no}\}$, where `yes` indicates hateful and `no` indicates benign (Lin et al., 2024).

**Supervised Fine-Tuning (SFT).** Given an SFT training set $\mathcal{D}_{\text{sft}} = \{(I_i, p_i, o_i^*)\}_{i=1}^N$, where $I_i$ denotes the $i$-th meme image, $p_i$ denotes the input prompt, and $o_i^*$ denotes the desired target text. Let the MLLM be $\pi_\theta$. SFT minimizes the negative log-likelihood of the target output:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(I_i,p_i,o_i^*)\sim\mathcal{D}_{\text{sft}}} \sum_{t=1}^{|o_i^*|} \log P_\theta(o_{i,t}^* \mid o_{i,<t}^*, I_i, p_i), \tag{1}$$

where $o_{i,t}^*$ denotes the $t$-th token in the target output $o_i^*$, and $o_{i,<t}^*$ denotes the tokens before it.

**Direct Preference Optimization (DPO).** Given a preference dataset $\mathcal{D}_{\text{dpo}} = \{(x_i, y_i^w, y_i^l)\}_{i=1}^N$, where $x_i$ denotes the input, $y_i^w$ denotes the preferred response, and $y_i^l$ denotes the non-preferred response. Let the reference model be $\pi_{\text{ref}}$. DPO optimizes the model by comparing the relative probabilities of $y_i^w$ and $y_i^l$:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \mathcal{D}_{\text{dpo}}) = -\mathbb{E}_{(x_i,y_i^w,y_i^l)\sim\mathcal{D}_{\text{dpo}}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_i^w \mid x_i)}{\pi_{\text{ref}}(y_i^w \mid x_i)} - \beta \log \frac{\pi_\theta(y_i^l \mid x_i)}{\pi_{\text{ref}}(y_i^l \mid x_i)} \right) \right], \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and $\beta$ controls the preference strength.

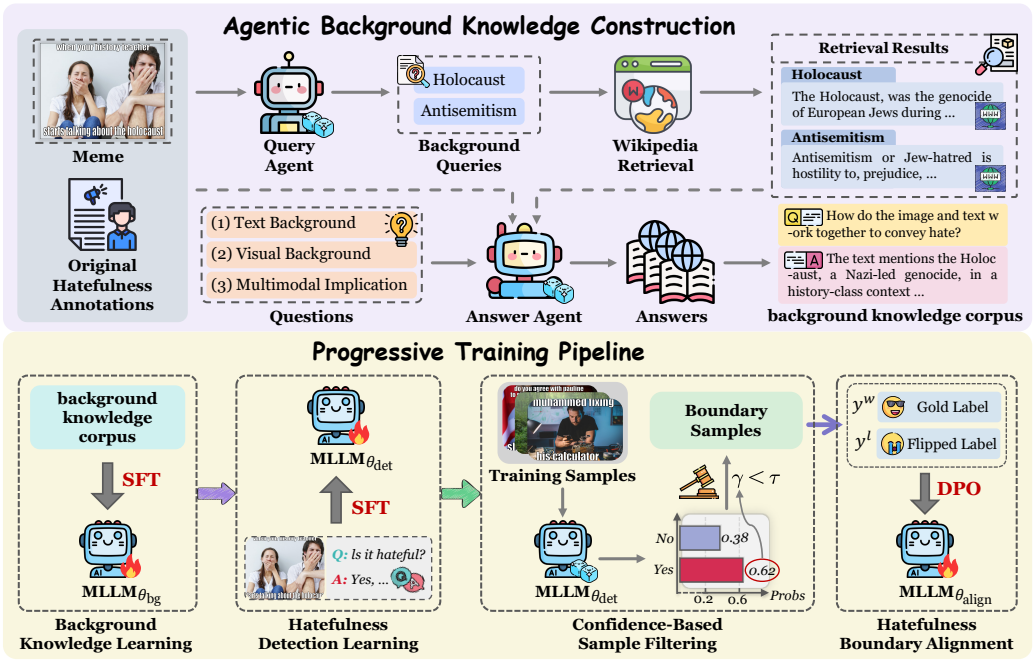



Figure 2: Overview of ProKDA. ProKDA first constructs a background knowledge corpus through an agentic pipeline, and then progressively optimizes the MLLM with knowledge learning, detection learning, confidence-based filtering, and DPO-based boundary alignment.

# 4 ProKDA Methodology

## 4.1 Agentic Background Knowledge Construction

Memes often rely on external background knowledge, such as historical events, public figures, and cultural symbols. For example, understanding "Holocaust" in Figure 2 requires knowledge of the event and its antisemitic context. Therefore, we design an agentic pipeline to construct background knowledge corpus for meme understanding.

Given the $i$-th meme sample, its image, text, and original hatefulness annotation are denoted as $I_i$, $T_i$, and $h_i$, respectively. Here, $h_i$ contains the hatefulness label and the attack target. First, a query agent $A_{\mathrm{q}}$ generates background knowledge queries for the current meme:

$$S_i = A_{\mathrm{q}}(I_i, T_i, h_i). \tag{3}$$

Here, $S_i$ contains one to two concise queries for retrieving key meme-related background concepts from Wikipedia:

$$R_i = \mathrm{Retrieve}(S_i). \tag{4}$$

We then use an answer agent $A_{\mathrm{a}}$ to convert retrieved content into structured question-answer pairs. $A_{\mathrm{a}}$ takes $I_i$, $T_i$, $h_i$, and $R_i$ as input, and answers three questions: what the text conveys, what the visual content conveys, and how the image and text jointly express the meaning. Let these three questions form a set $\mathcal{C} = \{c_{\mathrm{text}}, c_{\mathrm{vis}}, c_{\mathrm{mm}}\}$. The corresponding answer is:

$$a_i^j = A_{\mathrm{a}}(I_i, T_i, h_i, R_i, c_j), c_j \in C. \tag{5}$$

We compose the three answers into the background knowledge of the current meme:

$$b_i = \mathrm{Compose}(a_i^{\mathrm{text}}, a_i^{\mathrm{vis}}, a_i^{\mathrm{mm}}). \tag{6}$$

Here, $b_i$ denotes the background knowledge of the $i$-th meme, which connects textual and visual meanings to support the final judgment. The constructed background knowledge corpus is:

$$\mathcal{D}_{\mathrm{bg}} = \{(I_i, T_i, h_i, b_i)\}_{i=1}^{N}. \tag{7}$$

### 4.2 PROGRESSIVE KNOWLEDGE-TO-DECISION TRAINING

**Background Knowledge Learning.** Based on the background knowledge corpus $\mathcal{D}_{\mathrm{bg}}$, ProKDA first performs background knowledge learning on the initial MLLM $\theta_0$. For the $i$-th sample, we format its original annotation $h_i$ and background knowledge $b_i$ into the target response $y_i^{\mathrm{bg}} = \mathrm{Format}(h_i, b_i)$. The model takes the meme and the task prompt as input, and learns to generate $y_i^{\mathrm{bg}}$ through SFT. This stage enables the model to acquire meme-related background knowledge and understand how it supports the final hateful or non-hateful decision. The resulting model $\theta_{\mathrm{bg}}$ is used for subsequent hatefulness detection learning.

**Hatefulness Detection Learning.** Initialized from $\theta_{\mathrm{bg}}$, this stage uses SFT to adapt the model to the binary detection task, producing the hatefulness detection model $\theta_{\mathrm{det}}$. The model directly outputs a concise hatefulness judgment instead of detailed explanations, converting background understanding into detection capability for subsequent boundary-sample alignment.

**Confidence-Based Sample Filtering.** ProKDA selects boundary samples for subsequent DPO training based on model confidence. This process simulates the further decision review of ambiguous or uncertain samples in human annotation. We use the model $\theta_{\mathrm{det}}$ to perform inference on the whole training set, and identify low-confidence samples based on the prediction distribution over hateful and benign labels. Given a meme sample $M_i = (I_i, T_i)$, the model outputs the logits for hateful and benign judgments given the detection prompt $p_{\mathrm{det}}$:

$$\mathbf{z}_i = F_{\theta_{\mathrm{det}}}(I_i, T_i, p_{\mathrm{det}}), \tag{8}$$

where $\mathbf{z}_i = [z_i^{\mathrm{yes}}, z_i^{\mathrm{no}}]$ denotes the logits for the two labels. The label probabilities are obtained by softmax:

$$[p_i^{\mathrm{yes}}, p_i^{\mathrm{no}}] = Softmax([z_i^{\mathrm{yes}}, z_i^{\mathrm{no}}]). \tag{9}$$

The prediction confidence of the model on sample $M_i$ is defined as:

$$\gamma_i = max(p_i^{\mathrm{yes}}, p_i^{\mathrm{no}}). \tag{10}$$

We set $\tau \in [0.5, 1]$. Samples with $\gamma_i \geqslant \tau$ are classified with high confidence, while samples with $\gamma_i < \tau$ are treated as boundary samples, since they lie near the current decision boundary. Therefore, we construct the boundary sample set as:

$$\mathcal{D}_{\mathrm{bd}} = \{M_i \in \mathcal{D}_{\mathrm{train}} \mid \gamma_i < \tau\}. \tag{11}$$

$\mathcal{D}_{\mathrm{bd}}$ is used in the hatefulness boundary alignment stage, allowing the model to focus only on ambiguous samples that are difficult to judge reliably.

**Hatefulness Boundary Alignment.** Based on the boundary sample set $\mathcal{D}_{\mathrm{bd}}$, ProKDA applies DPO to refine the hatefulness decision boundary. This stage calibrates the model on low-confidence samples, improving its ability to handle ambiguous cases. Since hateful meme detection is a binary classification task, we directly construct preference pairs from gold labels. For each $M_i \in \mathcal{D}_{\mathrm{bd}}$, the gold-label response is used as the preferred response $y_i^w$, while the flipped-label response is used as the non-preferred response $y_i^l$. We then obtain the DPO training set:

$$\mathcal{D}_{\mathrm{dpo}} = \left\{ \left(M_i, p_{\mathrm{det}}, y_i^w, y_i^l\right) \mid M_i \in \mathcal{D}_{\mathrm{bd}} \right\}. \tag{12}$$

Starting from $\theta_{\mathrm{det}}$, we train the model on $\mathcal{D}_{\mathrm{dpo}}$ with the DPO objective to obtain the final model $\theta_{\mathrm{align}}$. This stage further calibrates the decision boundary on low-confidence samples.

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP

**Datasets and Tasks.** We evaluate ProKDA on three meme classification datasets: HMC, MAMI, and PrideMM (Shah et al., 2024). HMC covers different types of hateful memes, MAMI focuses on misogynous memes, and PrideMM focuses on LGBTQ+-related stances. We conduct binary classification on all datasets, classifying each meme as hateful or benign. We also evaluate model-generated explanations on HMC, since it provides human explanation annotations through HatReD (Hee et al., 2023). Detailed dataset descriptions and statistics are provided in Appendix B.

**Evaluation Metrics.** For classification, we follow existing work (Lu et al., 2025) and use accuracy, macro-F1, and weighted-F1 as metrics. For explanation quality, we use LLM-as-a-judge (Wang et al., 2023; Yang et al., 2023) and human evaluation to assess model-generated explanations against human references. Detailed evaluation settings are provided in Appendix E, and the results of explanation quality evaluation are reported in Section 5.6.

### 5.2 Baseline Models

Table 1 compares ProKDA with a comprehensive set of baselines based on Qwen2.5-VL-3B and Qwen2.5-VL-7B (Qwen Team, 2025). We briefly describe the baseline settings below, with implementation details and ProKDA prompts provided in Appendices C and D.

**SFT & DPO.** We build two SFT baselines to compare the direct detection and explain-then-detect paradigms. SFT (Explain-then-Detect) uses the background knowledge explanations constructed in Section 4.1 and gold labels as target responses, while SFT (Direct Detection) uses only gold labels for direct binary prediction. For DPO, we use the gold-label and flipped-label responses as the preferred and non-preferred responses, respectively. All SFT and DPO baselines use the same hyperparameter settings as ProKDA.

**MLLMs as auxiliary modules.** We compare ProKDA with representative methods that use MLLMs to provide auxiliary reasoning or knowledge for detection. ExplainHM (Lin et al., 2024) uses MLLM debates to generate reasoning for a small language model detector, while M2KE (Lu et al., 2025) uses multiple agents to obtain high-quality rationales as knowledge-enhanced information for the subsequent detector.

**MLLMs as core detectors.** We compare ProKDA with training-free and training-based methods that use MLLMs as core detectors. Training-free methods, including U-CoT+ (Pan et al., 2026), LoReHM (Huang et al., 2024), MIND (Liu et al., 2025), and ALARM (Lang et al., 2026), rely on prompting, retrieval, reflection, or agent experience summarization. Training-based methods, including MemeReason (Kmainasi et al., 2026), RA-HMD (Mei et al., 2025), and ExPO-HM (Mei et al., 2026), further optimize MLLMs through SFT, RL, or reasoning-quality rewards. We exclude closed-source models such as the OpenAI series because safety filtering may block responses to highly harmful memes.

### 5.3 Comparing ProKDA to Baselines

Table 1 compares ProKDA with existing methods on HMC, MAMI, and PrideMM. Background knowledge quality analysis is provided in Appendix F, and qualitative analysis and error cases are provided in Appendix H. We summarize the key observations below.

**Explain-then-detect methods hurt classification performance.** We first compare the two SFT baselines. SFT (Explain-then-Detect) consistently underperforms SFT (Direct Detection), indicating that jointly optimizing explanation generation and label prediction introduces task interference and hurts classification performance. Similarly, training-free explain-then-detect methods (#3-#6) also underperform the task-specific trained method RA-HMD. These results suggest that simply adding explicit rationales through CoT prompting or explanation-supervised post-training does not reliably improve hateful meme detection.

**Task-specific training is more important than model scaling.** Training-free methods generally underperform training-based methods when MLLMs are used as core detectors. Although ALARM uses Qwen2.5-VL-72B, its M-F1 on HMC is still 5.31 lower than ExPO-HM trained on Qwen2.5-VL-7B. The zero-shot results of Qwen2.5-VL-3B and Qwen2.5-VL-7B (#11, #17) are also clearly lower than their SFT and DPO versions. These results show that scaling alone is insufficient, and task-specific training remains crucial.

**Gold-label-only post-training shows a clear upper bound.** SFT (Direct Detection) and DPO improve over zero-shot inference on all three datasets, but still lag behind stronger training-based methods such as MemeReason and RA-HMD. The small gap between SFT (Direct Detection) and

Table 1: Comparison between ProKDA and baseline methods. Accuracy (ACC), macro-F1 (M-F1), and weighted-F1 (W-F1) are used as evaluation metrics (%). Bold and underlined values denote the best and second-best results, respectively.

| # | Model | HMC | | | MAMI | | | PrideMM | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | M-F1 | W-F1 | Acc | M-F1 | W-F1 | Acc | M-F1 | W-F1 |
| | *MLLMs as Auxiliary Modules* | | | | | | | | | |
| 1 | ExplainHM (LLaVA-v1.5-13B) | 75.60 | 75.39 | - | - | - | - | - | - | - |
| 2 | M2KE (LLaVA-v1.5-13B) | 75.76 | 75.62 | - | 75.85 | 75.71 | - | - | - | - |
| | *MLLMs as Core Detectors (Training-Free)* | | | | | | | | | |
| 3 | U-CoT+ (Qwen2.5-14B) | 72.50 | 72.41 | - | 79.90 | 79.89 | - | 71.60 | 71.37 | - |
| 4 | LoReHM (LLaVA-34B) | 65.60 | 65.59 | - | 75.40 | 75.28 | - | - | - | - |
| 5 | MIND (LLaVA-v1.6-34B) | 66.40 | 68.38 | - | 73.60 | 75.38 | - | - | - | - |
| 6 | ALARM (Qwen2.5-VL-72B) | 75.80 | 75.79 | - | <u>81.28</u> | 81.25 | - | - | - | - |
| | *MLLMs as Core Detectors (Training-Based)* | | | | | | | | | |
| 7 | MemeReason (Qwen3-VL-8B) | <u>81.20</u> | 79.00 | <u>81.00</u> | - | - | - | - | - | - |
| 8 | RA-HMD (Qwen2.5-VL-7B) | 80.80 | 80.10 | - | 81.00 | 81.00 | - | <u>78.00</u> | 77.80 | - |
| 9 | ExPO-HM (Qwen2.5-VL-7B) | - | <u>81.10</u> | - | - | <u>82.30</u> | - | - | <u>78.70</u> | - |
| 10 | Qwen2.5-VL-3B | | | | | | | | | |
| 11 | *Zero-shot* | 63.10 | 61.99 | 62.12 | 61.80 | 57.44 | 57.44 | 65.09 | 64.24 | 64.38 |
| 12 | *SFT (Explain-then-Detect)* | 67.00 | 65.22 | 65.38 | 69.70 | 68.63 | 68.63 | 66.86 | 66.81 | 66.84 |
| 13 | *SFT (Direct Detection)* | 72.20 | 71.07 | 71.18 | 77.50 | 76.99 | 76.99 | 74.36 | 74.35 | 74.36 |
| 14 | *DPO* | 71.60 | 70.24 | 70.36 | 76.40 | 75.65 | 75.65 | 74.75 | 74.75 | 74.73 |
| 15 | *ProKDA* | 80.20 | 80.19 | 80.20 | 78.70 | 78.13 | 78.13 | 75.74 | 75.72 | 75.70 |
| 16 | Qwen2.5-VL-7B | | | | | | | | | |
| 17 | *Zero-shot* | 63.40 | 60.19 | 60.42 | 56.30 | 47.31 | 47.31 | 58.19 | 51.10 | 51.58 |
| 18 | *SFT (Explain-then-Detect)* | 72.00 | 70.93 | 71.04 | 75.00 | 74.34 | 74.34 | 76.53 | 76.48 | 76.51 |
| 19 | *SFT (Direct Detection)* | 75.60 | 75.15 | 75.22 | 78.90 | 78.37 | <u>78.37</u> | 77.71 | 77.71 | <u>77.72</u> |
| 20 | *DPO* | 74.30 | 73.65 | 73.73 | 77.30 | 76.60 | 76.60 | 76.53 | 76.53 | 76.53 |
| 21 | *ProKDA* | **83.50** | **83.41** | **83.44** | **83.10** | **83.08** | **83.08** | **80.28** | **80.27** | **80.28** |

DPO suggests that gold-label-only optimization reaches a limited performance ceiling. Moreover, these methods focus on final labels rather than faithful explanations for memes, limiting their explainability in real-world moderation.

**ProKDA achieves stable improvements across datasets and model scales.** ProKDA based on Qwen2.5-VL-7B achieves the best overall performance on all three datasets, outperforming ExPO-HM and other baselines. On HMC, it improves W-F1 by 10.9% and 13.2% over the corresponding SFT (Direct Detection) and DPO versions, respectively. ProKDA based on Qwen2.5-VL-3B also consistently outperforms its SFT and DPO counterparts and even surpasses ALARM on HMC. Increasing the model scale further strengthens ProKDA (#15→#21), confirming its effectiveness across datasets and model scales.

### 5.4 Ablation Study of ProKDA Components

We conduct an ablation study on the three training stages of ProKDA on HMC to analyze the contribution of each stage. The results are shown in Table 2.

**Background knowledge learning provides the basis for progressive detection.** Background knowledge learning (Stage 1) provides a key foundation for subsequent progressive training. With only Stage 1, the model already outperforms zero-shot inference. Applying Stage 2 on top of Stage 1 (#4) also performs better than using Stage 2 alone (#3), with W-F1 gains of 4.91 and 6.57 for the 3B and 7B models, respectively. Similarly, introducing Stage 1 into the Stage 2+3 setting (#6 vs. #5) further improves W-F1 by 4.33 and 3.13. These results show that Stage 1 helps the model acquire meme-related background knowledge and connect it with implicit semantics, background

references, and attack intentions, providing a stronger basis for hatefulness detection. This observation is also consistent with Mei et al. (2026).

**Hatefulness detection learning and boundary alignment refine the decision boundary.** Hatefulness detection learning (Stage 2) adapts the model to the detection task. Using only Stage 2 improves W-F1 by 9.06 and 14.8 over zero-shot inference for the 3B and 7B models, while applying it after Stage 1 (#4) brings gains of 7.70 and 7.18. Boundary alignment (Stage 3) further improves performance, with gains of 4.11 and 1.65 when added on top of Stage 1+2 (#6 vs. #4). These results show that Stage 3 refines the decision boundary and helps the model handle ambiguous samples.

Table 2: Ablation study of ProKDA components.

| # | Components | | | Qwen2.5-VL-3B | | | Qwen2.5-VL-7B | | |
|---|---|---|---|---|---|---|---|---|---|
| | Stage 1 | Stage 2 | Stage 3 | Acc | M-F1 | W-F1 | Acc | M-F1 | W-F1 |
| 1 | - | - | - | 63.10 | 61.99 | 62.12 | 63.40 | 60.19 | 60.42 |
| 2 | ✓ | - | - | 68.50 | 68.42 | 68.39 | 74.70 | 74.63 | 74.61 |
| 3 | - | ✓ | - | 72.20 | 71.07 | 71.18 | 75.60 | 75.15 | 75.22 |
| 4 | ✓ | ✓ | - | 76.50 | 76.02 | 76.09 | 81.90 | 81.76 | 81.79 |
| 5 | - | ✓ | ✓ | 76.20 | 75.80 | 75.87 | 80.40 | 80.27 | 80.31 |
| 6 | ✓ | ✓ | ✓ | 80.20 | 80.19 | 80.20 | 83.50 | 83.41 | 83.44 |

## 5.5 Analysis of Confidence-based Sample Filtering

We analyze confidence-based sample filtering through threshold effects and training-set confidence distributions. Additional analyses on confidence-interval accuracy and qualitative examples are provided in Appendix G.

**Effect of Confidence Thresholds on Boundary Alignment.** Figure 3 shows the results under different confidence thresholds $\tau$. The optimal threshold depends on model scale: Qwen2.5-VL-3B performs best at $\tau = 0.8$, while Qwen2.5-VL-7B performs best at $\tau = 0.6$. We believe this is related to how well the model learns the training samples during hatefulness detection learning. For the 7B model, most samples already have confidence scores much higher than 0.8, indicating that it has sufficiently learned most samples and Stage 3 only needs to focus on low-confidence boundary samples. In contrast, Qwen2.5-VL-3B has a more dispersed confidence distribution, suggesting that the smaller model still benefits from including more medium-confidence samples for boundary alignment. Moreover, using all samples for Stage 3 often degrades performance by distracting the model from difficult cases. Confidence-based filtering instead preserves ambiguous boundary samples for more effective alignment.

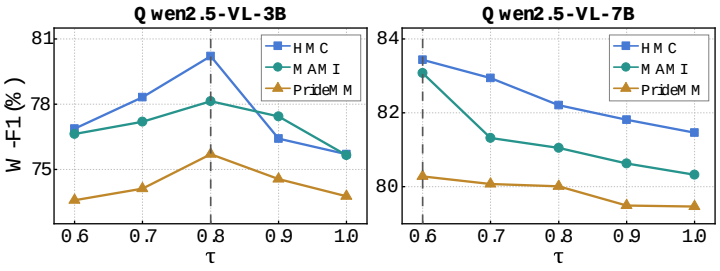



Figure 3: Performance under different confidence thresholds.

**Model Confidence Distributions on Training Sets.** Figure 4 shows the confidence distributions of Qwen2.5-VL-3B and Qwen2.5-VL-7B on the three training sets. The 7B model mainly concentrates in the [90,100] interval, while the 3B model is more dispersed. This suggests that the 7B model learns most samples more thoroughly, making low-confidence samples more indicative of boundary uncertainty. In contrast, the dispersed distribution of the 3B model suggests the need to include more medium-confidence samples for boundary alignment.

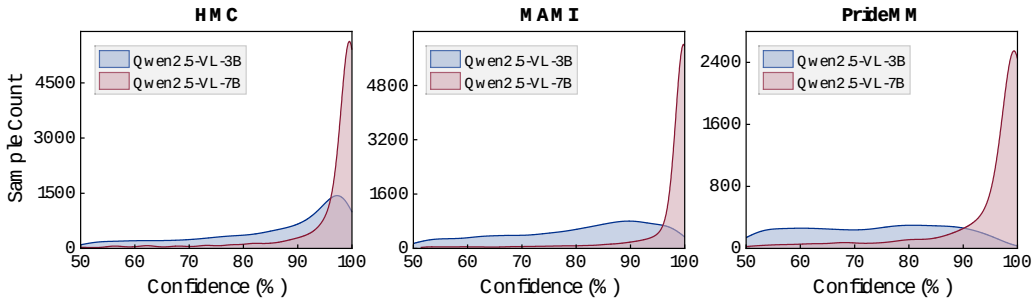



Figure 4: Model confidence distributions across different training sets.

Table 3: Explanation evaluation of hateful memes, including Informativeness (Info.), Readability (Read.), Soundness (Sound.), and Persuasiveness (Pers.).

| | GPT-4o mini | | | GPT-5 | | | Human | | |
|---|---|---|---|---|---|---|---|---|---|
| | Qwen | w/ Stage 1 | ProKDA | Qwen | w/ Stage 1 | ProKDA | Qwen | w/ Stage 1 | ProKDA |
| Info. | 2.93 | 4.08 | 4.25 | 2.20 | 3.62 | 3.83 | 2.35 | 3.74 | 3.94 |
| Read. | 4.34 | 4.62 | 4.78 | 4.50 | 4.73 | 4.87 | 4.47 | 4.75 | 4.79 |
| Sound. | 2.83 | 4.05 | 4.23 | 1.98 | 3.33 | 3.56 | 2.14 | 3.51 | 3.67 |
| Pers. | 2.73 | 4.05 | 4.21 | 2.08 | 3.53 | 3.74 | 2.39 | 3.82 | 4.06 |

### 5.6 Explainability Evaluation

In this section, we evaluate the quality of model-generated meme explanations. Since textual explanations are diverse and have no fixed standard answer, we follow (Lin et al., 2024; Lu et al., 2025) and use both LLM-as-a-judge and human evaluation. We evaluate four dimensions: informativeness, readability, soundness, and persuasiveness, each scored on a 5-point Likert scale (see Appendix E.4 for detailed prompts). GPT-4o mini and GPT-5 serve as LLM judges, while three computer science master's students with experience in hateful meme annotation conduct human evaluation. We evaluate Qwen2.5-VL-7B on the HMC test set with human gold explanation annotations (Hee et al., 2023), comparing the original model, the model after background knowledge learning (w/ Stage 1), and ProKDA. The results are shown in Table 3.

**ProKDA consistently improves the quality of model-generated explanations.** Based on Table 3, we make the following observations. First, w/ Stage 1 outperforms the original model on all four dimensions, showing that background knowledge learning improves meme understanding. Second, both w/ Stage 1 and ProKDA achieve clear gains in informativeness, soundness, and persuasiveness, indicating that the agentic background knowledge construction pipeline provides useful contextual evidence for explaining implicit hateful meanings. Third, the improvement in readability is relatively small, because the original model already has strong text generation ability. Finally, ProKDA further outperforms w/ Stage 1 on all four dimensions. This suggests that background knowledge alone is not enough: without explicit hatefulness detection learning, explanations may fail to focus on hateful intent. ProKDA addresses this by aligning background knowledge with detection decisions, making explanations more consistent with hateful intent and human references.

## 6 Conclusion

In this paper, we propose ProKDA, a progressive knowledge-to-decision alignment method for explainable hateful meme detection. ProKDA combines agentic background knowledge construction with three-stage progressive fine-tuning to align meme understanding with detection decisions. Experiments show that ProKDA achieves the best detection performance on multiple datasets and generates high-quality explanations. We discuss limitations and future work in Appendix I.

## Reproducibility Statement

We provide the LLM usage statement in Appendix A. Detailed dataset usage and statistics are provided in Appendix B. The experimental settings, implementation details, software versions, hardware configurations, and hyperparameter settings are described in Appendix C to support reproducibility. The prompts used in ProKDA and the evaluation prompts are provided in Appendices D and E, respectively. Additional analyses, including background knowledge quality analysis, confidence-based sample filtering analysis, and qualitative analysis, are reported in Appendices F, G, and H. Limitations, future work, and ethical considerations are discussed in Appendices I and J. The related implementation is provided in the following anonymous repository for reference: `https://anonymous.4open.science/r/ProKDA-Code-8544`.

## References


Mohammad Alafnan. The role of memes in shaping political discourse on social media. *Studies in Media and Communication*, 13:1, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Mapping memes to words for multimodal hateful meme classification. *arXiv preprint arXiv:2310.08368*, 2023.

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

Carlos Castaño Díaz. Defining and characterizing the concept of internet meme. *Revista CES Psicología*, 6:82–104, 2013.

Ke Chang, Hao Li, Junzhao Zhang, and Yunfang Wu. Nyk-ms: A well-annotated multi-modal metaphor and sarcasm understanding benchmark on cartoon-caption dataset. *arXiv preprint arXiv:2409.01037*, 2024.

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

Barbara Dancygier and Lieven Vandelanotte. Internet memes as multimodal constructions. *Cognitive Linguistics*, 28:565–598, 2017.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

DeepSeek-AI. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025.

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

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, and Ponnurangam Kumaraguru. Just KIDDIN' : Knowledge infusion and distillation for detection of INdecent memes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23067–23086, 2025.

Stefano De Giorgis, Ting-Chih Chen, and Filip Ilievski. M-quest – meme question-understanding evaluation on semantics and toxicity. *arXiv preprint arXiv:2603.03315*, 2026.

Ming Shan Hee and Roy Ka-Wei Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. *arXiv preprint arXiv:2502.11073*, 2025.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 5995–6003, 2023.

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

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

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1433–1445, 2023.

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. MemeGuard: An LLM and VLM-based framework for advancing content moderation via meme intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 8084–8104, 2024.

Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Isabella Garcia-Camargo, Emma S. Spiro, and Kate Starbird. Repeat spreaders and election delegitimization: A comprehensive dataset of misinformation tweets from the 2020 u.s. election. *Journal of Quantitative Description: Digital Media*, 2, 2022.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 2611–2624, 2020.

Mohamed Bayan Kmainasi, Mucahid Kutlu, Ali Ezzat Shahroor, Abul Hasnat, and Firoj Alam. Can thinking models think to detect hateful memes? In *Companion Proceedings of the ACM Web Conference 2026*, pp. 935–944, 2026.

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

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 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*, pp. 22105–22138, 2024.

Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. From shallow humor to metaphor: Towards label-free harmful meme detection via lmm agent self-improvement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 588–599, 2026.

Dita Lestari, Megi Primagara, Shinta Sari, Afifah Meilina, Silfiah Fauziah, Ade Sugesti, Aisyah Nasywa, and Ahmad Salwi. Meme culture: A study of humor and satire in digital media. *International Journal of Advanced Multidisciplinary Research and Studies*, 4:134–140, 2024.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 11214–11232, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9114–9128, 2023.

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

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Trans. Intell. Syst. Technol.*, 17:1–25, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024.

Olivia Shanhong Liu, Pai Chet Ng, De Wen Soh, and Konstantinos N. Plataniotis. Yes florence, i will do better next time! agentic feedback reasoning for humorous meme detection. *arXiv preprint arXiv:2601.07232*, 2026.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. MIND: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 923–947, 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. Is having rationales enough? rethinking knowledge enhancement for multimodal hateful meme detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 559–569, 2025.

Delfina S. Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 'toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47:100317, 2025.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 5333–5347, 2024.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 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*, pp. 23806–23828, 2025.

Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Pengda Qin, Yuhong Li, Da Chen, and Bill Byrne. ExPO-HM: Learning to Explain-Then-Detect for Hateful Meme Detection. In *International Conference on Learning Representations*, 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 27730–27744, 2022.

Fengjun Pan, Xiaobao Wu, Tho Quan, and Anh Tuan Luu. Read as you see: Guiding unimodal llms for low-resource explainable harmful meme detection. In *Companion Proceedings of the ACM Web Conference 2026*, pp. 1672–1683, 2026.

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

Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Naquee Rizwan, Subhankar Swain, Paramananda Bhaskar, Gagan Aryan, Shehryaar Shah Khan, and Animesh Mukherjee. See, explain, and intervene: A few-shot multimodal agent framework for hateful meme moderation. *arXiv preprint arXiv:2601.04692*, 2026.

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

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58:102664, 2021.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5597–5606, 2022.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 32–41, 2020.

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

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Evaluating gpt-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6255–6263, 2023.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2887–2899, 2022.

Hefei Xu, Le Wu, Chen Cheng, and Hao Liu. Multi-value alignment for LLMs via value decorrelation and extrapolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 34133–34141, 2026a.

Hefei Xu, Le Wu, Yu Wang, Min Hou, Han Wu, Zhen Zhang, and Meng Wang. Vc-soup: Value-consistency guided multi-value alignment for large language models. In *Proceedings of the ACM Web Conference 2026*, pp. 9699–9710, 2026b.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5490–5505, 2023.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3214–3225, 2021.

# A LLM USAGE STATEMENT

In the process of developing this work, we utilized LLMs for grammar correction and occasionally as a rewriting tool.

# B DATASET DETAILS AND STATISTICS

## B.1 DATASET DETAILS

**HMC.** The Hateful Memes Challenge (HMC) (Kiela et al., 2020) is released by Meta AI and contains 10,000 meme samples for binary hate speech detection. The dataset covers attacks against protected groups, such as race, religion, gender, and disability. It also constructs benign confounders by modifying the image or text to convert hateful samples into non-hateful ones, thereby challenging the model's ability to understand image-text interactions and multimodal hateful intent.

**MAMI.** Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022) is a multimodal misogynous meme detection dataset released in SemEval-2022 Task 5. It contains 10,000 training samples and 1,000 test samples. The dataset requires models to determine whether a meme is misogynous and further identify misogyny types, such as shaming, stereotype, objectification, and violence. Since the discriminatory meanings are often implicitly expressed through images, text, and their interactions, MAMI is widely used to evaluate the model's ability to understand multimodal gender-discriminatory content.

**PrideMM.** PrideMM (Shah et al., 2024) is a multimodal meme dataset for the LGBTQ+ Pride movement, containing 5,063 memes collected from Facebook, Twitter, and Reddit. The dataset is annotated for four tasks: hate detection, hate target classification, stance classification, and humor detection. The hate targets include undirected, individual, community, and organization. PrideMM focuses on hate, stance, and humor expressions in LGBTQ+-related online contexts, making it suitable for evaluating the model's ability to understand complex multimodal meme content.

## B.2 DATASET STATISTICS

**Binary Classification Statistics.** Table 4 shows the dataset statistics of HMC, MAMI, and PrideMM.

**Attack Targets.** For HMC, Table 5 reports the statistics of social categories, and we treat these social categories as attack targets. For MAMI, since the dataset focuses on misogynous content, we regard women as the attack target. For PrideMM, Table 6 reports the statistics of hate targets, and we use the annotated hate target categories as attack targets.

**Human Explanation Annotation Statistics.** For human-annotated meme explanation data, we use HatReD (Hee et al., 2023), which provides human rationale annotations for hateful memes in HMC. The training and test sets of HatReD are from the `train` split and `dev_seen` split of HMC, respectively. The detailed statistics are shown in Table 7.

Table 4: Statistical summary of datasets.

| | HMC | | MAMI | | PrideMM | |
|---|---|---|---|---|---|---|
| | #Benign | #Hate | #Benign | #Hate | #Benign | #Hate |
| Train | 5450 | 3050 | 5000 | 5000 | 2208 | 2120 |
| Test | 500 | 500 | 500 | 500 | 260 | 247 |

Table 5: Statistics of social categories in HMC.

| Social Category | train | dev_unseen | dev_seen |
|---|---|---|---|
| Religion | 1078 | 77 | 95 |
| Race | 1008 | 63 | 78 |
| Sex | 746 | 46 | 56 |
| Nationality | 325 | 20 | 26 |
| Disability | 255 | 17 | 22 |

Table 6: Statistics of hate targets in PrideMM.

| Hate Targets | Train | Test |
|---|---|---|
| Benign | 2208 | 260 |
| Undirected | 666 | 68 |
| Individual | 219 | 19 |
| Community | 986 | 122 |
| Organization | 249 | 38 |

Table 7: Dataset statistics of HatReD. HatReD only includes explanations for hateful memes, with the training and test sets derived from the `train` and `dev_seen` splits of HMC, respectively.

| | HatReD |
|---|---|
| Train | 2982 |
| Test | 246 |
| Total | 3228 |

## C Experiment Setup and Implementation Details

All experiments are conducted on two NVIDIA RTX 4090 GPUs, including zero-shot evaluation, SFT, DPO, and ProKDA. To ensure fair comparison, we keep the training settings consistent with ExPO-HM (Mei et al., 2026). Except for zero-shot evaluation, all fine-tuning experiments use LoRA, with $r = 64$ and $\alpha = 128$. We use LLaMA-Factory 0.9.5 for SFT and DPO training, and follow its official settings for other hyperparameters. All training tasks are run for 3 epochs, and the best checkpoint is selected based on validation performance.

For SFT (Explain-then-Detect), the target response follows the `<think>...</think> <answer>...</answer>` format, with the explanation in `<think>` and the final label in `<answer>`. The detailed evaluation prompt is provided in Appendix E.2.

**ProKDA.** In the agentic background knowledge construction stage, we use GPT-4o to implement the two agents. For the hyperparameter $\tau$, we set it to 0.8 for Qwen2.5-VL-3B and 0.6 for Qwen2.5-VL-7B. For baseline models, we mainly report the results from their original papers on the same datasets to ensure a complete comparison.

## D Detailed Prompt Design in ProKDA

In the agentic background knowledge construction stage, the prompts for the query agent and the answer agent are as follows:

**Prompt 1.** *The Prompt for Query Agent*

• Instruction:

You are a Wikipedia search query generator for multimodal meme analysis. Given a meme image, meme text, hateful label, and attack target(s), your task is to generate one or two concise search queries that can retrieve the most relevant background knowledge from Wikipedia.

Inputs:
- Meme image: <image>
- Meme text: {meme_text}
- Hateful label: {hateful_label}
- Attack target(s): {attack_target}

Requirements:
1. The queries should capture the key background knowledge needed to understand the hateful implication of the meme.
2. The queries should reflect the relationship between the attack target(s) and the stereotype, discrimination, dehumanization, or historical/cultural reference implied by the meme.
3. Prefer short Wikipedia-style keyword phrases rather than full sentences.
4. Prefer commonly used terms that are likely to match Wikipedia article titles or related concepts.
5. Avoid generic, overly broad, or overly long queries.
6. Do not simply repeat the meme text or the attack target.
7. Generate at most two queries, ordered by relevance.
8. Do not output explanations or reasoning.
9. Output json only.

Output format:
{
"search_queries": "[<query_1>, <query_2>]"
}

**Prompt 2.** *The Prompt for Answer Agent*

• Instruction:

You are a background knowledge QA generator for multimodal meme understanding. Given a meme image, meme text, hateful label, annotated attack target(s), and retrieved Wikipedia knowledge, your task is to generate three question-answer pairs that explain the meme from textual, visual, and multimodal perspectives.

Inputs:
- Meme image: <image>
- Meme text: {meme_text}
- Hateful label: {hateful_label}
- Attack target(s): {attack_target}
- Wikipedia knowledge: {wikipedia_knowledge}

Requirements:
1. Generate exactly three question-answer pairs.
2. The first QA pair should explain what the meme text conveys.
3. The second QA pair should explain what the visual content conveys.
4. The third QA pair should explain how the image and text interact to convey the overall meaning.
5. Use the retrieved Wikipedia knowledge only when it is relevant to understanding the meme.
6. The answers should be concise, informative, and grounded in the given annotation and retrieved knowledge.
7. Do not introduce unsupported external information.
8. Output JSON only.

Output format:
{
qa_pairs: [
{
"question": "What does the meme text convey?",
"answer": ""
},
{
"question": "What does the visual content convey?",
"answer": ""
},
{
"question": "How do the image and text jointly convey the overall meaning?",
"answer": ""
}
]
}

# E EVALUATION PROMPTS

## E.1 HATEFUL MEME DETECTION EVALUATION

For hateful meme detection evaluation, we use a unified evaluation prompt, as shown below:

**Prompt 3.** *The Prompt for Hateful Meme Detection Evaluation*

• Instruction:

Is it hateful? The answer should be Yes or No, if the meme is hateful than output Yes; if the meme is Benign than output No.

## E.2 EXPLAIN-THEN-DETECT EVALUATION

For SFT (Explain-then-Detect) evaluation, we use a separate prompt that follows the explain-then-detect format, as shown below:

**Prompt 4.** *The Prompt for Explain-then-Detect Evaluation*

• Instruction:

Is it hateful? Output the thinking process in `<think> </think>` and final answer in `<answer> </answer>` tags. The answer should be Yes or No, if the meme is hateful than output Yes ; if the meme is Benign than output No. The output format should be as follows: `<think>` ... `</think> <answer>...</answer>` Please strictly follow the format.

## E.3 BACKGROUND KNOWLEDGE QUALITY EVALUATION

In the background knowledge quality evaluation, we use LLM-as-a-judge to evaluate the meme explanations generated by ProKDA and GPT-4o. The evaluation prompt is designed as follows:

**Prompt 5.** *The Prompt for Background Knowledge Quality Evaluation*

• Instruction:

Your task is to evaluate which model provides a better explanation for the given meme.

You will be given:
1. The meme image.

2. A human reference rationale explaining why the meme is hateful.
3. Explanation A generated by Model A.
4. Explanation B generated by Model B.

Please compare Explanation A and Explanation B from the following two dimensions:
1. Informativeness: Evaluate which explanation provides more sufficient, specific, and relevant information.
2. Consistency with the human reference rationale: Evaluate which explanation is more consistent with the human reference rationale.

Human reference rationale:
{human_reference}

Explanation A:
{explanation_A}

Explanation B:
{explanation_B}

Output strictly in the following format:
Informativeness: [A/B/Tie]
Consistency with human reference rationale: [A/B/Tie]
Reason: [1-2 sentences]

### E.4 Meme Explanation Evaluation

In the explainability evaluation, we use the following prompt for LLM-as-a-judge evaluation:

**Prompt 6.** *The Prompt for Meme Explanation Evaluation*

• Instruction:

Your task is to evaluate how well the model-generated explanation semantically matches the human reference rationale for a hateful meme.

Evaluation dimensions:
1. Informativeness: Whether the model explanation provides sufficient and relevant information.
2. Readability: Whether the model explanation is fluent, grammatically correct, and clearly structured.
3. Soundness: Whether the model explanation is logically valid and consistent with the meme content and the human reference rationale.
4. Persuasiveness: Whether the model explanation convincingly explains why the meme is hateful.

Each dimension should be scored on a 1-to-5 Likert scale, where 1 indicates the poorest quality and 5 indicates the best quality.

Human reference rationale:
{human_reference}

Model-generated explanation:
{model_reasoning}

Output strictly in the following format:
Informativeness: [1-5]
Readability: [1-5]
Soundness: [1-5]

Persuasiveness: [1-5]

Explanation: [1-2 sentences]

### E.5 Qualitative Analysis Evaluation

In the qualitative evaluation, our prompt is designed as follows:

**Prompt 7.** *The Prompt for Qualitative Analysis Evaluation*

• Instruction:

Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

## F Background Knowledge Quality Analysis

We analyze the quality of the background knowledge generated in Section 4.1, with results shown in Figure 5. We randomly sample 100 hateful memes from HMC and compare ProKDA with GPT-4o, which directly generates explanations from memes and human labels. We evaluate the explanations using LLM-as-a-judge and human evaluation on informativeness and consistency with human explanations (Hee et al., 2023). For each sample, evaluators choose the better explanation or mark a tie. GPT-4o mini and GPT-5 serve as LLM judges, while two computer science master's students experienced in hateful meme annotation and evaluation conduct human evaluation. We report the average preference ratio and shuffle explanation order to avoid positional bias.

**Agentic background knowledge construction outperforms direct explanation generation.** ProKDA consistently outperforms GPT-4o in both informativeness and consistency. Under LLM-as-a-judge evaluation, ProKDA obtains preference ratios above 80% on both dimensions. Human evaluation shows the same trend, while the preference ratios of GPT-4o remain below 10%. The tie ratios exceed 10%, mainly because some memes contain limited information, allowing both methods to cover the necessary content. Overall, these results show that ProKDA generates more informative explanations that better align with human reference explanations, providing a solid foundation for subsequent training.

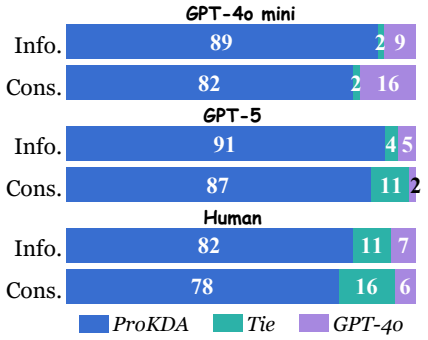



Figure 5: Preference-based evaluation of generated background knowledge.

## G Analysis of Confidence-based Sample Filtering

**Training-set Prediction Accuracy across Confidence Intervals.** Figure 6 shows the training-set accuracy of Qwen2.5-VL-3B and Qwen2.5-VL-7B across confidence intervals on the three datasets. Both models achieve the highest accuracy in the [90,100] interval, with most results above 95%, showing that high-confidence predictions are more reliable. In contrast, accuracy drops clearly in the [50,60) interval, indicating that low-confidence samples are difficult to distinguish. This trend shows that confidence reflects how well the model learns each sample and helps select difficult samples with higher training value.

**Qualitative Analysis of Samples across Confidence Levels.** We further analyze representative HMC examples from different confidence intervals using Qwen2.5-VL-7B. As shown in Figure 7, we randomly select one hateful and one benign meme from each of the low-confidence interval

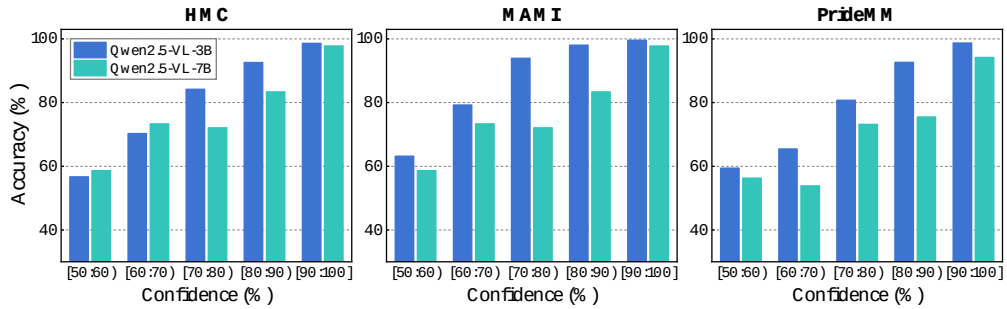



Figure 6: Training-set prediction accuracy across confidence intervals.

$\gamma_i < 0.6$ and the high-confidence interval $\gamma_i > 0.9$. High-confidence examples are easier to judge, while low-confidence examples are more ambiguous: the hateful meme requires joint reasoning over text, vision, and background knowledge, and the benign meme contains misleading sensitive words or attack cues. These examples suggest that low-confidence samples better reflect difficult boundary cases, supporting their use for hatefulness boundary alignment.




(a) $\gamma_i < 0.6$ (b) $\gamma_i > 0.9$

Figure 7: Examples of low- and high-confidence samples.

## H Qualitative Analysis of ProKDA

### H.1 Case Study

Figures 8–13 show representative cases where ProKDA corrects the mispredictions made by the baselines, demonstrating its advantages in both prediction accuracy and explanation quality.

In Figures 8–10, Qwen2.5-VL-7B fails to effectively perform joint reasoning over the meme text and visual content, and therefore fails to identify the hateful intent. In contrast, both the model with Stage 1 and ProKDA capture key background information and make correct predictions, showing that background knowledge learning helps improve the model's ability to understand memes.

In Figures 11–13, both Qwen2.5-VL-7B and w/ Stage 1 make incorrect predictions, indicating that recognizing only partial background information is still insufficient for understanding complex meme samples. In contrast, ProKDA further connects background knowledge, visual cues, and textual content, thereby accurately inferring the implicit hateful intent and generating more reasonable explanations.

### H.2 Error Analysis

When analyzing the cases misclassified by Qwen2.5-VL-7B, w/stage 1, and ProKDA, we find that these errors mainly stem from highly implicit hateful memes. Such samples usually do not contain explicit insulting words or clear attack targets. Instead, their hateful meanings rely on joint reasoning over textual implications, visual cues, and the overall interaction within the meme. Since the potential attack target is often hidden in visual identity, metaphor, or contextual associations, these models struggle to identify the attack intent and link it to a specific protected group. Therefore,

even ProKDA fails to correct these cases. These samples are inherently ambiguous, and human annotators may also misclassify them Kiela et al. (2020). Figures 14–16 show three representative cases.

As shown in Figure 14, the meme uses the seemingly meaningless absurd statement "potato is my favorite color of the alphabet" to implicitly portray the person in the image as extremely unintelligent. The model fails to recognize the negative depiction of cognitive ability in the absurd text, and also fails to link this depiction to stereotypes associated with people with intellectual disabilities.

As shown in Figure 15, the meme combines the image of a goat with the sexually suggestive text "thank god my ass needed a break", implicitly portraying men as using goats as sexual objects to satisfy their needs. This constitutes derogation and dehumanization of men as a group. The model fails to identify the implicit target of sexual humiliation in the interaction between the visual content and the text, and therefore misclassifies it as ordinary humor.

As shown in Figure 16, the meme appears to show an ordinary scene in which two people are making "vegetable soup" in a kitchen. However, when considered together with the depiction of people with intellectual disabilities in the image, the word "vegetable" is used as a derogatory metaphor that implicitly compares this group to "vegetables", thereby constituting an attack. The model fails to identify the implicit depiction of people with intellectual disabilities during visual analysis, and therefore fails to recognize the discriminatory meaning produced by the connection between the characters and "vegetable soup".

Future research can further construct corpora that contain more implicit attack targets and implicit hateful expressions. This can enhance the model's understanding of complex memes, contextual cues, and the mapping between protected groups and hateful meanings, thereby improving its ability to detect highly implicit hateful memes.

## I Limitations and Future Work

However, existing MLLMs and ProKDA still struggle with highly implicit hateful memes, especially in identifying implicit attack targets and their protected groups. Future work will build meme corpora with more implicit hateful samples to improve reasoning over complex hateful intent.

## J Ethical Statement

**Social impact.** ProKDA can help automatically identify and mitigate harmful online content, reducing the spread of hate speech. By providing both predictions and explanations, the system can help build safer online environments and reduce the workload of human content moderators. We believe such systems have positive value for promoting respectful communication and healthy digital communities.

**Intended use.** We will apply strict access control to the model release, making it available only to researchers who agree to the terms of use. These terms clearly state that the system can only be used for hate speech detection and prevention. Any use that encourages, tolerates, or spreads hate speech or other harmful content is strictly prohibited.

**Misuse risks.** Although ProKDA is not designed to introduce bias, its training data may contain social biases or annotator biases (Pramanick et al., 2021), which may further affect model predictions. To reduce the risk of unfair or disproportionate content moderation, human oversight should be included in real-world deployment.

**Deployment considerations.** Judgments of hateful content are often influenced by cultural background, linguistic context, and subjective interpretation. The same expression may have different meanings in different contexts. Therefore, when deploying such systems in multicultural and multilingual settings, careful review and adaptation based on specific community norms are required.

**Dataset use.** The HMC, MAMI, and PrideMM datasets used in this study are constructed for research purposes and aim to support research on online hate speech detection. We strictly follow the terms of use specified by the authors of each dataset.

**Q:** What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Hateful**




**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** It is a humorous observation about the versatility of a sandwich maker that doubles as an ironing board. The humor comes from the unexpected and practical dual use of a common kitchen appliance. The meme does not contain any elements that promote hatred or discrimination against any group. Instead, it plays on the idea of multitasking in a lighthearted way.

**w/ Stage 1**

**Prediction:** Yes

**Explanation:** The image depicts a woman lying on grass with her shadow cast in a way that resembles a cat, which is used as a visual cue to mockingly suggest that she is a "sandwich maker" who doubles as an "ironing board." The text reinforces this dehumanizing comparison by framing the woman's physical appearance and posture as a tool for household chores, reducing her identity to a stereotype of domestic servitude. Together, the visual and textual elements ridicule women by implying that their primary value lies in fulfilling traditional gender roles associated with cooking and cleaning, while simultaneously objectifying them through the metaphor of a "sandwich maker" and "ironing board". This combination constitutes an attack by perpetuating harmful stereotypes and undermining the dignity of women through humor that trivializes their humanity and societal contributions."

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts a woman lying on grass with her shadow cast in a way that resembles a cat, which is used as a visual cue to mockingly suggest that she is a "sandwich maker" who doubles as an "ironing board." The text reinforces this dehumanizing comparison by framing the woman's physical appearance and posture as a tool for household chores, reducing her identity to a stereotype of domestic servitude. Together, the visual and textual elements ridicule women by implying that their primary value lies in fulfilling traditional gender roles associated with cooking and cleaning, while simultaneously objectifying them through the metaphor of a "sandwich maker" and "ironing board". This combination constitutes an attack by perpetuating harmful stereotypes and undermining the dignity of women through humor that trivializes their humanity and societal contributions.

Figure 8: Case study 1: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Q:** What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Hateful**




**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The meme is not intended to be hateful but rather to express a critical perspective on security measures following a significant historical event. The meme does not use any overtly offensive language or imagery to convey hate. Instead, it uses the gravity of the event and the subsequent discussion about security to make a point about the nature of safety and security measures. The intent seems to be more about raising awareness or debate rather than expressing hatred towards any group or individual.

**w/ Stage 1**

**Prediction:** Yes

**Explanation:** The image depicts the 9/11 attacks, a catastrophic event where 19 individuals hijacked planes and crashed them into the Twin Towers, resulting in the deaths of thousands. The text references this tragedy by stating that these individuals were allowed into the U.S., implying negligence or error in immigration policies. It further generalizes this incident by suggesting that allowing everyone into the country compromises safety, framing immigrants as potential threats. Together, the visual and textual elements evoke the emotional weight of 9/11 while using it as a metaphor to vilify immigrants, portraying them as untrustworthy and dangerous. This combination constitutes an attack by leveraging a historical tragedy to reinforce xenophobic stereotypes and fear-mongering about immigrants.

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts the 9/11 attacks, a tragic event where 19 individuals hijacked planes and crashed them into the Twin Towers, resulting in the deaths of thousands. The text explicitly references these 19 people as ""allowed into the US who should not have been,"" framing their presence as a mistake that led to the tragedy. By associating the victims with the attack, the meme implicitly targets Muslims, as many of the hijackers were identified as being of Muslim origin. The combination of the visual cue—the iconic imagery of the 9/11 attacks—and the textual accusation creates a direct association between Muslims and terrorism, perpetuating a harmful stereotype that unfairly generalizes an entire group based on the actions of a few individuals. This reinforces prejudice and dehumanization, making the meme a hateful expression.

Figure 9: Case study 2: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Hateful**




**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The text "abnormal vs normal" is likely meant to be humorous or thought-provoking rather than to promote hate. The image shows two women in different styles of clothing, but there is no indication that one style is inherently better or worse than the other. The meme is more about contrasting fashion choices rather than making a judgment about who is abnormal or normal. The humor comes from the juxtaposition of the two styles, which can be seen as a playful commentary on individuality and personal expression.

**w/ Stage 1**

**Prediction:** Yes

**Explanation:** The image depicts two women standing side by side in an alleyway, with one wearing a hijab and the other dressed in Western-style clothing. The text "abnormal vs normal" juxtaposes these two figures, implying that the woman wearing the hijab represents "abnormality" while the other represents "normalcy." This visual and textual pairing reinforces a harmful stereotype that equates the hijab with deviance or inferiority, contrasting it against Western attire as the standard of normalcy. By framing the hijab-wearing woman as abnormal, the meme perpetuates a discriminatory narrative that marginalizes Muslims and their cultural practices, fostering prejudice and exclusion.

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts two women standing side by side, with no visible differences in their appearance or behavior that would suggest any form of abnormality. The text "abnormal vs normal" explicitly labels one of the women as "abnormal," which is a derogatory and dehumanizing statement. By juxtaposing the two women in this manner, the meme implies that the individual wearing a hijab is inferior or deviant compared to the other woman, who is labeled as "normal." This combination of visual and textual elements reinforces harmful stereotypes and perpetuates discrimination against Muslims, particularly those who wear hijabs, by suggesting that their identity or practices are inherently abnormal or unacceptable.

Figure 10: Case study 3: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Hateful**




**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The meme uses a humorous comparison between the appearance of certain types of bombs and the concept of "crispy rice." The text "i like my rice crispy please" is a playful reference to the preference for crispy rice cereal, which is a popular breakfast food. The meme juxtaposes this with images of bombs that have a similar appearance to crispy rice, suggesting a humorous and light-hearted take on the visual similarity rather than any form of hate or discrimination.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image depicts various models of nuclear bombs, which are typically associated with destruction and violence. The text, "i like my rice crispy please" humorously juxtaposes the serious imagery of weapons with a casual, lighthearted request for crispy rice, a common food item. This contrast creates an absurd and ironic tone, as the phrase trivializes the destructive power of the bombs by framing it in the context of a mundane preference. The interaction between the visual and textual elements highlights the incongruity, making the meme humorous rather than hateful.

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts the atomic bombs dropped on Hiroshima and Nagasaki during World War II, which caused immense destruction and loss of life. The text, "i like my rice crispy please" uses a casual tone to trivialize the horrific events by likening the mass killings to the act of cooking rice, specifically in a way that implies preference for "crispy" results. This combination of visual and textual elements dehumanizes the victims of the bombings, reducing their suffering to a grotesque joke. By juxtaposing the imagery of war with a mundane food-related phrase, the meme conveys a hateful meaning that mocks the gravity of historical atrocities, perpetuating insensitivity toward the Japanese community affected by these events.

Figure 11: Case study 4: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Hateful**




**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The meme is making a humorous statement about the stereotype that men should be responsible for household chores, specifically loading the dishwasher. The text "a real man loads the dishwasher every night" is meant to be a playful exaggeration of this stereotype. The image of a person lying on a bed with their feet up does not contribute to any form of hate or discrimination. Instead, it adds to the humor by contrasting the idea of a "real man" with the act of doing housework, which is often associated with women in traditional gender roles.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image shows a person lying on a bed with their feet prominently displayed, suggesting relaxation or exhaustion after a day's work. The text, "a real man loads the dishwasher every night," humorously challenges traditional gender stereotypes by implying that men should contribute to household chores, specifically dishwashing, as part of their responsibility. Together, the visual and textual elements create a playful commentary on gender roles, using the image of someone resting to humorously suggest that even men who might prefer relaxation still have a role in domestic tasks. This meme does not express hatred but rather uses humor to promote a more inclusive and balanced perspective on household responsibilities.

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts a man lying on a bed, which visually reinforces the stereotype of inactivity or passivity associated with men. The text, "a real man loads the dishwasher every night," uses the term "dishwasher" as a metaphor for women, reducing them to objects of domestic utility and implying that their primary role is household chores. Together, the image and text create a harmful narrative that dehumanizes women by equating them to appliances. This combination constitutes an attack on women by reinforcing gender-based stereotypes and promoting misogynistic attitudes.

Figure 12: Case study 5: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Hateful**




**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The image shows a U.S. Border Patrol agent interacting with a group of people who appear to be seeking asylum or refuge at the border. The text "family trip in mexico" is a humorous and ironic caption that plays on the idea of a family vacation, which contrasts sharply with the serious and somber reality depicted in the image. This juxtaposition is meant to highlight the absurdity and the stark contrast between the intended purpose of a family trip and the harsh realities faced by those seeking safety and protection. The meme uses humor to draw attention to the issue rather than to spread hate.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image depicts a group of individuals, including children, standing near a U.S.-Mexico border fence, with a Border Patrol agent visible in the foreground. The text, "family trip in Mexico" uses irony to highlight the juxtaposition between the idea of a leisurely family vacation and the serious context of migration and border enforcement. Together, the image and text create a satirical commentary on the complexities of migration policies and the human cost of border control, without expressing hateful sentiment toward any group.

**ProKDA**

**Prediction:** Yes

**Explanation:** The image depicts a group of migrants, including adults and children, standing near a U.S.-Mexico border fence, with a Border Patrol agent present. The text, "family trip in mexico" sarcastically frames their presence as a vacation, trivializing the serious and often dangerous circumstances of migration. By juxtaposing the visual of a vulnerable group with the dismissive language, the meme dehumanizes the migrants, reducing their struggles to a joke. This combination of imagery and text conveys a hateful message by mocking the plight of refugees and migrants, perpetuating harmful stereotypes and insensitivity toward their hardships.

Figure 13: Case study 6: predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Hateful**




**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The text "potato is my favorite color of the alphabet" is a play on words. It's a humorous statement that uses the word "potato" as a pun, implying that the person finds the letter "P" in the alphabet to be their favorite color. This is a light-hearted joke and does not convey any form of hate or discrimination. The meme image and text work together to create a funny and relatable statement without any negative connotations.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image features a person with a neutral or slightly amused expression, which complements the humorous tone of the text. The meme text, "potato is my favorite color of the alphabet," uses wordplay by referencing the letter "P" in the alphabet, which sounds similar to "potato." This creates a playful and absurd connection between food and language, making the joke lighthearted and amusing. The image and text create a harmless and entertaining meme that relies on linguistic creativity rather than any offensive or harmful intent.

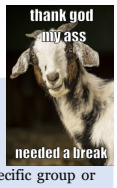
**ProKDA**

**Prediction:** No

**Explanation:** The image features a person with a neutral or slightly amused expression, which complements the humorous tone of the meme text. The text, "potato is my favorite color of the alphabet," uses wordplay by referring to the letter "P" as "potato," a common internet meme trope that humorously misidentifies letters in the alphabet. The visual and textual elements interact to create a lighthearted and playful meme, as the image's neutral demeanor does not detract from the humor but rather complements it by suggesting a casual, everyday context for the joke. There is no indication of malice or targeting any individual or group, making the meme entirely non-hateful.

Figure 14: Error case 1: incorrect predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Hateful**




**Q**: What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The image of a goat with the text "thank god my ass needed a break" is a humorous take on the phrase "thank God I needed a break," which is often used in a lighthearted manner to express relief or gratitude for taking a break from something. The humor comes from the unexpected and somewhat absurd context of a goat being the subject of such a statement. The meme does not target any specific group or individual but instead uses a playful twist on a common expression.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image features a goat with a humorous expression, which is often used in memes to convey a sense of relief or satisfaction. The text, "thank god my ass needed a break," uses informal language and a double entendre, implying gratitude for a physical or emotional release, likely tied to the goat's natural behavior. This meme does not target or demean any individual or group, instead relying on humor derived from the goat's appearance and the wordplay in the text.

**ProKDA**

**Prediction:** No

**Explanation:** The image features a goat with a neutral or slightly amused expression, which serves as the visual focal point. The text, "thank god my ass needed a break," uses informal language and humorously anthropomorphizes the goat, implying it is expressing relief or satisfaction about taking a break. The interaction between the goat's expression and the text creates a lighthearted and relatable tone, as if the goat is sharing its personal experience of needing rest. This meme does not target or demean any individual or group, making it non-hateful in nature.

Figure 15: Error case 2: incorrect predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.

**Q:** What is the meme trying to convey? Is this meme hateful? If yes, how do the meme image and text work together to convey hate?

**Hateful**




**Qwen2.5-VL-7B**

**Prediction:** No

**Explanation:** The image shows two individuals in a kitchen setting, one of whom appears to be wearing an apron, suggesting t hey might be cooking or preparing food. The text "today we are making vegetable soup" indicates that they are engaged in a cooking activity, likely a shared experience. There is no content in the image or text that conveys any form of hate or discrimination.

**w/ Stage 1**

**Prediction:** No

**Explanation:** The image depicts two individuals in a kitchen setting, with one person leaning casually against a coffee machine while the other stands nearby, wearing an apron, suggesting a collaborative or professional environment. The text, "today we are making vegetable soup," humorously implies that the individuals are preparing food, possibly for a meal or event. The interaction between the visual cues and the text creates a lighthearted tone, emphasizing teamwork or shared effort in cooking. This meme does not convey any hateful meaning.

**ProKDA**

**Prediction:** No

**Explanation:** The image shows two individuals in a kitchen setting, both wearing aprons, suggesting they are engaged in food preparation. The text, "today we are making vegetable soup," provides context for their activity, indicating a collaborative effort in cooking. Together, the visual and textual elements convey a simple, positive message about teamwork and shared responsibility in preparing a meal. There are no elements in the image or text that imply negativity or mockery, ensuring the meme does not convey any hateful meaning.

Figure 16: Error case 3: incorrect predictions and explanations by Qwen2.5-VL-7B, w/ Stage 1, and ProKDA.